6-2021

Strategic Default and Moral Hazard in Real Estate: Insights from Machine Learning Applications

Arka Prava Bandyopadhyay

The Graduate Center, City University of New York

Strategic Default and Moral Hazard in Real Estate: Insights from Machine Learning Applications

by

Arka Prava Bandyopadhyay

A dissertation submitted to the Graduate Faculty in Business in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2021

STRATEGIC DEFAULT AND MORAL HAZARD IN REAL ESTATE: INSIGHTS FROM MACHINE LEARNING APPLICATIONS

by

ARKA PRAVA BANDYOPADHYAY

This manuscript has been read and accepted by the Graduate Faculty in Business
in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date Yildiray Yildirim
Chair of Examining Committee

Date Karl Lang
Executive Officer

Supervisory Committee:
Yildiray Yildirim
Linda Allen
Liuren Wu
Johannes Stroebel

THE CITY UNIVERSITY OF NEW YORK

Abstract

Strategic Default and Moral Hazard in Real Estate: Insights from Machine Learning Applications

by

Arka Prava Bandyopadhyay

Adviser: Yildiray Yildirim

Strategic default has been the achilles heel in academic finance for decades. By definition, whether a default has occurred due to strategic motive is unobservable. Moreover, a household has only so many avenues of conducting a strategic default. I use the context of commercial mortgages as property value as well property cashflow co-determine the default decision of these borrowers. I tease out the different strategic aspects of default from the ones emanating from liquidity constraints. The recent advances in Deep Neural Network (DNN), the advent of big data and the computational power associated with it has enabled me to disentangle the motive of default.

Also, agency conflicts of brokers during origination of a mortgage loan and the *moral hazards* thereof has been documented based on the *soft* information about the borrowers. However, there have been few, if any paper, which retains the soft information about the borrowers, post origination, during the life of the loans. There has been a plethora of research about the biases generated towards foreclosures and other adverse outcomes post securitization for the last decade. But the soft information about the borrowers obtained by the brokers have been lost during the pooling process in securitization and there have been famous papers on the loss of information during the securitization process which happens at arms' length from the original lender. I bridge this gap by using novel data on proprietary call transcripts (textual data) between borrowers and servicers. I am also in the process of procuring audio files which can capture mood, content, tone of these communications.

My dissertation documents the use of machine learning (ML) techniques in commercial and

residential real estate to answer long-standing questions, which could not previously be answered due to paucity of data and computational resources. In the first chapter, I run a horserace of Deep Neural Network with other ML models and parametric models to provide a new identification strategy to disentangle liquidity-constrained default and incentives for strategic default. The second chapter attempts to answer the most pressing current socio-economic issue in the United States. Specifically, I compute the social, racial and dollar cost of the CARES Act and find these adhoc policies are as expensive as direct payment of $2,000 to households, if not worse. Finally, in the third chapter I create a novel framework to ingest quantified time-varying soft information from call transcript text data about borrowers in ML models on hard information. I alleviate the information asymmetry between the borrowers and issuers, increase mortgage market efficiency and mitigate the conflict of interest between master servicers and special servicers.

There has been recent literature on the applications of supervised, unsupervised and reinforcement learning in mainstream academic finance. But, very little work is done in the highly illiquid opaque real estate literature using the cutting edge methods in Machine Learning. I take a fresh look at some of the long-debated questions in the literature using some of the machine learning techniques. I am also able to able to use the current COVID-19 pandemic as an exogenous shock for robustness check in most of my current research.

ACKNOWLEDGEMENTS

During the years of my study at the Baruch College in the City University of New York, I have been privileged to know many great people and work with some of them. I have developed the most valuable connection with my advisor, Prof. Yildiray Yildirim, William Newman Chair of Real Estate Finance. Prof. Yildirim's depth of knowledge and generosity in spending time with me have been crucial. His courage in tackling seemingly obscure problems and enthusiasm for exploring new avenues of research will always remain inspiring to me. This dissertation would have not been possible without his constant support.

I sincerely thank Johannes Stroebel, Linda Allen, Liuren Wu, for being on my candidacy, thesis committee and for providing me with great and timely feedback. My humble appreciation goes to Professor Peter Carr at NYU Tandon for the detailed, insightful discussions and Karl Lang for helping me get into the PhD program. I am also thankful to my mentors Prof. Srinivasa Varadhan, Prof. Jaydeep Mukherjee, Prof. Sukhamay Kundu, Prof. Jimmie Lawson, Prof. Alex van der Minne, Prof. Elaine Worzala, Prof. Agostino Capponi, Prof. Piet Eichholtz, Prof. Sumit Agarwal, Prof. Brent Ambrose, Prof. Thies Lindenthal, Prof. Guillaume Haeringer, Prof. Wim Vijverberg, Dr. Xudong An, Dr. Jeff Fisher, Dr. Frank Nothaft, Dr. Joseph Nichols, Dr. Anthony Pennington-Cross, Dr. Michael LaCour-Little, my current co-authors Jiro Yoshida, Patrick Smith, Lilia Maliar, Enrico Onali, Dongshin Kim, Raj Nahata, Mariya Letdin for their insightful discussions, my classmate Joonsung (Frank) Won for thought provoking conversations during our time as PhD mates.

I am highly obliged to professors Pavel Krivenko, Waldo Ojeda, Sophia Gilbukh, Albert Zevelev and EunKyu Lee in the Real Estate department for their kindness. My deepest gratitude goes to my family. My grandfather Prof. Bimal Bihari Banerjee, dad Late Ranjan Banerjee, mom Aheri Banerjee, my sister Raikamal Banerjee and brother-in-law Sougata Chatterjee, maternal uncle Rishiraj Maity, grand uncle Bipul Banerjee and several more have been the greatest inspiration in my life. Last but certainly not the least, my beloved wife Srabonti Roy Bandyopadhyay has helped me drive

this thesis home with her unconditional love and dedication. I dedicate this thesis to my family as a whole-hearted appreciation of everything they have sacrificed for me.

I also want to thank my professional colleagues and acquaintances with whom I have grown as a professional and who have guided or co-operated with me throughout. I would like to thank Davis Edwards and Rajan Tiwari from Deloitte & Touche, Vishwanadha Reddy, Mohd. Zakir Hossain and Yong Wang from Santander Bank, Basile Maire and Eugene Yeboah from UBS, Guang Yu and Indrani De from Nuveen, Josh Harrison and Brett Benson from Rocktop Partners.

To achieve my research goals and to answer long-standing research questions, I have used unique proprietary data for all 3 chapters of my PhD thesis. For Chapter 1, I have used proprietary commercial loan performance data from Trepp, from whom I have obtained the written permission to use the data for academic publication and my PhD thesis. For Chapters 2 and 3, I have used proprietary administrative textual data from a private equity firm Ascension Data & Analytics. I have obtained written permission to use anonymized version of this data for my PhD thesis and publication purposes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

DEEP LEARNING FOR DISENTANGLING LIQUIDITY-CONSTRAINED AND STRATEGIC DEFAULT

We disentangle liquidity-constrained default and the incentives for strategic default using Deep Neural Network (DNN) methodology on a proprietary Trepp data set of commercial mortgages. Our results are consistent during the severe Financial Crisis (2008) and the plausible economic catastrophe ensuing from COVID-19 pandemic (2020-2021). We retrieve the motive of default from observationally equivalent delinquency classes by bivariate analysis of default rate on Net operating income (NOI) and Loan-to-Value (LTV). NOI, appraisal reduction amount, prepayment penalty clause, balloon payment amongst others co-determine the delinquency class in highly non-linear ways compared to more statistically significant variables such as LTV. Prediction accuracy for defaulted loans is higher when DNN is compared with other models, by increasing flexibility and relaxing the specification structure. These findings have significant policy implications for investors, rating agencies and the overall mortgage market.

1.1 Introduction

We reconcile the long-standing debate between two competing theories of mortgage default using Deep Neural Network (DNN) as an alternative identification strategy disentangling competing motives that produce observationally equivalent results (see Indarte 2020 for Regression Kink Design approach). Riddiough 1991 claimed that mortgage default in triggered by life events that reduces borrower cash flows. Foster and Order 1984 opine that default is caused by negative equity (when option to sell the home in the future is worth less that her loan obligations) as borrowers treat their homes like a financial asset. We create a framework in which both of the above can be observed, and more importantly strategic default can be disentangled from liquidity-constrained default. The advent of large scale use of DNN and the computational resources for handling big data has enabled us to address this central research question, impossible even ten years back. Identification strategy has been defined in the literature, vis-a-vis uniqueness for *parametric* models in economics. DNN, with sufficient tuning, can provide a unique non-parametric model which serves the same purpose as canonical identification strategy in a high-dimensional setup. The variable importance (in terms of marginal contribution) of the economically meaningful variables is consistent even during dire economic conditions, ensuring the stability of the non-parametric model and providing a pseudo-identification strategy.

Although defaults are observed, one cannot observe the strategic element of a default as the strategic defaulters are pooled together with borrowers who cannot afford to pay. [1] But, to the best of our knowledge, none have been able to disentangle who defaulted strategically (which can lead to spatial clustering of default) and the intent to default strategically (which can lead to aggregate contagion). [2] Although, moral hazard is time-invariant, but the incentive of a borrower for moral hazard needs to be *triggered*. This is just one special case among the several elements of strategic behavior we document. We use several key covariates, e.g., Net Operating Income, Appraisal Reduction Amount, Prepayment Penalty Clause, Balloon Payment at Maturity, Non-Recoverability, etc. to identify when moral hazard is triggered vis-a-vis higher order non-linear interactions during severe stress in the 2008 Financial Crisis.

Contemporaneous LTV is a proxy for a single default trigger based on property value. Our model incorporates another crucial trigger based on contemporaneous property income (NOI). [3] Net Operating Income (NOI), a key indicator for an investment property's financial standing, is the income from the property after subtracting the operating expenses and vacancy losses (before principal and interest payments, capital expenditures, depreciation, and amortization). The vari-

[1] Bajari, Chu, and Park 2008 estimate the probability of strategic default via a structural model of double trigger, namely, both cash flow and negative equity considerations. Similarly, in a survey of a sample of U.S. households, Guiso, Sapienza, and Zingales 2013 use survey to ask a direct question about a person's willingness to default across different severities of negative equity, keeping the level of wealth and other individual characteristics constant, thereby separating contagion effects from sorting effects, by asking questions about social and moral aspects of default.

[2] Ganong and Noel 2020 try to identify the reason for default from life-events and adverse cashflow events in the context of Residential Mortgages

[3] This is in line with Foote et al. 2009: when equity is negative but above a threshold, default occurs with negative income shock, although our context is commercial real estate (CRE).

ation in the *ability to service debt*, measured by Debt Service Coverage Ratio, i.e., $DSCR = \frac{NOI}{ScheduledPayment}$ provides the identification strategy in disentangling liquity-constrained and strategic defaulters. Commercial Real Estate (CRE) borrowers with $NOI > 0$ & $DSCR < 1$, do not have the ability to service the debt obligation in a given month as they are *liquidity-constrained*, whereas, CRE borrowers with $NOI > 0$ & $DSCR > 1$ have the immediate available liquidity, but may choose to default *strategically*. [4]

We list the possible combinations of LTV and NOI that can *disentangle* Liquidity-constrained Default and the incentives for Strategic Default behavior in Figure 5.1. For a given Loan-to-Value (LTV) bucket, if default rate monotonically increases in Net Operating Income (NOI), we call those defaulters strategic, as their ability to pay increases with NOI but they still increase their default rate. When $LTV > 1$, the borrower is insolvent and Default Rate increases with NOI in the Bivariate Heatmap in Figure 5.2b. In fact, DNN algorithm can identify the threshold of NOI^* (**percentile 6.5** in Trepp data) which disentangles the cases (1) and (2) in Figure 5.1. When $LTV^*(0.82) < LTV < 1$, the borrower passes on the NOI risk to the lender to maintain equity and there is high default across all borrowers, in anticipation of the abrupt jump to insolvency ($LTV > 1$) in the event of appraisal reduction. When $LTV^{**}(0.6) < LTV < LTV^*(0.82)$, there is no risk of negative equity, hence the borrower negotiates the loan and strategic default starts from low NOI^{**} (**percentile 4.0** in Trepp data) in Figure 5.2b which disentangles cases (3), (4) in Figure 5.1. In the LTV bucket $LTV^{***}(0.2) < LTV < LTV^{**}(0.6)$, the borrower default rate is very low as there is neither liquidity-constraint (interest payment is mostly complete) or any incentive for strategic default. When $LTV < LTV^{***}(0.2)$, the loan is close to maturity, hence all borrowers default at a significantly higher rate, due to change in underwriting standard towards maturity or the inability refinance while balloon payment is looming. NOI^{***} can be assumed to be (**percentile 0**, since all borrowers are strategic defaulters and hence case (6) in Figure 5.1 is

[4] Institutional details on Property Type: We assume that property owners manage properties in isolation and do not cross-finance. This is reasonable to assume for investment-type, ring-fenced properties. Some property owners are more constrained for financing than others, e.g., some industries face shock independent of real estate market and suddenly have trouble repaying debt on their buildings. It is reasonable to assume for consumption-type properties held directly be user-firms. In this case, the differences between industries could provide variation in ability.

mostly not realized.

The importance of these strategic (contractual) variables is captured only in DNN, which cannot be captured in Multinomial Logistic Regression, Lasso, Ridge and even in Distributed Random Forest and Gradient Boosting Machine, as evidenced by the higher ranking of NOI over statistically significant LTV vis-a-vis variable importance tables. [5] We then test the robustness of this higher ranking of NOI over LTV, by leaving out other strategic contractual features (year_month, prepayment penalty, balloon payment, occupancy, appraisal reduction, etc.) one at a time. We further test the robustness of the ranking order of NOI over LTV during the Financial Crisis of 2008 by training the DNN on data from 1998-2006 and testing on 2007-2008 data. We further test this order during COVID-19, the current ongoing pandemic and find the exact same results. This proves beyond any doubt, that NOI (and other contractual features mentioned above) are more important than LTV, even during the dire economic circumstances.

There are often situations in which there are no or few good quasi-natural experiments. [6] We exploit the massive proprietary data set on commercial mortgages from Trepp to disentangle liquidity-constrained default (from lack of Net Operating Income (NOI)) from defaults motivated by strategic behavior, as evidenced by default rate increasing monotonically with increasing NOI, for certain LTV buckets. The context of commercial mortgages borrowers is appropriate to document strategic default as these borrowers are institutions [7] and not households. [8] The commercial borrowers are savvy businessmen and hence their delinquency behavior is possibly much more P&L - oriented [9] based on mortgage *contractual features* (prepayment penalty clause, balloon

[5] This could be because: (i) High NOI implies Low Cap Rate for value-added or opportunistic properties compared to core properties; (ii) debt yield from the frothiness of local market; (iii) market cycle channel; (iv) income deficiency related to occupancy.

[6] Differences in Differences require an exogenous treatment and parallel pre-trend, Regression Discontinuity Design requires randomness around one observable characteristic, Instrumental Variables require rigorous explanations on plausibility and ruling out alternative channels, etc.

[7] In an institution, the responsibilities for payment of debt obligation are diffused. The blame for non-payment is not bourne out on one person, but on the institution. Hence there is agency conflict and real possibilities to discern strategic behavior.

[8] Ganong and Noel 2020 find only 3% strategic default for households. Also they define a default as strategic only when the property is under water.

[9] Our definition of strategic behavior is not the same as bourne out of strategies in game theory, but is more in line with the mortgage default literature. The Profit and Loss (P&L) is summarizes the revenues, costs, and expenses incurred during a specified period, and hence P&L management becomes important for businesses from a strategic

payment indicator) and *financial constraints*, such as Net Operating Income (NOI), emanating from the unbalance in terms of the amount and time lag between cost of funding and income cash flows, and much less from macroeconomic conditions, supply and demand in the local geography. We contribute an alternative and new DNN approach (see drawbacks and inconsistencies of the other causal mechanisms described in Black, Kim, and Nasev 2012).

Our paper uses big data with medium frequency to alleviate concerns raised in Manski 2004.[10] We overcome the challenge and observe beliefs [11] and actions in the same data. We implore this novel DNN approach which is much more accurate and robust than Survey measures of (intended) actions and inferring beliefs [12] from actual actions, which assumes that beliefs affect actions. The DNN model can calibrate the thresholds across key variables like NOI, LTV, etc. beyond which there are sharp changes in borrower behavior. The flexibility of not having a pre-specified structure to a model helps us capture univariate and bivariate visualizations of the impact of nonlinearity and higher order interactions.

Goldsmith-Pinkham, Sorkin, and Swift 2018 create of a Bartik instrument by interacting local exogenous industry shares (differential exposure to common shocks) with national industry growth rates. Our DNN model can seamlessly exploit the heterogeneous shares of property-type and measure the differential exogenous exposure to the common COVID-19 shock. The variation in cashflow (NOI) from lessee to borrower across property types provides the exogenous heterogeneity across local geography. We no longer need to argue about the plausibility of the identifying assumption, as the flexible nature of DNN provides an alternative identification strategy that is so robust that it weathers 2008 and current crisis from COVID-19 ongoing pandemic.

Beyond the identification strategy from several variables in this big data setting, our DNN

viewpoint.

[10] Beliefs are essential in appreciating the inter-temporal decisions regarding financial choices. The traditional benchmark has been *rational expectations* based on all (publicly) available information, but has scanty evidence in data (Manski 2004).

[11] The key questions of active research are how beliefs are formed, how different beliefs affect behavior and what are the implications in macroeconomic models and asset pricing.

[12] We contribute to the growing literature of time-inconsistent beliefs or time variation in average beliefs. Giglio et al. 2019 conduct a variance decomposition of beliefs by heterogeneous individual fixed effects, which cannot be explained by observable demographic characteristics. Our paper extends the literature on wealth redistribution between optimists/pessimists (Geanakoplos 2010), which is a model with constant difference in beliefs.

methodology also extends the scope of ”Frailty Model”[13]. DNN not only captures latent time-fixed macroeconomic effect but also loan specific idiosyncratic effects beyond what has been captured in prior literature in Commercial Mortgages. We include 29 variables from Trepp in our DNN model along with state-level macro variables like unempoyment, GDP growth and 2-Year & 10-Year treasury rates and recently created indices. These 33 variables capture the loan-specific unobserved effects and the macroeconomic variables proxy for unobserved common latent variables. Morover, using the DNN, we can capture the highly non-linear interaction among the covariates. We conduct a horse racing among all the models based on misclassification errors for 7 delinquency classes and conclusively show that DNN has the hightest accuracy of prediction along with Gradient Boosting Machine (GBM) in Table 5.4. Since GBM is a greedy algorithm, the variable importance is not robust (in Figure 5.10b) and hence we choose DNN (in Figure 5.10c) as the best model due to it's interpretability along with the accuracy of prediction.

Our findings capture the interplay of borrower behavior, various risk triggers and the overall mortgage market. They significantly differ from the findings of Campbell and Dietrich 1983, Cunningham and Capone 1990, Deng 1999, Elul et al. 2010, Foote et al. 2009, Heimer and Imas 2018, Heimer and Simsek 2019 and others. These prior studies have used loan-to-value ratio, loan age and other loan level variables, as major predictors of borrower behavior. We test whether by adding macroeconomic variables, we can delve into the realm of omitted variable bias found in all hedonic models. We extend the literature on hedonic models by systematically different macro variables which are exogenous in the hedonic regression model beyond the characteristics (used as covariates) and can explain a lot of the unobserved effects Childs, Ott, and Riddiough 1996. Other than 2Yr Treasury Rate and State Unemployment Rate, the macroeconomic variables do not directly affect the strategic delinquency behavior and timing. National interest rates, e.g., 2-Year and 10-Year Treasury rates impact occupancy of commercial properties directly, as well as through state-level GDP. The unemployment is also captured at the state level. The local State

[13] Duffie 2009 created an MCMC methodology that updates the posterior distribution of unobserved risk factors based on Bayes' rule whenever defaults cluster at a given point in time. In the event forecasting literature, such a dynamic unobserved covariate's effect is termed ”frailty”. Yildirim 2008 disentangle the probability of long-run survivorship and the default timing using a mixture model.

level Unemployment Rate in urban centers and occupancy of lessees in commercial properties are highly correlated, because the lessees are the job-creators locally. We claim that all of the above can be captured by NOI, since both occupancy and unemployment rates affect the NOI from the property, as is bourne out of the variable importable in Figure 5.10c where NOI is much higher in importance and internalizes the effect of local unemployment and occupancy through non-linear interactions among these variables.

We also add more broadly to the literature on neural networks. Shallow neural networks have been used in areas of financial economics. Bansal and Viswanathan 1993 use neural networks to estimate the pricing kernel. Hutchinson, Lo, and Poggio 1994 devised the use of neural networks as a nonparametric method of option pricing. Brown, Goetzmann, and Kumar 1998 use neural networks to predict stock markets. Swanson and White 1997 conduct macroeconomic forecasting vis-a-vis neural networks. Lee, White, and Granger 1993 construct tests to capture nonlinearities observed in time series models with the advent of neural networks. Granger 1995 and Kuan and White 1994 study the nonlinear nature of financial time series via neural networks. Khandani, Kim, and Lo 2010b and Butaru et al. 2016 investigate a plethora of machine learning models for estimating financial default. Recent applications of DNN in financial economics include Klabjan 2007 who model market movements. Heaton, Polson, and Witte 2017 use DNN for portfolio selection. The purpose for a deep learning model is bourne out of the need to have transparency and accountability Albanesi and Vamossy 2019.

1.2 Commercial Real Estate Vs Household Finance

The residential real estate bubble from 2004 (emergence) to 2008 (burst) has generated an enormous volume of literature and also relevant policy prescriptions. Surprisingly, the price impact on commercial real estate (potential bubble) has been ignored in comparison. Since the inception of securitization as a means of financing commercial real estate (CRE) mortgages from 1998, the sophisticated B-piece investors have been outbid beyond sustainable long-run fundamentals, over time (from 2004) by investors who ”originated to securitize”, thereby, resulting in decline in underwriting standards in CRE (Levitin and Wachter 2012). Despite some overlap in multi-family property type, Commercial and Residential Real Estate (RRE) is markedly different markets and hence has attracted dissimilar government (e.g., GSE) intervention. One specific difference we will focus on in this paper is the unobservable strategic default behavior of commercial mortgage borrowers, which is different from the residential counterpart due to the nonlinear relationship between multiple property level financials and the mortgage terms. In particular, our goal is to disentangle liquidity-constrained default and the incentives for strategic default based on the Debt-Service Coverage Ratio, a.k.a., DSCR. The the **ability** to pay is turned on when $DSCR \geq 1$ but the *willingness* to pay and motives for strategic default demonstrate the need for a Deep Neural Network (DNN) methodology.

Non-recourse RRE and CRE have an implicit put-option structure equivalent to the repurchase of the loan with the value of the property as the strike, wherein the borrower can meet the debt obligation by the surrender value of the property. This is the theoretical reason for the **LTV** being the primary driver of default behavior in previous literature (e.g., Ambrose and Jr. 2012, Ambrose, Capone, and Deng 2001). Since RRE is both investible and consumable, tax-deduction acts as an incentive, and the foreclosure and recourse laws act as disincentives for strategic default for households. Although the individual CRE loans are much bigger in size compared to their residential counterpart, the partially amortizing structure, defeasance, yield maintenance clauses discourage refinancing and/or curtailment, and hence CRE is exposed to **strategic** default, where the borrowers may choose to stay in 90-120 days delinquency bucket strategically, not being liable to be in foreclosure and not having to be REO. We see a huge surge in these loans in this delinquency bucket after 2008 Financial Crisis in Figure 5.4.

Commercial mortgages, at least the ones which are securitized into a Commercial Mortgage-backed Security (CMBS), are generally used to finance properties which have a stream of income. Therefore, a borrower's default decision depends on not only the asset value (i.e., borrower equity) but also the property liquidity (i.e., property income). Hence, a rational borrower, with a negative equity position, should not default when the net cash flow from property income is positive and enough for debt-service, i.e., $NOI > 0$ & $DSCR > 1$. So, a model for a rational borrower's default decision is co-determined by both property value and property income. Commercial mortgages are partially amortizing (7-12 year term and 25-30 year amortization schedule), i.e., a balloon payment is due when the mortgage matures. Borrowers generally refinance the current mortgage to fund the balloon payment, which may be **strategic**, towards maturity due to increased interest rates or the underwriting standards having tightened, even for a borrower making payments on time.

The lender makes a judgement about the riskiness of the borrower in terms of continued payment towards a loan obligation and underwrites the risk vis-a-vis the mortgage rate in Figure 5.5. Although, the lender is well aware of the reputation and past loan repayment behavior of the borrower, the type of the borrower is noisy and hence this leads to **adverse selection**. The lender makes an actuarily fair **take-it-or-leave-it** offer, adding some risk premium, to the borrower at origination. Moreover, there is competition among borrowers in the same business of leasing income-producing commercial properties. Due to the search friction for the lessee in Figure 5.5, there is considerable uncertainty about the NOI coming from the lessee's rent payment. The borrowers have no bargaining power in terms of loan pricing, but they could use the act of strategic delinquency as an **insurance policy** against the premium they had to pay at origination via the mortgage rate. We assume that act of strategic delinquency of some borrowers can be captured by the first-order stochastic dominance of the cumulative default rate (higher default rate for strategic defaulters than liquidity-constrained defaulters) of the bad-type over the good-type, for different buckets of LTV in Figure 5.2b. There could be moral hazard from the lessee in Figure 5.5, in terms of continued rent payment and servicers are used by the borrower as a commitment device, since the lease is not negotiation-proof.

The commercial lender (debtholder) and the commercial borrower (equityholder) enter into a **contract** (firm) for the business of leasing/renting out the property to a lessee. The prospect of debt renegotiation increases the expected payoff to shareholders in default, and induces them to anticipate the timing of default by strategic default (Favara, Schroth, and Valta 2012), hence increasing the bargaining power of the commercial borrowers (equityholders in this context). There is a subtle difference between **firm** strategic default and the strategic default of the **CRE borrower**, as the CRE borrower usually have different LLCs for each property and hence their strategic default is more property-centric, specifically in terms of the rent payment from the lessee in the property.

Our results indicate that the effects of property income, prepayment penalty clause and balloon risk are significant to assess total credit risk.[14] The estimation of the economic parameters is a nontrivial problem, given the massive sparsity and the paucity of CMBS data from a historical perspective. The empirical mortgage literature identified a linear combination of variables for the commercial mortgage credit and prepayment risk including creditworthiness and free cash flow of the entity, current leverage ratio, loan age, interest rates, and CMBS indices (e.g. Furstenberg and George 1969, Curley and Guttentag 1974, Campbell and Dietrich 1983). The commercial mortgage performance data, however, tell a different story. The presence of nonlinear effects obviates the need for a more general form but it is difficult to identify all the factors and their mutual interactions. Instead of specifying a functional form for commercial mortgage performance, we include all possible factors and let the data dictate the model, which also allows for highly nonlinear interaction terms between factors. Since, our data set is nationally representative, the pooled model computes an estimate of aggregate default risk in the commercial mortgages especially well for 2007-2009. Our estimation result provides a ranking of individual commercial mortgages in terms of their delinquency behavior and can be aggregated to a systemic measure of default risk in the commercial sector.

The remainder of the paper proceeds as follows. We first motivate a toy theory model in Section 1.3, provide details on the big data in Section 3.2 and provide descriptive statistics and conduct rigorous exploratory analysis to give an idea of the trends in data. In Section 1.5, we motivate Naive Bayes, Multinomial Logit & Ordered Logit, in Section 1.6 Distributed Random Forest, Gradient Boosting Machine models and provide empirical results and list the deficiencies in each of them. In Section 1.7, we describe the DNN model and motivate how the DNN model

[14] The simultaneous inclusion of four significant risks: market, credit, prepayment (Christopoulos, Jarrow, and Yildirim 2008) and liquidity (Ambrose and Sanders 2003) makes CMBS modeling extremely hard intractable. The CMBS loan pool tranche cash flows and corresponding allocation rules, the prepayment restrictions and the prepayment penalties are heterogeneous across the various CMBS trusts.

can alleviate most of the issues in the earlier models. We also point out the key findings of the paper in this section and how they differ from the earlier literature. We test robustness of NOI and LTV order during the ongoing COVID-19 pandemic in Section 4.1.1 and document industry-level heterogeneity. In Section 3.6, we provide concluding remarks followed by references.

1.3 A Simple Theoretical Motivation

We provide a simple model framework, in line with Guiso, Sapienza, and Zingales 2009, to motivate that the ”Optimists” (or Strategic Defaulters) would prefer to maintain a consistently higher LTV during the good portion of business and economic cycle. They will then have the option to strategically default in the future. Whereas, ”Pessimists” (Non-Strategic or Liquidity-constrained Defaulters) would prefer to continually reduce LTV, in anticipation of different forces increasing LTV in the future and also to alleviate the consequences of default in the event they are liquidity-constrained. This differential behavior across the cohort of borrowers will price in their heterogenous beliefs (π) in the expectation of occupancy of the property. The differential behavior of these two types of borrowers across different LTV buckets are explained in Figure 5.2b, based on NOI and based on other variable interactions in the other subfigures in 5.2. The motivation for using Net Operating Income (NOI), Debt-Service Coverage Ratio (DSCR), Balloon Payment, Scheduled and Unscheduled Payments, Occupancy is explained with the toy model below.

In residential market, while negative equity, in nonrecourse states, is necessary for strategic default ibid., it is by itself not sufficient. There are several frictions that make defaulting unappealing in nonrecourse states. Consider a borrower who owns a property worth A_t at time t and faces a bequest mortgage balloon payment of B_T. The borrower will not default as long as $A_t > B_T$ from a financial standpoint. However, for strategic default, there are aspects to look into beyond the financial loss from defaulting. Not defaulting renders a borrower the benefit of defaulting in dire conditions in the future. The intertemporal substitution of default choice is co-determined by timing of appraisal reduction, non-recoverability depending on whether the master servicer/special servicer has ceased advancing for the related mortgage loan. Also, by defaulting she plausibly faces higher borrowing cost in the future due to differential credit-rationing by the lender, since lenders are generally NPV-neutral and default is a deadweight loss for them. Let K_t be the net benefit (opportunity cost of cash) of not defaulting at t. Then a rational borrower will not default if $A_t - B_T + K_t > 0$.

If the commercial borrower is not constrained a bequest lump sum payment due in the near future, then her decision of strategic default is far more involved, because of the same intertemporal choice and the trade-off therein, i.e., the decision to default today Vs postponing the default decision later. Also, the option to default later depends on the borrower's ability (DSCR) to serve her debt obligations, which has high correlation with the probability of occupancy and positive cash flow from the lessee in the property. If the property is vacant or if the lessee does not pay up, the borrower is likely to default later and is unable to use the value of the option. Let $V_T = A_T - B_T + K_T$, where T is the balloon payment date. Then the value Bajari, Chu, and Park 2008 of not defaulting at T-1 is:

$$V_{T-1} = a_{T-1} - m_{T-1} - B_T + K_{T-1} + (1 - \pi_{T-1})E_{max}(V_T, 0) \tag{1.1}$$

where a is the monetary value of the cashflow and the serviceflow experienced from time T - 1 and T, m is the total (scheduled and unscheduled) mortgage payment between T-1 and T, π_{T-1} captures the probability of vacancy of the property (i.e., not having a lessee, and E is the expectation operator. The value of not defaulting at a generic date t can be deduced from backward induction:

$$V_t = a_t - m_t - B_T + K_t + (1 - \pi_t)E_{max}(V_{t+1}, 0). \tag{1.2}$$

From the above equation, the decision of strategic default at t can be captured by the relationship below:

$$StrategicDefault = F(A - B, a, m, \pi, K). \tag{1.3}$$

The functional form of $F(.)$ is extremely difficult, if not impossible to pin down. Even locally, to define $F(.)$ piecewise using Implicit Function Theorem, one would need the partial derivative of $F(.)^{-1}$ with respect to the shortfall $A - B$, the monetary value of the cashflow and the serviceflow a, the scheduled and unscheduled mortgage payments m, belief about the property occupancy π, non-monetary benefit K to be well-behaved. We show in 5.3, this is not the case. The LTV is a function of the shortfall $A - B$, the NOI is a function of the cashflow and the serviceflow a, Debt-Service Coverage Ratio (DSCR) is a function of $\frac{a}{m}$, occupancy is the expectation of the belief π, non-monetary benefits K are mostly unobservable. This is further clouded by the fact that Recourse Laws are not strictly implemented in most states. Bankruptcy Laws need to be fairly strong in a state to reinforce recourse laws.

We give indication from the data, how non-linear the interactions among the above variables can be in Figure 5.2 and hence resort to the most flexible yet robust DNN methodology in Section 1.7.

1.4 Data

We have monthly proprietary novel data set of 91,767 loans (only US loans) from January 1998 to September 2016 from Trepp[15] We exclude CRE loans, as our research focused on NOI generated from income-producing properties only found in CMBS loans.

We include the variables used in previous CMBS literature, like An, Deng, and Gabriel 2009, Ambrose and Sanders 2003 and preclude the following key loan-specific variables: log(original balance), LTV, time ofamortization, time to maturity, lockout, lockout expiration, corporate bond-credit spread Titman, Tompaidis, and Tsyplakov 2005, yield curve, mortgage-treasury rate spread, region dummy, seasonal/quarter dummy, among others.

We finally decide to use loan-to-value (ltv), occupancy rate (occ), tranche loan-to-value, (securltv), tranche weighted average cost (securwac), annualized gross rate (actrate), outstanding scheduled principal balance at end of current period (obal), derived most recent net operating

[15] Trepp is the leading provider of analytics, information, and technology to the global CMBS, commercial mortgage finance, and banking industries. Trepp is the largest commercially available database containing detailed information on over 1,800 deals and more than 100,000 loans, which support close to $800 billion in securities. Deal coverage includes North American, European, and Asian CMBS, as well as Commercial Real Estate backed CDOs.

income (noi), outstanding legal remaining outstanding principal balance reflecting defeasance of the loan as of the determination date (balact), securitization balance of the loan predged to the trust (face), most recent appraised value else securitization appraised value (appvalue), total amount of principal and interest due (actpmt), regularly scheduled principal to be paid to the trust (curschedprin), principal prepayments and prepayments (full or partial), discounted payoffs, and/or other proceeds resulting from liquidation, condemnation, insurance settlements (curunschedprin), interest basis of an adjustable rate loan (pmtbas), net proceeds received on liquidation of loan (liqproceeds), expenses associated with the liquidation (liqexpense), difference between Net Proceeds (after Liquidation Expenses) and Current Beginning Scheduled Balance (realizedloss), amount received from a borrower as a pay off a loan prior to the maturity or anticipated repayment date (pppenalties) as the loan-specific variables. Age of the property is include as a control in addition to the age of the loan. We add age^2 as as a control variable too to capture the non-linear relationship of aging of the loan with the delinquency classes. We calculate ”time to maturity” to extract any strategic default behavior closer to the realized maturity of the loans.

We use the loan vintage (to capture if origination and underwriting standards have an effect on the delinquency class of the loans), 51 states in USA (msa, county, zip have severe missing values, hence the identification comes at a state level), property type (we bucket thousands of property types into 8 unique types), fixed/floating as dummy variables. We use ”Number of Properties” (numprop) in a deal as a deal-specific variable. We control for refinance pipeline and/or balloon payment by assigning a dummy if a loan is within 3 months threshold to its original scheduled maturity date. We use MIT Commercial Index, National Council of Real Estate Investment Fiduciaries (NCREIF) regional property value indices. Additionally, we include state-level quarterly GDP (converted to monthly), monthly historical unemployment data by state and historical interest rates of different maturities.

For the classifications models to generate realistic results and capture the marginal contributions of the features in a scale-free way, we convert numerical variables like: $x :\longrightarrow \frac{x-min(x)}{max(x)-min(x)}$. This keeps the distributional characteristics of the numerical variables, but makes them all scale-free so that their marginal contributions towards the output can be uniform. We avoid the other more frequently used *standardization* technique where $x :\longrightarrow \frac{x - Mean(x)}{StdDev(x)}$ as it converts all variables into standard normal. The entire valuable information, e.g., skewness, kurtosis and all distributional characteristics are lost in this imposition of normal distribution across feature space.

The summary statistics for the cleaned data containing 9,617,333 observations of continuous variables is provided in Table 5.1. ”One hot encoding” technique converts categorical variables as binary vectors without any order.

1.5 Parametric Models and Empirical Results

Our first set of empirical results are based on **parametric** models: Naive Bayes, Multinomial (with Lasso and Ridge regularization) & Ordered Logit harnessing the unprecedented size of our sample set and the heterogeneity in the incentives of default and beliefs we investigate. The models calculate the accuracy of prediction for 7 different delinquency states starting from Current/Performing classes **W0_30D** which includes ”loans with payments not received but still in grace period or not yet due”, Late/Non-Performing classes **W30_60D**, **W60_90D** which includes loans with ”Late Payment beyond 30-days but less than 60 days, beyond 60-days but less than 90-days, Default state **W90_120D** ((within 90 to 120 days of delinquency), Liquidation Proceedings & Final Resolution state **B120D** (beyond 120 days of delinquency), combined together as ”limbo” loans. We add further states in the **PrfMatBal** (Performing, Mature and Balloon Payment due) and **NPrfMatBal** (Non-Performing, Mature and Balloon Payment due) classes to capture the incentives delay in resolution for foreclosed loans to REO/prepaid. Although **PrfMatBal** is a performing loan, but it can be anywhere between 0-90 days of delinquency. **PrfMatBal** are also close to maturity, rendering itself vulnerable to strategic behavior from changing interest rate environment and underwriting standards. Hence, **PrfMatBal** is treated as a separate delinquency status. The same argument holds even stronger for **NPrfMatBal** loans. We motivate below the reasons why these parametric models misrepresent the risk for the delinquent loans in this context.

A **Naive Bayes** classifier estimates the conditional a-posterior probabilities of a categorical variable given independent covariates using the Bayes rule. The assumption of **independence** of the covariates is key to the success of the Naive Bayes (NB) classifier. We see that **W0_30D**, **W30_60D** & **W60_90D** classes have less mis-classification in Table 5.4 error in NB than other models, since the assumption of independence among the co-variates holds until a loan is in these classes. This analysis is still kept in the paper to motivate why we eventually need DNN as a means of avoiding this strong assumption of independence among the covariates.

Ordered Logit exploits the natural order of delinquency classes and computes transition probabilities in that order. Ordered Logit does not allow all the back transitions from a worse delinquency state to a better delinquency state, which can be shown in a Finite State Automaton. **Multinomial Logit** assumes Independence of Irrelevant Alternatives (IIA).2, which is not true in this situation as we will see in the next section. Suppose, hypothetically, there are two choices given to a borrower to be either **within 30 days of delinquency** or **between 90 days and 120 days of delinquency**, which is not true in this situation as we will see in the next section. Suppose, hypothetically, there are two choices given to a borrower to be either **within 30 days of delinquency** or **between 90 days and 120 days of delinquency**. Clearly, the borrower would like to stick with the first choice, as the second choice classifies him/her in the default category and is detrimental for her credit-worthiness from a lender's perspective. Now suppose, one more choice for being in **30 days to 60 days of delinquency** is given to the borrower, s/he may choose to rather be in this new state instead of less than 30 days of delinquency and may **strategically** miss one payment if there is a great investment opportunity for him/her in that one month horizon. In fact, none of the models (except Naive Bayes) can distinguish these three classes (**W0_30D**, **W30_60D** & **W60_90D**) and considers all of them as **Current Loans** in Table 5.4.

The granularity of delinquency classes brings out the gradual transition of loans into adverse states rather than simply having a cutoff for default which would imply that we are assuming that loans ”Within 30 days delinquency”, ”Between 30 days and 60 days delinquency” and ”Between 60 days and 90 days delinquency” have the same default risk. If all the loans which are less then 90 days delinquent had the same default risk, a borrower would only pay off just before 90 days delinquency in order to avoid default and facing derogatory consequences. The fact that the above three buckets represent different default risk categories imply that the borrower's default behavior will change when she/she is between 30 days and 60 days of delinquency compared to the situation when all the above three categories are bucketed together as "Non-Default".

In Table 5.4 the row labels are the predicted classes and the column labels are the actual classes. As is evident from the Sensitivity and Error , the Multinomial Logistic Model can correctly classify the Current or "**W0_30D**" really well, but the Specificity is really low, i.e., the model cannot classify the loans that are **not** in "**W0_30D**" correctly vis-a-vis the "**W0_30D**" class. Also the error rates for the classes "**W30_60D**", "**W60_90D**" are 100% which means the model cannot identify any those classes correctly. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly the Multinomial Logistic Model. In fact, some of the risks (Current Note Rate, LTV, Unemployment Rate, etc.) are misrepresented in Multinomial Logit, e.g., if local Unemployment increases, the *Current* Commercial Loan Default should increase (Table 5.2). **Lasso** and **Ridge** do not improve the performance of Multinomial Logit in Table 5.4.2.

1.6 Vanilla Machine Learning Models & Empirical Results

In the current section, we parallelize Random Forest and implement adaptive gradient boosting after bagging. We finally implement DNN in Section 1.7 and compare the prediction on different mortgage states on the holdout sample.

1.6.1 Distributed Random Forest

The confusion matrices of the delinquency classes for in-sample/training set are calculated for the entire data in Table 5.3 and also subsample in Table 5.3 until the December, 2006 for stress testing the robustness for Out-of Sample Prediction during the Financial crisis in Figure 5.10. As is evident from the Error in Table 5.4, the **Distributed Random Forest** Model can correctly classify the Current or "**W0_30D**" **completely** in Table 5.4. Also the error rates for the classes "**W30_60D**", "**W60_90D**" are 98% which means the model cannot identify any those classes correctly but better than Multinomial Logit Model. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly but better than the Multinomial Logistic Model in Appendix .3.

As is evident from the **Out-of-Sample** Errors in Table 5.4, the Distributed Random Forest Model can correctly classify the Current or "**W0_30D**" **completely**. here the column labels are the predicted classes and the row labels are the actual classes.Also the error rates for the classes "**W30_60D**", "**W60_90D**" are 100% which means the model cannot identify any those classes any better than Multinomial Logit Model. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly but better than the Multinomial Logistic Model **Out-of-Sample**. The Out-of-sample predictions worsen during the Financial Crisis. [16]

1.6.2 Gradient Boosting Machine (GBM)

As is evident from the **In-Sample** Errors in Table 5.4, the **Gradient Boosting Machine** can correctly classify the Current or "**W0_30D**" **completely**. Also the error rates for the classes "**W30_60D**", "**W60_90D**" are almost 100% which means the model cannot identify those classes any better than Multinomial Logit Model. Similarly, the classes "**W90_120D**" and "**B120D**" are also identified very poorly but better than the Multinomial Logistic Model **In-Sample** in Figure 5.3. We also attach the Variable Importance for GBM during the using data before Financial Crisis in Figure 5.10.

Here the column labels are the predicted classes and the row labels are the realized delinquency classes. The out-of-sample predictions for GBM perform as good as DNN in our preliminary analysis. This methodology sums the importances over each boosting iteration (see the gbm package

[16] Along with training a model that classifies accurately in a hold-out sample, one needs to be able to interpret the model results. Feature importance, the coefficients of linear models, is the crucial tool to identify important features. Almost all random Forest (RF) routines also provide measures of feature importance via permutation importance. Permutation importance is obtained by randomly shuffling each predictor variable by estimating the effect on model accuracy. This technique is model-agnostic because of independence from internal model parameters even while using Lasso or Ridge regularization in the presence of highly correlated features.

The prediction accuracy on the out-of-bag data is recorded for each tree. Each predictor variable is permuted and the same routine is repeated. The difference of the two are then averaged across all trees, and further standardized by the standard error. If the standard error is equal to 0 for a variable, the division is not done. here is the Variable Importance table 5.10a for the Random Forest Model Khandani, Kim, and Lo 2010a. The Variable Importance for Out-of-Sample predictions during the Financial Crisis in Figure 5.10 give similar results.

vignette).4.

1.7 DNN for disentangling Default Incentives

DNN is a form of machine learning with multiple layers that learns multiple levels of representations for different levels of abstraction Sirignano, Sadhwani, and Giesecke 2016. It captures associations and discovers regularities within sets of patterns; it is suited for high volume, high dimensional data. It performs well when the relationships are dynamic or non-linear in Figure 5.2, when the standard regression models perform very poorly. No assumptions on normality, linearity, variable independence are needed.

We use a multi-layer feedforward DNN, trained via stochastic gradient descent and tuned by back-propagation. Each computational node trains a cache of the global model parameters on its locally available data with asynchronous multi-threading and contributes to the global model via model averaging across the DNN. We tune both and Optimizer and Model-specific Hyperparameters (described in Appendix .5.2). We use SMOTE technique to reduce class imbalance. According to Appendix .5.3 [17], we use Variable Importance to compare the most significant marginal contributions of the features (described in Appendix **??**).

1.7.1 Model Results

We motivate the highly strategic delinquency behavior of the savvy commercial borrowers/ business-owners from two different angles. We provide evidence from the Trepp data in Figure 5.6a that from 2012, the number of loans have remained flat but the outstanding balance of loans have steadily increased until 2016. This could have serious implications. There can only be two possibilities: if the same loans stay and there is no origination at all, and further if the outstanding balance is increasing, it means there is serious delinquency in the loans and the servicers are unable to secure the payment from the borrower and all these loans could potentially become limbo

[17] As is clear from the similar counts of the loans of different categories in the in-sample confusion matrix in Table 5.3, we have **undersampled** the W0_30D class/Current Loans to alleviate the class imbalance problem. The Out-of-Sample predictions across different delinquency classes are as good as GBM in Table 5.4.

loans. Figure 5.6b furthers the narrative. From mid-2014, the age of the loans is decreasing and the time-to-maturity is increasing. This could mean that from mid-2014, there are an equal number of originations to the number of maturing loans. But the fact that the Outstanding Balance is increasing in this entire period could only mean that the same loans are getting rolled over to new contracts, when balloon payments are missed during maturity.

Figure 5.6c clearly shows that LTV (widely used in previous literature and used by most banks/asset managers for credit risk calculations) is flat throughout the data horizon. The interest rate is decreasing almost monotonically in the data and there seems to be no sensitivity of LTV to interest rate. This means LTV is probably not the right way to think about credit risk. It could also be that the commercial borrowers **target** LTV. They strategically make payments towards their obligation so that the ratio of ”Book Value of Loan” and the ”Value of the Property” remains relatively stable over time. It would make sense for them to do this as banks/asset managers use LTV at origination as the primary determinant of creditworthiness of the borrowers. Further, the Contemporaneous LTV (CLTV) is used to calculate LGD (Loss given Default or 1-Recovery Rate). So, CLTV could also be targeted and there is no evidence of voluntary deleveraging from the borrower inspite of widely changing macro-economic conditions, e.g., interest rate.

Figure 5.6d corroborates that the NOI monotonically increases in the data and the occupancy is almost 100% in the entire data. So, there may be strategic saving of internal cash flow from income producing properties. Because of the strictly increasing NOI level, the strategic dominance of NOI over other factors can have disastrous aggregate macroeconomic consequences. To capture this, we try different methodologies like vanilla models (Naive Bayes, Multinomial and Ordered Logistic) and machine learning models (Distributed Random Forest, Gradient Boosting) and finally Deep Neural Network (DNN) and find that DNN is best positioned to address the above issue and does capture NOI as the most significant strategic variable from the Variable Importance (VI) tables of the models. This difference does not stem from sample bias. This is the core reason for our choice of big data for training all the models. Also, Trepp is the largest provider of CMBS data and hence the sample is representative of the entire market and does not have any selection bias.

We normalize Net Operating Income (NOI) as a percentage in the pooled data for loans and create histograms of relative frquency of the number of loans in different delinquency classes with respect to the different percentiles of NOI. We see a sheavy support for the relative frequency across all the delinquency classes at the NOI percentages 5%-7%. We call them **dominant** NOI buckets. We show the distribution of different delinquency classes with all the NOI buckets including the dominant ones (see figure 5.7a). The significant heterogeneity across the delinquency classes and the highly non-linear effect of NOI towards the strategic choice of the borrower to be in a specific delinquency class is not bourne out of this diagram. Beyond the above dominant buckets, we see highly non-linear **strategic** behavior for commercial mortgage borrowers to choose different delinquency classes for different buckets of NOI. To visualize this, we zoom in and remove the dominant buckets and form the **rescaled** (without dominant NOI bucket masses) relative frequency histogram across all delinquency classes.

Figure 5.7b highlights the complex relationship that exists between the percentage of loans across the different delinquency classes "Within 30 Days"(**W0_30D**), "30 Days to 60 Days"(**W30_60D**), "60 Days to 90 Days"(**W60_90D**), "90 Days to 120 Days"(**W90_120D**), "Beyond 120 Days" (**B120D**) and the buckets of net operating income (NOI) excluding the dominant NOI buckets, which can be incentivized by the macro-economy. The sensitivity varies significantly in a highly non-linear way in both magnitude and sign. There is a **U-shaped** choice between NOI buckets 37%-45%for the borrowers in the delinquency class W90_120D. This means that when a borrower is already beyond the default threshold of 90 days, but less than the cutoff of 120 days, they are incentivized to stay there for a while and time their future payments based on cash flow. Since these NOI bucktes are higher, the borrowers make some profit from the income generated from the property, but they still stay at the same delinquency classes and do not pay-off the earlier missed payments to come back to the Current State (less than 30 days of delinquency). Similarly, the borrowers in delinquency class B120D choose to be in lower NOI buckets in a non-linear way. This is because of the lack of net cash flow income for them to be able to pay off the earlier missed payments. They end up in a vicious cycle of making less money from the property and becoming worse off in terms of their creditworthiness. We call them ”limbo” loans as these loans stay in this state for a while before they are resolved. The sensitivity estimates generated by vanilla models can misrepresent the influence of risk factors because of naive choice of linear specification. This can make it difficult to make economic conclusions from the borrower behavior. In our approach, the relationship is entirely dictated by data, thereby minimizing model misspecification and bias of the variable estimates.

The accuracy of predictions change dramatically, if NOI is taken out. We also conduct a robustness check by leaving out each of the *strategic* variables from the DNN model. When year and month fixed effects are taken out in Figure 5.9, NOI loses its importance significantly! This clearly indicates that NOI is not a statistically significant variable by itself. It is used strategically by borrowers when clustering of macro-economic events happen and when NOI is taken out, the constraint variable like prepayment penalty clause and voluntary prepayment variable like current unscheduled pricinpal payment show up higher in the variable importance in Table 5.4 than LTV. Similarly, when Prepayment Penalties are taken out of the list of variables. When Balloon Payment constraints are taken out of the list of variables.

The neural networks literature provides us two methods for constructing Variable Importance (VI) scores, the Garson algorithm, later modified by Goh 1995, and the Olden algorithm Olden, Joy, and Death 2004. Both algorithms use the network’s connection weights to generate these importance scores. The Garson algorithm utilizes all weighted connections between the nodes to estimate VI. Olden’s algorithm, sums the product of the raw connection weights between each input and output neuron across all hidden neurons. For DNNs, we use a similar but more interpretable method due to Gedeon 1997 which considers the weights in connection to only the first two hidden layers. For Deep Learning, there is no impact of scaling, because the numbers were already scaled. hence, the relative importance is the same as the absolute importance in Figure 5.9.

Unlike model-specific approaches, model-agnostic interpretation via VI methods are more flexible. We further investigate model-agnostic methods for quantifying global measures of importance using three different approaches: 1) PDPs, 2) ICE curves, and 3) permutation Greenwell et al. 2018.

As is clear from the preliminary analysis, the Net Operating Income (NOI), the Prepayment Penalty clause and the Balloon Payment triger are significantly high in the variable importance table 5.4. NOI is even higher than LTV as found in the VI tables for other previous models. This provides evidence on how these three less statistically significant features contribute much more towards the classification, via highly non-trivial and non-linear interactions with more statistically significant variables.

Our DNN Variable Importance table in Figure 5.9 shows that NOI is the key endogenous feature for understanding strategic delinquency behavior of the commercial mortgage borrowers. We intend to further investigate how prepayment penalty clause and indicator for balloon payments co-determine the strategic delinquency behavior along with the NOI using Shapley values by capturing the marginal contributions.

To test the robustness and stability of our DNN, we present the Variable Importance Plots of Predicted Default Rate from June 2006 to December 2008 with several features in Distributed Random Forest (DRF) in Figure 5.10a, Gradient Boosting Machine (GBM) in Figure 5.10b and DNN in Figure 5.10c, trained on data before June 2006 and motivate why we need a highly non-linear model and also why we allow for high-dimensional interaction among the borrower-specific, macroeconomic, spatial, vintage effects in the features. Time-to-Maturity, Geographical cross-correlation, NOI, Appraisal Reduction, Bankruptcy Flag, Property Type, Non-Recoverability, Appraised Value supercede Securitized LTV in the Variable Importance chart for DNN in Figure 5.10c. Morover, Balloon Payment supercedes LTV, corroborating the robustness of our DNN model. DRF captures non-lineary of the covariates but still ranks LTV much above NOI and other strategic variables in Figure 5.10a, even after tuning and bagging. GBM is a greedy algorithm and hence finds more occurances of local minima for LTV and hence ranks LTV higher than NOI in Figure 5.10b, even after boosting.

1.8 Conclusion

Using DNN, non-linearities of dependence of the response and interactions among features can be captured, without specifying the relationships apriori. DNN provides an alternative identification strategy, specially when there are no available quasi-natural experiments. Net Operating Income, Prepayment Penalty Clause, Appraisal Reduction, Non-Recoverability, Bankruptcy Flag, Liquidity Proceeds, Liquidity Expense and Balloon Payment Indicator co-determine the strategic delinquency behavior of a commercial mortgage borrower. Loan-to-Value is unable to capture this Strategic behavior as obviated by the Variable Importance charts since statistical significance cannot capture the non-linear effect during Financial Crisis. Hyperparameter Tuning during the implementation of DNN is still an art and not a science. The classification of critical delinquency states of systems when the agent decisions are endogenous while the data is highly unbalanced across states can only be captured through DNN.

CHAPTER 2

COST OF MISALIGNED CARES ACT: OVERCROWDING, SELECTIVE VERIFICATION AND UNINTENDED RACIAL CONSEQUENCES

I utilize a novel data on proprietary servicer call transcripts to investigate *overcrowding* of mortgage forbearance program (13% with only 1.5% unemployed) contained in the CARES Act. I document *selective verification* of unemployment status (financial hardship) by the servicer. I also discern *unintended consequences* (disparate impact) of 2.4% for Inbound and 2% for Outbound communications for African American borrowers (without the servicer having race information) by the servicer to reduce ex-post risk. My finding sheds light on the poor-targeting of Government programs (FHA, VA, USDA) during exacerbated income shocks from COVID-19 and estimates a $5.76 Trillion exposure from plausible non-payment of residential mortgage debt obligations from forbearance.

2.1 Introduction

> ”This bill is free money for everyone. Proponents don’t care if you’re fully employed or own your own house or own your business. ”Free money for everyone,” they cry. And yet if free money were the answer, if money really grew on trees, why not give more free money? Why not give it out all the time? Why stop at $600 a person? Why not $1,000? Why not $2,000?” — Senator Rand Paul (Dec 21, 2020) [1]

I allude to the heated debate regarding the amount of relief and the mechanisms of providing relief in the current socio-political environment, charged with racial tension and unforeseen wrath of the COVID-19 pandemic. Senator Rand Paul recently derided COVID-19 relief aid that Senate Leader Mitch McConnell negotiated with the Democrats. President Trump signed the Omnibus

[1] Source: https://www.wsj.com/articles/notable-quotable-rand-paul-11608925235

Bill that includes $600 relief per person before the year end. We already know what happened when money was directly handed out as relief to households and business owners in March 2020. Chetty et al. 2020 track economic activity at a granular level [2]. Using these data, they study the mechanisms through which COVID-19 affected the economy and conclude that these bailout money is mostly used as precautionary savings and not spent on consumption, which was the desired outcome to stimulate the economy.

I investigate the other mechanism of relief provided by the U.S. Government to the real economy, namely, the Coronavirus Aid, Relief and Economic Security Act (”CARES Act”) and find that it was no better. I try to tease out the tension between giving away money directly to the pocket of households Vs providing relief through forbearance, foreclosure moratorium and eviction moratorium. The former exacerbates the budget deficit and the latter seems more targeted and cost-efficient. I show that it is not the case and the latter leads to more overcrowding by certain borrowers who don’t need relief, selective verification of financial hardship by the servicer and unintended racial consequences (disparate impact) for African Americans.

Government intervention has been the mechanism to attenuate large unexpected shocks like COVID-19 or housing crashes like the 2008 financial catastrophe. I investigate the effectiveness or lack thereof of such Government interventions selectively applied on Government programs by Housing and Urban Development (HUD). Specifically, I find irrational (opportunistic) behavior among borrowers with Govt.-backed loans vis-a-vis spatial overcrowding and rational (logical and conservative) behavior from borrowers with Conventional loans. Conventional borrowers with lower income and more financial constraints are the hardest hit by the wrath of COVID-19 and hence apply for forbearance and conventional borrowers with relatively higher income and stable jobs do not take up forbearance even though they may be affected in the short-term. On the other hand, the borrowers with Govt-backed loans opportunistically apply for forbearance from almost all income brackets in Figure 13 and I provide evidence in Figure 14 that some of these borrowers avail forbearance even though they are not *unemployed* or have not had any *curtailment of income*.

[2] statistics on consumer spending, business revenues, employment rates, and other key indicators disaggregated by county, industry, and income group in real time using anonymized data from private companies

Hence, the Govt.-backed loan borrowers spatially overcrowd forbearance applications (see Figure 27). Moreover, I show that servicers are much more lenient towards borrowers with Govt.-backed loans and stringent with borrowers with Conventional loans by verifying the employment status thereby scrutinizing the forbearance applications of the latter. The CARES Act and the uniqueness of the data set allows me to disentangle both of these information asymmetries in the same set up. Specifically, I use the servicer call transcripts to extract *soft* information about the borrower and create a narrative retrieval apparatus via Inbound/Outbound calls [3] capturing the intent of those communications. For marginal borrowers who have missed a couple of payments and whose loans are about to become non-performing, I see a significant spike in Foreclosure Moratorium by the end of March 2020 in Figure 15. In April 2020, the servicers face a crucial choice whether to approve these marginal borrowers in their forbearance applications or advise them to avail the foreclosure moratorium. I see that, by May 2020, most of these borrowers have been dissuaded and informed about their ineligibility (due to adverse delinquency status) by the servicer. This could have serious implications of a looming housing crisis if there is a massive surge of foreclosure after foreclosure moratorium ends.

The first few cases of the global COVID-19 pandemic in the United States were diagnosed in early March 2020. The global COVID-19 pandemic precipitated a growing public health crisis and necessitated President Trump to sign the CARES Act into law on March 27, 2020. The CARES Act[4] contained numerous fiscal stimulus programs and policy directives designed to aid households and businesses negatively affected by the government mandated shutdown (business closings) and social distancing restrictions imposed after March 15, 2020. Of course, there is a lot heterogeneity in terms of the implementation of the shutdown orders DLima, Lopez, and Pradhan 2020 and actual implementation of the mask and social-distancing policies across counties and states, but these are not the prerogative of this paper. Instead, I investigate the effect of the CARES Act as a point-in-time Government policy and the implications thereof. In order to protect households from

[3] Inbound communications are initiated by the borrower and Outbound communications are initiated by the servicer. The dialogue between the borrower and the servicer is recorded in call transcripts, email exchanges and/or physical mails.

[4] For details, check https://www.govtrack.us/congress/bills/116/s3548.

unemployment or income curtailment resulting from government ordered business shutdowns, Title IV of the CARES Act stipulated a foreclosure moratorium and created a payment forbearance program for federally-related mortgage loans.[5] One important feature of the forbearance program is that it does not require that borrowers prove financial hardship or be in a delinquency status before requesting forbearance. Indeed, I show in Tables 18 and 19, that borrowers having performing Govt-backed (PL_Gov) loans strategically take advantage of the CARES Act and apply for forbearance (13.18%) even though they were not unemployed (1.48%) or did not suffer from major financial hardship (4.73%). After a forbearance application is approved, lenders are prohibited from collecting accrued interest, late fees, convenience fees, or other charges associated with the missed payments. The Department of Housing and Urban Development clearly demarcated the rules governing the forbearance program for Federal Housing Administration (FHA) loans on April 1, 2020 and because of this I see most Forbearance applications approved on April 9th by this specific servicer. There is another peak of forbearance applications when the Housing Government Sponsored Enterprises (GSEs), Fannie Mae and Freddie Mac, announced their forbearance plans covering conventional mortgages on April 21, 2020. While the CARES Act specifically targets payment relief to FHA/VA loans and conventional mortgages backed by the GSEs, it does not clearly indicate/delegate specific rules non-government backed (or private-label) mortgages, which leads room for interpretation and discretion by the servicer.

The typical forbearance is approved for 6 months (or two consecutive 3 month blocks). The borrower can choose to preempt the forbearance approval by starting to pay if she regains her financial status and ability to continually pay from a new job or other source of income. The borrower can also choose to use the 6 months forbearance approval, even if she regains her ability pay before the end of 6 months. Beyond the first 6 months, the borrower can be granted another forbearance period of 6 months. The CARES Act stipulates as maximum of one full year from the time a borrower first applied for forbearance relief. In practice, no servicer would require a

[5] Section 4022 specifies that lenders must grant a minimum 60-day foreclosure moratorium beginning March 18, 2020 on all Federally backed mortgage loans. The section also requires that lenders create a 180-day forbearance program for borrowers experiencing direct or indirect financial hardship due to the COVID-19 crisis.

lumpsum payment afterwards. There will mostly be a partial payment plan after the end of the forbearance spell (similar to a loan modification or an additional refinance loan). The CARES Act mandated that the servicer inform at least the Gov-backed borrowers about their eligibility of availing forbearance if their loans were performing ex-ante. There is a grey area for marginal borrowers who are about to become non-performing in terms of their payment ex-ante. This ambiguity is the crux of the negative amplified outcome of the misaligned CARES Act. Servicers can choose to offer Foreclosure Moratorium to marginal borrowers who are not yet formally in foreclosure and/or bankruptcy proceedings. Servicers can also offer loan modification. This leaves a discretionary room for the servicer. The conflict of interest between the master and special servicer is well-documented in the real estate finance literature. This incentive incompatibility of the special servicer with the investor and the issuer can plausibly lead to a surge of foreclosures in the near future and a vicious cycle thereafter depending on whether the economic recovery is tick-mark shaped or W-shaped.

In this paper, I analyze the differential impact of the CARES Act forbearance provision that specifically targeted government backed loans (FHA/VA and GSE mortgages) but not mortgages originated outside these government agencies. I utilize a novel administrative dataset obtained from a mortgage servicer that comprises a portfolio of FHA/VA and private-label mortgages. First, I manually identify a set of keywords for identifying Inbound/Outbound communications and borrower reported unemployment status in Section .6.1 in the Appendix. Because of unique and real-time nature of this almost-daily administrative transcript of the communication between the borrower and the servicer, I am able to track the borrower-noted unemployment status, which is much more accurate than the aggregated estimates of unemployment from Bureau of Labor Statistics. Also the reason for these communications and the incentive compatibility between the borrower and the servicer can be captured from Inbound/Outbound communications. I use these indicator variables in my regression specification. There are several aspects of reporting Unemployment status via Inbound/Outbound communications. Also, Inbound/Outbound communications can take place due to different reasons. I do not distinguish these different aspects of Unemployment Status, Inbound/Outbound communications in the regression specification as it would require big data for exploiting such a rich specification. Also, I use logistic regression for number of forbearance applications on several variables including but not limited to Inbound/Outbound communications and borrower reported unemployment status, which is a linear model and hence cannot capture the non-linearity of the different aspects of these indicator variables and their interactions. Instead, I bring to bear an application of natural language processing (NLP) technology in order to identify whether the borrower or servicer initiated the forbearance process. I also identify whether the borrower indicated financial distress (e.g. job loss) as a motivating factor in requesting forbearance. To the best of my knowledge, the use of transcripts of communications between the servicer and borrowers to identify requests for forbearance and financial distress has not been pursued before in academia. In contrast, typical studies in the mortgage literature that examine mortgage default of modification rely on datasets derived from servicer records containing hard coded data Agarwal, Chang, and Yavas 2012 and Mayer et al. 2014. In other words, access to servicer-borrower communications is not available in most mortgage performance data. As a result, my study provides a unique insight into the initial process of requesting mortgage payment relief that has heretofore been unavailable to researchers.

In line with the design of the CARES Act forbearance program, there is a higher incidence of forbearance with government backed loans in response to communication initiated by the servicer (denoted as ”outbound”). The CARES Act required that servicers proactively reach out to borrowers with details about the forbearance program. The Act leaves a grey area and does not stipulate that servicer proactively contact private-label or non-government back loans. A positive increase in forbearance in the private label set follows from a borrower initiated (denoted as ”inbound”) communication. Unlike government-backed mortgages, the servicer is able to demand that the borrower prove financial hardship before granting forbearance for Non-Gov borrowers. Consequently, I find a lower incidence of forbearance within this set of loans following communication with the servicer. The endogeneity from the strategic overcrowding of forbearance applications by Gov-backed performing loan borrowers and the endogeneity emanating from selective verification by the servicer *undo the CARES Act*. To overcome these endogeneity issues, I implement Differences-in-Differences approach towards the end of the paper. However, the CARES Act does help some financially constrained borrowers and, at the same time, does not bail out the servicers.

Based on the available data, I find evidence for the following research questions. I formulate them as conjectures/claims which *undoes* the CARES Act (not technically hypotheses, as I am not rejecting the null) and corroborate them using logistic regression with and without fixed effects and Differences-in-Differences approaches (for alleviating endogeneity concerns) in the following sections of the paper.

Proposition 1: Borrowers having Gov-backed performing loans are **overcrowding** forbearance applications, even if they are not unemployed/have any financial hardship from curtailment of income.

Proposition 2: Servicers **selectively verify** the employment status of Non-Gov loan borrowers and dissuade marginal borrowers by offering loan modification and/or foreclosure moratorium, to preempt/prevent them from availing forbearance.

Proposition 3: Servicers' behavior have **unintended distributional implications** towards African American and Hispanic borrowers, their forbearance applications are accepted only in dire financial conditions.

Proposition 4: **Poorly-targeted** CARES Act helped some borrowers and did not bail out servicers; still some borrowers overcrowded and some servicers prevented certain borrowers from availing forbearance.

There has been a plethora of research following the inception of the CARES Act. An et al. 2020 provide evidence that lower-income and minority borrowers have relatively higher nonpayment rates during the COVID-19 pandemic. McManus and Yannopoulos 2021 find mortgage forbearance in the COVID-19 period are similar to those during natural disaster. CoibionGorodnichenkoWeber study the impact of large one-time transfers to individuals from the CARES Act on the consumption, saving and labor-supply decisions. Carroll et al. 2020 model responses of households to past consumption stimulus packages and find, during the lockdown, most types of spending are undesirable and/or impossible. They also opine the jobs lost during the lockdown may be permanently gone. Humphries, Neilson, and Ulyssea 2020 document how COVID-19 impacted small business owners and how these effects have evolved since CARES Act. Boar and Mongey 2020 documented that many unemployed workers received benefits that exceeded wages. Akee et al. 2020 dissect the US Department of the Treasurys distribution of first-round CARES Act funds to Indian Country in terms of relief funds. Petrosky-Nadeau 2020 investigate the existence a reservation level of payments in which an individual is indifferent between accepting and refusing an offer under the increased unemployment insurance and extended duration of relief provided by the CARES Act. Neilson, Humphries, and Ulyssea 2020 explore information frictions and the ”first-come, first-served” design of the Paycheck Protection Program (PPP). Baker and Judge 2020 explore the counterfactual where critical forgivable loan program by the government and find this debt relief alone will not provide the cash they need to retain employees, pay timely rent, etc. Wilson and Stimpson 2020 claim that the adverse policy environment has made immigrant communities particularly vulnerable to community spread of COVID-19. Capponi, JIA, and Rios 2020 show the existence of a self-reinforcing feedback loop between foreclosures and growth in house prices.

Racial implications in Real Estate Literature have been studied from various perspectives in Cashin 2008, Bayer, Ferreira, and Ross 2016, Denton 2017, Spalding 2008, Jackson 1980, Pace, Barry, and Sirmans 1998, Schafran and Wegmann 2012. To the best of my knowledge, the unintended racial implications of selective verification of the unemployment status (financial hardship) has not been studied previously in this literature.

2.2 Data

I utilize a proprietary administrative dataset containing detailed information on residential mortgage performance that was collected from daily mortgage servicing logs.[6] The data consists of the servicing records spanning the period from January to May 2020 for 19,418 loans that were active

[6] This dataset was provided by a private equity firm that focuses on real estate investments.

as of January 2020. The data contains a rich set of variables that provide information about the borrowers and their loans. For example, the data records the loan-to-value ratio (LTV) at origination, the loan's current interest rate, balance and appraisal, whether the loan is a fixed-rate or adjustable-rate mortgage, the loan purpose (cash out refinance, home improvement, rate-term/vanilla refinance, or purchase), the property type (modular home, single family, multi-family, condominium, townhouse, or planned unit development (PUD)) and occupancy status (owner-occupied, second home, or investment property), the borrower's credit (FICO) score at origination, loan modification flag, and amount of any corporate advances paid by the servicer on behalf of the borrower.[7] For a subset of the data, I have the employment industry (mostly Small and Medium Enterprise for the borrowers in this portfolio) and credit tradeline information, which provides a proxy for the household liability (mortgage, credit card, auto loan, student loan, etc.).

As typical in mortgage servicing data, my dataset contains detailed *hard* information on each loan's payment status. Using this information, I define loans as being performing (PL) or non-performing (NPL). I classify loans as performing if their payments are less than 60-days delinquent and non-performing if payments are 60-days or more delinquent. The data indicates whether the loan was originated as part of a federal government-backed insurance program (Federal Housing Administration (FHA), Veterans Administration (VA), or US Department of Agriculture (USDA)) or if the loan was originated as a conventional or non-conforming mortgage.[8]

In addition to the typical information collected from mortgage servicing tapes used in prior

[7] The borrower makes monthly payments comprising principal, interest, taxes, and insurance (PITI). The TI part is usually put into an escrow account. The servicer then draws down that escrow account to pay the taxes and insurance premium on behalf of the borrower. That account can typically hold up to 2-yrs of TI funds. The servicer can earn a float on those funds. There is a separate reserve fund set-up where the servicer deposits part of the PI to hold in reserve in case the borrower misses a payment. This reserve is funded out of the monthly servicing fee that the servicer deducts from the PI before passing it to the investor.

Corporate advances are expenses paid by the servicer and recoverable from the borrower. Typical corporate advances include attorney or court fees associated with a foreclosure or required insurance premiums paid on behalf of the borrower. The servicer passes all these costs through to the investor. Each month, the advances they make are netted out of the remittance that goes to the investor. If the pool doesn't generate enough cash to cover the advances (rare), then the investor has to write a check to cover the advances. On a particular loan, advances balances get paid down from the cash that comes in – either if the borrower makes a payment or the loan liquidates when the borrower is delinquent and there are advance balances which get paid down first before any cash is applied to PI.

[8] Conventional mortgages refer to loans eligible for purchase by Fannie Mae or Freddie Mac while non-conforming refers to jumbo mortgages or subprime mortgages that are not eligible for purchase by the government sponsored enterprises (GSEs).

studies e.g. Cordell et al. 2015; Agarwal et al. 2018; Buchak et al. 2018; Kruger 2018; Conklin et al. 2019, the unique feature of this dataset is that it contains transcripts documenting the communication between the borrowers and the servicer call centers. These transcripts contain real-time (almost daily) loan status updates and thus provide a preview of the loan status variables contained in typical mortgage servicing records. Thus, using these servicer comments, I create a time series of several important indicator variables to capture the borrower's payment intention or financial stress. For example, I search the transcripts for the keywords "COVID" and "forbearance" to identify if and when a borrower had a conversation with the servicer regarding forbearance options emanating from CARES Act (enacted in March 2020).[9] To capture financial stress arising from plausible employment interruption, I use natural language processing (NLP) techniques to identify whether a borrower is unemployed in Appendix .6.1. I also identify borrowers who are experiencing income disruption via the keywords "curtailment of income". This is a stronger indicator than unemployment for a COVID based forbearance application and, more importantly, is a proxy for the borrower's inability to pay. I also scan for the words "foreclosure" and "moratorium" to identify borrowers who are currently in a foreclosure moratorium status, a relief channel for non-performing loans. Finally, I identify whether the servicer comments originated with the borrower (inbound) or from the servicer (outbound), detailed in Appendix .6.1. My regression specification is very rich and I use Forbearance Applications as my dependent variable and Inbound/Outbound Communications and Borrower noted Unemployment as explanatory variables. However, I do not use interactions of Borrower noted Unemployment and Inbound/Outbound Communications in the regression specification. Similarly, there are several aspects of Inbound/Outbound Communications which captur the rationale and the incentives of those communications. I showcase these details via NLP separately instead of directly invoking them in the regression. I provide the t-SNE diagrams[10], which are 2-dimensional projections of word clouds similar to the words

[9] See the Appendix for a complete list of key words used to denote various aspects of the mortgage servicing calls.

[10] t-SNE is a non-linerar algorithm for visualizing high-dimensional data van der Maaten and Hinton 2008, via two-dimensional maps, a faithful representation of those points in a lower-dimensional space. Secondly, t-SNE has a tuneable parameter, perplexity, which tracks the balance between local and global components of data, in a sense, ”a guesstimate” about the number of close neighbors each point has.

”Unemployed”, ”Inbound” (IB) and ”Outbound” (OB) in Figure 16. The strategic element of the forbearance applications is higher for the inbound comments. The creation of these flags (dummies) is uniquely able to determine the delinquency status of the borrower and her propensity to apply for forbearance.

As evident in Figure 16, there are several clusters in the t-SNE, which necessitates a deeper dive into interaction of unemployment with IB in Figure 17, unemployment with OB in Figure 18 and unemployment per se (without IB and OB) in Figure 19. In Figure 16, one can notice 4 clusters. On the South-East corner, words related to ”ob” (in violet) which are not so much related to unemployment per se but *continual renegotiation* between the borrower and the servicer. On the North-East corner, the cluster represents words related to ”ib” and ”inbound” (in sea-green) where the borrower seems to be making the case for loan modification and other *offers* that they can avail from the servicer. In the North-West corner, the two subclusters are entangled, one of them highlights the *occupancy* and related issues emanating from unemployment and the other specifically relates curtailment of income with the *intent* of the borrower. The use of these keywords in defining the ”Unemployment”, ”Inbound” and ”Outbound” flags (dummy variables) helps me tease out the tension among these aspects of borrower and servicer behavior, which have not been explored previously in academia, to the best of my knowledge. Figure 17 details 3 of the 4 clusters in Figure 16 on the facets of ”Inbound” and ”Unemployment”. The *partial adjustment* and *payment disputes* in the North-East corner point out aspects of financial hardship and renegotiation related to Inbound calls from the borrower. The other entangled cluster captures several borrower aspects related to *intent* and servicer response to *offer* the borrower more favorable terms for the loans. The *refusal* of forbearance for certain borrowers is also captured in the South-West corner of Figure 17. Figure 18 captures the selective *verification* of the unemployment status of certain borrowers by the servicer. The southern part of Figure 18 underscores the typical outbound conversations related to borrower financial health and dire personal circumstances and the ensuing renegotiations. Finally, Figure 19 encapsulates all words related to ”Unemployement” in a giant cluster.

I also find clear indications of multiple facets of IB and OB which can capture the reason for the communication between the borrower and the servicer and also their incentive compatibility. I create 3 sub-clusters of IB t-SNE for: (1) IB and Financial Hardship in Figure 20, (2) IB and Family, Property, Loss in Figure 21, (3) IB and Legal Issues in Figure 22. In the same chain of thought, I create 3 sub-clusters of OB t-SNE for: (1) OB and Loan Modification in Figure 23, (2) OB and COVID in Figure 24, (3) OB and Legal Issues in Figure 25. Figure 20 captures several aspects of the Inbound communications related to financial hardship, e.g., liquidating Vs keeping property by the borrower, change in owner occupancy due to employment transfer and distance of the property from the new job, inability to sell the property, borrower illness, etc. Not all of the above are verified for borrowers with Gov-backed loans and hence this category of Inbound communications heavily contributes towards the opportunistic/strategic elements of borrower forbearance applications. Figure 21 points out Inbound communications related to marital matters like marriage/divorce/death of spouse, excessive obligations, casualty loss, etc. Figure 22 is more comprehensive and nuanced to the legal aspects of Inbound communications, e.g., prior bankruptcy, ownership transfer, business failure, leniency from military service, non-payment by the tenant, payment disputes, etc. Figure 23 captures the keywords related to selective verification by the servicer, e.g., decline, payment dispute, disposition, suspense, reapply, ineligible, denial, flag, intermittent, etc. Figure 24 directly captures Outbound communications related to COVID-19 and as one can see there are not many words related to forbearance, since the servicers approve forbearance applications from the borrowers who fall under the purview of the CARES Act and try to dissuade other borrowers when the borrowers initiate conversations related to forbearance and/or foreclosure moratorium. Figure 25 details the Outbound communications related to the specific servicer attributes such as performance, borrower indication, involuntary, representation, temporary, title, signal, silence, commitment, satisfaction, judicial, document, identification, etc.

Table 18 reports the descriptive statistics for the mortgages as of the April 2020 servicer reporting date. Panel A summarizes the statistics for all loans while Panels B and C summarize the data based on whether the mortgages are government-back loans (Panel B) or non-government program loans (Panel C). The average loan had an origination amount of approximately $96,700 on a property with an appraised value of approximately $120,000. The average loan-to-value ratio at origination was approximately 80%. Since the data consists of first-mortgages, second mortgages, and home-equity loans/lines of credit, the average loan amount is lower than samples comprising exclusively first-mortgages. The mean borrower credit score at origination was 613, reflecting the higher proportion of subprime borrowers in the portfolio (65% of the sample). Panels B and C reveal significant differences in the government (FHA/VA) and conventional (non-government) loans. For example, FHA/VA mortgages had higher origination loan-to-vale ratios than conventional loans (96% versus 72%) and higher average current balances ($122,308 versus $56,648, respectively). The geographic distribution FHA/VA loans in the dataset in Figure 26 is consistent with the distribution of FHA market shares reported in Ambrose and Pennington-Cross 2000 and Ambrose, Pennington-Cross, and Yezer 2002. Across all loans in the sample in Panel A, the call center logs indicate that 7% of the loans were flagged for a Covid-19 related forbearance. In addition, approximately 3% of the borrowers indicated an employment problem and 5% reported having a serious income issue (curtailment of income). During April, 29% of all borrowers were contacted by the servicer (outbound) while 22% of the borrowers initiated contact with the servicer (inbound). Panel B reveals significant differences in call center activity for government and non-government loans. I see that 42% of government loan borrowers experienced a call center initiated contact (outbound) with the servicer and 32% initiated (inbound) contact with the servicer. In contrast, 23% of the non-government borrowers experienced a call center initiated contact and 17% initiated a contact with the servicer. Consistent with the FHA being more aggressive and faster in responding to the Covid-19 crisis, I see that 11% of these borrowers had a Covid-19 related discussion with the servicer as compared to 4% of the non-government loan borrowers.

Table 19 goes one step further on Panels B and C in Table 18, by creating separate buckets form Performing (PL) and Non-Perfoming (NPL) loans among Govt-backed (Gov) and Non-Gov-backed (Non-Gov) loans. In Panel A, PL and Gov is highest group applying for COVID-19 compared to unemployed borrowers in the group. Number of incoming calls is also very high for this group. For Gov-backed PL loans, COVID-19 Forbearance applicants (13.18%) is much higher than unemployed (1.48%) and Curtailment of Income (4.7%). So, at least 8% of these borrowers are definitely strategic. The number of Inbound calls is also much higher for Gov-backed loans, leading to the possibility of strategic behavior. The original LTV is much higher for Gov-backed PL loans, still the current mortgage rate for Gov-backed loans is much lower and their FICO scores are relatively higher. In Panel B, PL and Non-Gov still has higher COVID-19 Forbearance applicants than the number of unemployed borrowers in the same group, but much less than PL and Gov, although their current delinquency status is much better. For Non-Gov PL loans, only the unemployed borrowers are applying for COVID-19 Forbearance. This happens to be same as Curtailment of income, which means the only source of income for these borrowers is from employment. In Panel C, NPL and Gov takes advantage of Foreclosure Moratorium, as they are mostly ineligible for Forbearance applications. DLQ for NLP is 4 .37 (in Panel C) and 4.39 (in Panel D), meaning 120+ day delinquent, hence they are ineligible for Forbearance. The servicer may be letting the borrower know their ineligibility for Forbearance since the servicer has corporate advances in place. In particular, the servicer increases their call volume and frequency for NPL non-GOV cases since corporate advances are the highest in that bucket.

The spatial distribution of key variables in this paper provide stronger evidence of the strategic behavior of borrowers in PL_Gov group. There are plenty of Non-Gov loans in Las Vegas in Figure 26 depicted by yellow color and Las Vegas was one of the major fatalities of the COVID-19 pandemic as the entire state runs on gambling and tourism revenue which were shut down abruptly. However, the forbearance applications of residents of Las Vegas were overcrowded (in Figure 27) by residents from the northern mountain states who arguably were affected much less severely by the first major hit of the COVID-19 during March - April. The geographical distribution of the curtailment of income in Figure 28 also paints a similar picture in April 2020 data, where the residents of only a few pockets were facing severe financial hardship, but forbearance applications were rampant from all over the United States by the opportunistic/strategic PL_Gov borrowers. The distribution of Inbound calls and Outbound Calls provide preliminary evidence of strategic behavior by the borrowers and the selective verification by the servicer respectively in my data.

2.3 Servicer Perspective and Institutional details

If the forbearance is extended for another 3 months after June 2020, this could have serious cash flow implications for the investor (bond-holder). This is crucial since after 4 months of Forbearance, the servicer is not required to make any advances to the investor. So, essentially, after 4 months, the investors would have to take the hit, if the borrower decides to be delinquent and not make timely payments after 6 months of Forbearance. It is a high Cash flow risk whose downside is not protected. If the borrower was 2 payments down or in foreclosure, they are still being reported to the Credit Bureau as 2 payments down or in foreclosure, and they cannot refinance until they become current. They would qualify to be considered for a modification. From the servicer's perspective, all of the non-government insured loans will need to demonstrate that they have been impacted by the pandemic (i.e. borrowers are unemployed). The purpose of the loan does not avid the CARES Act. Certain loans would not be covered, such as second homes and second liens. However, second liens undergo an analysis on whether there is enough equity to initiate foreclosure, and second homes would require full workout submissions. All of this is somewhat moot since there most states have moratoriums on referral to foreclosure, foreclosure sale, and eviction (excluding vacant properties).

Another key question for Non-Gov loans (e.g., Conventional Loans) in this portfolio is if the servicer is indifferent to the outcome: Foreclosure Vs Forbearance, since the advancing costs are taken care of by the investor. It is fair to say the the servicer has no exposure to either PI (principal interest) advances or corporate advances once the loans are acquired by the PE firm (provider of this data). On the government portfolio, the guarantors (HUD, VA and USDA) have insisted on forbearance being a preferred decision before foreclosure. On the conventional portfolios, although there is no guarantor dictating decision paths, the CFPB (Consumer Financial Protection Bureau), the CARES Act, and many states (such as NY and CA) do provide regulatory guidance over offering workout opportunities (forbearance) over foreclosure. They also indicate that the borrower owns certain responsibilities in requesting that help that is more expansive then what the

GSEs require, which has allowed the servicer to be more insistent on documentation of financial hardship instead of wanting to take time off from making payments. The servicer does insist on documentation of a lost income rather than a just a phone call request for a forebearance plan. Also, from a business perspective, the servicer earns service fee income on both a loan in foreclosure or on forbearance, so the servicer is indifferent on the path based on similar income and similar expenses. However, if they can ultimately cure the default with a forbearance that turns into a modification, then the servicer would prefer that solution, i.e., they will ultimately make more income. The servicer would also incur additional staffing costs for loans that go through the Foreclosure/REO process than a loan that goes through the forbearance reinstatement process, again slightly favoring the forbearance, where the servicer will ultimately incur less staffing expenses.

Related to the above aspects, how the servicer is incentivized/contracted may lead to making more financial gains in one choice Vs the other. Of course, the servicer has to consider the local jurisdictions/rules in place for offering loss mitigation options before starting foreclosure. This becomes crucial, since 70-75% of the borrowers who had applied for Forbearance are extending their non-Payment from being in Forbearance. So, there is a clear choice to be made by the servicer for Non-Gov loans (which are not directly under the purview of the CARES Act). The incentives of the servicer for Foreclosure Vs Forbearance paths from being a performing loan, need not be aligned with the interest of the PE firm acquiring these loans. Although the guarantors have no authority over the PE firm portfolio, the government and state agencies/regulators do, and their preference leans towards the borrower (voter). The servicer is definitely risk adverse with picking a fight with a regulator. Even if a request for an extended forbearance is denied, there is a limit to moving forward with the foreclosure process. Properties that are vacant can usually proceed with foreclosure (referral or sale), but not in all states, and in some counties, the Courts have not reopened to allow movement.

2.4 Univariate Analysis & Empirical Model Specification

The comparison of differences in Covid-19 forbearance responses and call center activity across various loan types is captured in Table 2, which shows that significant differences exist between performing and non-performing loans for government-back and non-government back mortgages. Government-backed performing loans (columns 2-4), had approximately the same rate of in-bound and out-bound communication (15% versus 17%) and for the non-performing government-backed loans, outbound rate (51%) is significantly higher versus inbound (29%). The difference is not surprising since servicers are required to attempt to contact borrowers who have missed monthly payments in an effort to mitigate potential losses associated with default. It is interesting to note that only 9% of non-performing borrower communications mention Covid-19 forbearance relief in contrast to the 13% rate observed in performing borrowers. I also find several interesting insights from Columns (5) through (7) that compare performing and non-performing non-government loans. First, similar to the government loans, I see that outbound communication is significantly higher in the non-performing set than in the performing set (54% vs 10%). Again, this is to be expected since non-performing loans require direct servicer intervention whereas servicers typically respond to borrower requests for some action in the performing loan group. In comparing the differences between government and non-government mortgage portfolios, I see that borrowers with performing loans in the FHA/VA and conventional portfolios have approximately the same unemployment indicator, which is again consistent with borrowers who are current on their mortgage payments having low employment problems. However, the Covid-19 forbearance rates are significantly higher in the government backed loan group than in the non-government portfolio. Thus, given that employment issues are roughly equal between government and non-government borrowers, the higher Covid-19 forbearance rates in the government portfolio is evidence of strategic forbearance requests coming from the borrowers who are current on their government insured mortgages.

I test for evidence of strategic forbearance, controlling for differences across borrowers and mortgages, with the following logistic regression framework:

$$\begin{aligned} Pr(F_i) &= \Phi(\delta_1 IB_i + \delta_2 OB_i + \delta_3 G_i + \delta_4 P_i + \delta_5 IB_i \times G_i \\ &\quad + \delta_6 OB_i \times G_i + \delta_7 IB_i \times G_i + \delta_8 OB_i \times P_i + \gamma U_i' + X_i' \beta + \kappa_c) \end{aligned} \tag{2.1}$$

The parameters δ_1 through δ_8 are the primary coefficients of interest and capture the differential effects on the probability of the borrower entering forbearance based on whether the borrower (δ_1) or servicer (δ_2) initiated contact during April 2020. The interaction terms (δ_5 through δ_8) thus capture the differential servicer incentive effects based on the type of loan (government insured or non-government) and the loan payment status (performing versus non-performing).

2.5 Multivariate Analysis

Table 20 reports the regression results for early forbearance. Columns (1) and (2) report the results for all loans with and without county fixed effects, respectively. Not surprisingly, the positive and statistically coefficients (at the 1% level) in all columns 1-4 in Table 20 for inbound and outbound confirm that the probability of forbearance increases with any communication with the servicer, whether initiated by the borrower or servicer. This is to be expected since forbearance requires an active request on the part of the borrower and thus necessitates communication with the servicer. The estimated coefficients for inbound and outbound in my preferred specification that includes county fixed effects (column 2) reveal that the probability of forbearance is essentially the same regardless of whether the servicer or borrower initiated the contact.

Turning to the differential between FHA/VA in column (3) or private-label mortgages in column (4), the estimated coefficient is not statistically significant indicating no difference in the incidence of requesting forbearance between these borrower groups (all else being equal). However, it is interesting to see that the coefficient for the variable indicating whether the loan is currently performing is also not statistically significant. Thus, I do not find evidence that borrowers who were already in financial difficulty and were delinquent on their loan prior to the pandemic are taking advantage of the forbearance option. In fact, the interaction of inbound with performing is statistically significant and indicates that requests for forbearance are most likely arising from borrowers who were current on their loans at the onset of the pandemic.

Columns 1-2 in Table 20 show that Inbound calls for Performing loans (IB_PL) is positive and significant with and without county fixed effects. Same is true for Outbound Calls for Performing Loans (OB_PL) in Columns 1-2. Essentially, both the borrower and the servicer makes more communications if the loan is performing. I test this with triple interaction too and find the same results as a robustness check. IB_PL is insignificant in Gov-only case in Column 3, since there borrowers get Forbearance by default from CARES Act. In the Non-Gov case, however, IB_PL is significant in Column 4, i.e., the borrowers have to actively communicate with the servicer to apply and be considered for forbearance. Similarly for OB_PL, columns 1-2 are significant. For Gov-only loans, OB_PL is significant as the servicer has to let the borrower know about the availability of forbearance option and have to ask the borrower if they want to enjoy the benefits delaying the payments in the future via forbearance. The servicers also respond to the borrower inquisitions in the situation where the borrower initiates the forbearance conversation. For Non-Gov loans, OB_PL is insignificant, however, I feel that this behavior can also be strategic/selective from the servicer's perspective based on the loan-type, i.e., Cash-Out Refinance or Purchase only loans. Cash-out loans comprise of almost all of the PL foreberance cases for non-Gov loans - 409 out of 461 in April 2020. This is why I separate the sample to only cash-out and no-cash-out in Table 21. IB_GOV and OB_GOV are insignificant without the interaction with the delinquency status of the loans. Curtailment of Income is an important determinant of Forbearance applications for the Liquidity-constrained borrowers. Unemployment Status per se is not as good of an indicator for Forbearance like Curtailment of Income.

Cash Out loans in this PE firm portfolio are distressed loans, which they bought at a significant discount in 2017. No promissory notes are attached to these Cash Out loans. These borrowers have lower FICO scores and are subprime. But from the time, they were acquired by the PE firm, they have mostly remained performing loans. Because the cash-out loans were bought at such discount and they comprise of subprime borrowers which mostly have been making payments, the focus has been less on re-couping corporate advances, since the servicer and the PE firm both share the profit in a liquidation/exit event. Because most Cash Out loans are non-Gov, the borrowers have to provide evidence to the servicer to get Forbearance approved. This is corroborated by positive significance of IB_PL. Both the OB_PL and OB_GOV are positive significant, indicating the servicer calls are significant for GOV PL borrower. On the other hand, the borrower calls significant when they have GOV loan.

Both the IB_GOV and IB_PL are positive but insignificant in Columns 2-4 with County fixed effects. This means, the borrower makes less calls when they have GOV and PL, as they are provisioned for forbearance under the CARES Act. OB_PL is positive significant in Column 2-3, since servicer makes mandatory calls in GOV loan case if the Borrower is PL. OB_PL is negative highly significant for Non-Gov loans (which are not Cash-Out), i.e., the servicer makes more calls for non-GOV borrowers to let them know they have to verify their unemployment status and financial hardship, otherwise they are ineligible for forbearance. For Govt.-backed loans, the CARES act legally binds the servicer to approve those borrowers and continue making advances. The servicer is legally obliged to advance for 4 months into Forbearance. From the data, I find that 70-75% are extending Forbearance. Hence, the servicer will have no incentive to reject Forbearance extensions as they have no skin in the game beyond 1 more month (first Forbearance was approved for 3 months).

2.6 Race Implications of Forbearance

I find evidence that African American borrowers have been able to avail the leniency from forbearance, but much less than the white borrowers. This is especially important in the current socio-economic and political climate charged with racial tension. The servicers have sophisticated Machine Learning models to profile the borrowers and hence, even without the race information, the Machine Learning models uncover structural relationships among Loan performance variables and/or triangulate from otherwise excluded characteristics Fuster et al. 2017, implying plausible statistical discrimination. However, the servicer do not have the race information per se and it is difficult, if not impossible to impute the race (or ethnicity) from the name of the borrower. [11] I argue that I demonstrate evidence of disparate impact and not disparate treatment. This still has a serious social cost and creates divide and inequality in the current socio-political and racially charged environment, exacerbated by the unforeseen pandemic ensuing from the novel COVID-19.

Among both conventional and FHA *performing* loans, the percentage of *Inbound* Forbearance applications is higher for White borrowers compared to the African American (henceforth Black) borrowers, by $(53.5\% - 29.41\%) \simeq 24\%$ and $(50.47\% - 48.10\%) \simeq 2.4\%$ respectively in Figure 29 and Table 30. Among both conventional and FHA *performing* loans, the percentage of *Outbound* Forbearance applications is higher for White borrowers compared to the African American (henceforth Black) borrowers, by $(40.54.5\% - 27.78\%) \simeq 13\%$ and $(45.00\% - 43.18\%) \simeq 2\%$ respectively in Figure 31 and Table 32. NPL loans do not mostly fall under the purview of forbearance and can be considered for foreclosure moratorium or loan modification depending on the loan balance and borrower past behavior.

I test the above observation in the model in Table 22, I add an indicator whether a borrower is Black (African American) and interact the dummy variable with IB, OB, IB_PL and OB_PL to make further inferences beyond Table 21. Firstly, for Non-Gov Black borrowers, there is a huge negative significance towards forbearance, with and without county fixed effect in Table 22. This clearly shows, all else equal, Black borrowers are prevented (unintentionally dissuaded from the race perspective) against availing forbearance applications because of adverse loan performance behavior, since the servicer does not have the race information of the borrowers. Also, for Non-Gov Black borrowers, IB_PL_Black is positive and significant with no fixed effects. With county fixed effects, the result is also economically significant, however the statistical significance is not bourne out in the county fixed effect due to small sample size when grouped by county.

[11] I am undertaking a new project in 2021 where I will obtain the confidential and proprietary audio files of some of these communications between some of these borrowers and the servicer legal representatives. The accent, tonal quality, emotional content, responsiveness, co-operation, willingness to negotiate, etc. can be captured in that paper. The idea is to design a mechanism through borrower responsiveness and co-operation, which is stricter than a carrot but more lenient than a stick.

OB_PL is still negative for Non-Gov loans with race in the specification, providing robustness of the selective behavior by the servicer, as detailed in the previous section Table 21. OB_Black for Non-Gov loans is highly positive and significant, with and without county fixed effects. This implies the servicers verify the Unemployment status or financial hardship for Black borrowers and then and then only are those Black borrowers are approved on their forbearance applications. OB_PL_Black is negative and significant without county fixed effects, which means Black borrowers who have performing loans are dissuaded by the servicers with their forbearance applications. Typically a borrower who has a performing loan ex-ante should not be in financial distress ex-ante. Essentially, the servicer is trying to reduce the ex-post risk for Black borrowers. If the Black borrowers are really in financial distress, they need to provide hard evidence to get their forbearance approved. On the flip side, these borrowers are not encouraged by the servicer to apply for forbearance and are mostly pre-empted by adding an income verification clause.

In it important to point out some nuances of columns (5), (6) in Table 22. Column (5) is without county fixed effect and I see warnings that some fitted probabilities of forbearance are numerically 0 or 1. This implies perfect predictability for a whole lot of people, specially in counties from which I have lower number of loans and hence lower forbearance application rates. This is plausibly related to the commonalities in a small county, i.e., if one person applies for forbearance, there everyone in the vicinity also does and vice versa. Also, the supply chains in smaller counties are heavily affected by the economic activities from big cities, which are mostly coastal in USA and hence affect heavily by COVID-19. The African American community amplifies these small county commonalities further more and hence their behavior is uniform and predictable from the same county. Also, smaller counties presumably have fewer industries and fewer number of jobs available, which skews the forbearance application behavior one way or the other. This is the reason, I evaluate forbearance rate with county fixed effect in column (6). This enables me to observe forbearance application and acceptance for African American people across county level heterogeneity, before and after CARES Act.

2.7 Diff-in-Diff for Borrower & Servicer behaviors

To address the endogeneity concerns from the strategic overcrowding of forbearance applications by Gov-backed performing loan borrowers and the endogeneity emanating from selective verification by the servicer, I use Difference-in-Differences approach. The treatment group comprises of the Govt.-backed loans and the control group contains the Conventional/Private Label Non-Govt. backed loans. The treatment time is end of March 2020, when the effect of CARES Act is internalized. In Table 23, the first variable is the interaction between treatment group and treatment time. In all columns (1)-(6), the interaction is statistically significant, implying significant causal impact of CARES Act on the treatment group. This clearly addresses the endogeneity concerns.

Columns (1) (without county fixed effect) and (2) (with county fixed effect) in Table 23 capture the Diff-in-Diff estimates for all loans, controlling for Inbound/Outbound calls and their interactions with Performing Loans and Gov dummies. The unique feature of this specification is the real-time ”Borrower Noted Unemployment” and ”Borrower Noted Income Curtailment” which gives me a unique chance to capture the borrower financial health. The strong positive significance of the interaction term provides evidence of the average positive treatment effect on the treated. Then I delve deeper to confirm the narrative of the paper. In columns (3) and (4), I explore the Diff-in-Diff approach to capture the strategic forbearance applications of the borrowers from their Inbound communications. The volume of Inbound communications after CARES Act increases the likelihood of forbearance applications. This overcrowds several worthy borrowers who really need the forbearance because of their financial hardship, e.g., borrowers in Las Vegas who are heavily impacted by the shutdown of the gaming industry cannot all avail forbearance, whereas, borrowers in mountain zone in Montana and other areas are applying for forbearance although the local economy was not affected by COVID-19 immediately in March/April.

Similarly, I conduct a Diff-in-Diff analysis in columns (5) (without county fixed effect) and (6) (with county fixed effect) to capture the selective verification of of borrowers by the servicers vis-a-vis Outbound communications. The robust positive significance of the interaction term again provides sound evidence to the narrative that the servicers try to pre-empt certain borrowers with Non-performing loans and other marginal borrowers from availing the forbearance options. Instead, the servicers encourage them for loan modification or foreclosure moratorium options.

For robustness checks, I also create Non-Gov dummy in Table 24 as the Outbound communications of the servicers for selective verification of unemployment status mostly targets the Non-Gov borrowers. I still find strong positive statistical significance for the interaction of term of the Treatment time and the treated (here treatment group in Non-Gov). Finally, because of multi-collinearity issues discusses in Section 2.6, I exclude those counties which have one or less loans for the whole time horizon of 5 months. Table 25 again shows strong statistical significance among counties with more than one loans for the interaction term. These results prove beyond any doubt the causal impact of strategic/opportunistic behavior on the borrower's part on forbearance selective verification by servicer on dissuading (denial with provision of alternatives like foreclosure moratorium or loan modification) forbearance applications.

2.8 Cost-benefit analysis of CARES Act

I conduct a simple Back-of-Envelope Calculation to give an estimate of the cost and benefit of the CARES Act. As of the end of May 2020, the US Residential Mortgage Debt was **$ 11.1 Trillion**[12]. Around 4 million Gov borrowers availed Forbearance by the end of May 2020[13]. Forbearances are approved for 6 months at a time. So the, next cost of forbearance for Gov-backed borrowers nationally is:

(Availed Forbearance) * (P&I) * (Forbearance period)/ (Principal Payment)
= 4,000,000 * $721.18 * 6 / 3 = **$ 5.76 Trillion**

Here, I use a mortgage calculator and use the terms of Gov-backed of mortgage from the

[12] Source:https://www.housingwire.com/articles/u-s-mortgage-debt-hits-a-record-15-8-trillion/: :text=Outstanding%20U.S.%20mortgage%20debt%20rose,according%20to%20the%20Federal%20Reserve

[13] Source:https://www.cnbc.com/2020/05/07/4-million-homeowners-in-cares-act-mortgage-forbearance-program.html

Summary Statistics Table 18 detailed in Figure 33 in Black Knight August 2020 report. The latest updates on forbearance and foreclosure moratorium are provided by Housingwire. [14]

2.9 Looming Plausible Housing Crisis

United States is still in the middle of the COVID-19 pandemic and no one can say for certain what the future holds. In these highly charged socio-economic and political circumstances charged with racial tensions, the possibility of an upcoming housing crisis cannot be overlooked. I take recourse to a recent article by Black Knight August Research[15]. Although the number of Forbearance applications have gone down in Figure 34 and weekly new forbearance plans have declined and flatted in Figure 35, there is still a sizable amount of forbearance plans extended in Figure 36. Moreover, the evictions of non-paying renters/homeowners is temporarily halted by President Trump at least till Jan 31st, 2021. As evident in Figure 37, the status of loans leaving COVID-related Forbearance plans are not current. After the forbearance period ends, they become delinquent loans, which can subsequently be offered foreclosure moratorium or loan modification to cure them. There could be a significant surge in foreclosure in the near future if these borrowers are able to avail foreclosure moratorium and get temporary relief from ban in eviction for renter and homeowners alike. FHA and VA have worse delinquency than other GSE and Private label loans. So, the marginal borrowers whose loans are about to become Non-performing can eventually end up in a vicious cycle of foreclosure and lack of employment which in turn increases foreclosure. So, a misaligned CARES Act to provide economic relief to the borrowers due to a public health crisis could be amplified into a much bigger housing crisis in the near future. Also, the exposure of 5.76 Trillion USD, as explained in the previous section could become permanent and it may take years, if not a decade, of slow economic recovery to reduce this national debt.

[14] Source: https://www.housingwire.com/articles/heres-what-mortgage-forbearance-looked-like-in-2020/

[15] Black Knight's August Report on Forbearance is provided here: https://cdn.blackknightinc.com/wp-content/uploads/2020/10/BKI$_{M}M_{A}ug2020_{R}eport.pdf$

2.10 Ongoing Research and Future Work

The effect of securitization on loan modifications (or lack thereof) has received a great deal of attention among academics and policymakers. Piskorski, Seru, and Vig 2010 find securitization induce a foreclosure bias. Agarwal et al. 2011a and **Kruger2018** find securitization reduces the likelihood of renegotiation and increases the likelihood of foreclosure. One critical assumption in the extant literature is that the quality of the portfolio and securitized loans are similar. The literature relies on hard information at the time of origination to control for loan quality despite the fact that it is a time-varying attribute. This may be a concern if portfolio borrowers and securitized borrowers had similar credit scores and incomes at origination, but securitized borrowers were more likely to lose their job. The change in employment status would partially explain the higher likelihood of foreclosure that was mistakenly attributed to the securitization process. I address this issue by quantifying and incorporating soft information in Chapter 4 from servicer call logs. An even more acute consequence of another future paper is to run a horserace between different loan dispositions of forbearance, foreclosure moratorium, loan modification and possibly partial loan forgiveness, conditional on securitization in the upcoming democratic Biden administration.

2.11 Conclusion

In this paper, I provide insights on the lack of effectiveness of the mortgage forbearance program contained in the CARES Act. I demonstrate that the CARES Act created a huge untenable exposure for the government in 2021 and this dollar cost and social cost is higher than the current heated debate on handing out money directly to households. Utilizing a novel administrative dataset coming from a mortgage servicer, I examine the communications between borrower and servicer in order to shed light on the probability that a borrower will request forbearance. In line with studies looking at mortgage modifications during the Great Financial Crisis period Demyanyk and Van Hemert 2011, Mayer et al. 2014, I provide evidence to suggest that borrowers are strategically taking advantage of the CARES Act to request mortgage forbearance, the servicers are selectively verifying the unemployment status/financial hardship for Non-Gov borrowers. I find that this selective verification by the servicer precludes unintended distributional implications based on race for African American and Hispanic borrowers. My results document strategic behavior in response to CARES Act policy which was an ad-hoc response during the advent of COVID-19 in the United States. The economic costs of strategic behavior/selective verification are significantly large relative to the potential gains to borrowers, lenders, and servicers from these policies. My results highlight the misalignment of ad-hoc policies on Government programs such FHA, VA, USDA of the HUD, due to the non-verification (for Gov borrowers) and selective verification (Non-Gov borrowers) of unemployment status/financial hardship. I am able to conduct this research due to the narrative retrieval apparatus I have created from the novel administrative data on the communication between the borrower and the servicer. The flawed interpretations and the erroneous inference thereof of the special servicers can be mitigated using Machine Learning/Natural Language Processing techniques. More work must be done to assess the overall costs/benefits of such forbearance policies and their effectiveness in preventing foreclosures, or else, surge in foreclosures will inevitably lead to the next housing crisis.

CHAPTER 3
QUANTIFYING SOFT INFORMATION, MORTGAGE MARKET EFFICIENCY & ASSET PRICING IMPLICATIONS

I provide a novel framework for machine learning models to ingest quantified soft information during the life of a loan, using cutting-edge natural language processing techniques on salient unstructured text. This soft information, from servicer call transcripts, is not restricted to mere positive/negative sentiments and provides efficiency and alleviates the information asymmetry between the lender (and/or issuer) and the borrower. Proprietary servicer comments are hardly accessible and offer the soft information for real-time delinquency status of the mortgages. I investigate whether the special servicer invoked by the investor can utilize the valuable comments from the master servicer. The time-varying soft information about the borrower's financial condition, health of the loan and the property condition from these master servicer comments renders the predictive power and has asset pricing implications. Given this valuable information, the special servicer may choose to use this information, as I anecdotally see with several private equity investors. The well-known unresolved conflict of interest between the master and special servicers can be resolved, thereby reducing moral hazard and increasing efficiency and transparency.

3.1 Introduction

One of the key functions of the fixed-income (especially mortgages) finance profession is measuring the delinquency status of a loan (mortgage) and developing quantitative models that predict the distribution of future outcomes. This enables lenders, consumers, investors, and policymakers with the information on the structural context needed to allocate resources efficiently. Banks have a repository of information about borrowers creditworthiness, collected over time through frequent and personal contacts (*relationship banking*) between the prospective borrower and the loan officer. Banks have also been using available public records with the advent of technology.

This private/soft information is valuable to the banks for its lending decisions during underwriting. However, the soft information gets blurred during the surveillance process after the loan has been issued, since the financial or personal situation of the borrower may change over time. My novel proprietary data allows me to quantify the time-varying soft information[1] about borrowers, previously inaccessible to academia. I provide the first quantitative measures of this soft information, beyond positive/negative sentiments, about the borrowers, so that this information can be transmitted during the securitization process and there is no information loss during securitization as pointed out by DeMarzo2004. [2]

The situation is substantially more complicated for securitized loans. Figure 38 describes how soft information can be used to gain insights about the changes in borrower financial condition and life events in real time and utilize the information for improving loan performance predictions through the feedback loop enabled by servicer call transcripts. MBS (Mortgage-backed securities) securitizations are not balance-sheet transactions and are done at arms length via an SPV. [3] Going forward, the delinquent loans can be kicked off (bought out by the issuer) from the pool via *Early-Buy-Out* (EBO) [4] or *Reps and warranties*. Otherwise, the bank can choose to do nothing, in which case these delinquent loans become limbo loans. Inaction on limbo loans

[1] Soft information goes beyond the one-dimensional measure of sentiment. Soft information is difficult to transmit and quantify (mostly unobservable to the issuer); however, I can extract soft information from the call transcripts (text data) between the borrower and the servicer. The context under which soft information is collected and the collector of the information are intertwined. Natural Language Processing (NLP) renders a narrative retrieval device to capture the soft information along with the context. Hard information is quantitative, easy to store and distribute in impersonal ways, and its information content is independent of its collection methodology. Hard information has been used historically to predict mortgage loan delinquency status even in the latest cutting-edge models. I predict ex-post default rate more accurately and hence conclude soft information along with hard information is critical in identifying the health of loans.

[2] DeMarzo2004 argues that pooling of assets (loans) has an information destruction effect for the intermediary that prevents her from fully exploiting individual asset information. Asset servicers are employed to monitor information about the assets, to ensure that timely performance of loan payments and to provide confidence to investors.

[3] During the ramp up period, an MBS manager invests the proceeds of the issuance into assets. In addition to the ramp-up period, the typical MBS has a warehousing period before the securities issuance when the sponsor starts collecting collateral. There is a sample selection (around 10%) which the sponsor removes from the MBS and keeps the remaining for due diligence. Some delinquent loans are kept in a pool, and the banks hire special servicers to monitor them.

[4] When Ginnie Mae loans become close to 90 days delinquent, the issuer has teh right to buy them out early from the Ginnie pool and sell them in the secondary market to Private Equity, who bid for these loans and buy them at a discount. This has provided a new way of reducing the skin in the game for the issuer after the implementation of *Risk Retention* rule from Dodd-Frank Act from early 2017.

(see LindaTangJRERLimbo2015) has reduced dramatically after change in securitization structure from Risk Retention rule in 2017 (see GhentRR). This may further be exacerbated by exogenous changes, e.g., regulations regarding Ginnie Mae (VA/FHA/USDA) claims process and/or Comprehensive Capital Adequacy and Review (CCAR) regulations in the last 3-4 years. In the current COVID-19 pandemic, I use soft information and address the choice between forbearance, foreclosure moratorium, loan modification and loan forgiveness, conditional on securitization in a future paper.

I claim that the aforementioned quantified soft information can resolve the conflict of interest between the master and special servicers (see Mayer and Gan 2006). This can have an unforeseen reduction of moral hazard in the opaque residential real estate market, thereby increasing efficiency and transparency. The servicers (see Ambrose, Sanders, and Yavas 2008)[5] collect and document monthly loans payments, manage escrow accounts and monitor the underlying properties' condition. These master servicers prepare loan payment reports for the trustee and transfer collected funds to the trustee for payment to investors. The special servicer come into the picture for a 60 days past due loan, and she can modify the terms of the loan via loan extension or loan restructuring or foreclose on the loan and sell the underlying collateral.

The special servicer is responsible for work-out of the loans forwarded by the master servicer, but is contractually obligated to maximize the interests of the investors. The optimal solution would be for the special servicer to cure the the loan to performing status.[6] The special servicer is not concerned with the borrower's position (preempting the borrower to negotiate an outcome other than foreclosure), but rather may undertake actions (modification, foreclosure, etc.) that maximize the position of the first-loss investor and guarantee the timely cash flow payments to the senior investor. This mechanism contrasts directly with the role of a master servicer (the administration

[5] ”There are three types of servicers. The subservicer is typically the loan originator in a conduit deal who has decided to sell the loan but retain the servicing rights. The subservicer sends all payments and property information to the master servicer. The master servicer oversees the deal and ensures the servicing arrangements are maintained. In addition, the master servicer facilitates the timely payment of interest and principal to the investor. When a loan goes into default, the master servicer has the responsibility to provide for servicing advances.”

[6] Special servicers are generally compensated by a percentage of the outstanding balance of the loans that they serve plus a fixed fee. Unlike the master servicer, the special servicer generates more profit if a particular loan goes into default.

of troubled loans originated and retained by traditional lenders).

Generally, a master servicer has a management system (Ops), which typically manages these delinquent loans. They also have a capital market (CM) function. CM and Ops groups have different priorities. The objectives of CM include NPV based advances, managing capital constraints, etc. which may not be in line with the primary objective of monitoring delinquent loans for which the servicer (Ops) was hired by the bank, to begin with. The loans for which escrow advances have been made and about two-thirds of the Corporate Advances have been identified as deadweight losses are the candidates for *buy-out* from the issuer. This is a new mechanism that the issuers and lenders are plausibly using to avoid *risk-retention* for the adverse loans in the MBS. On the other hand, the legal servicer comments, used in this paper, are from the Ops, which has valuable *soft* information both about the borrower life events and financial condition & the property condition.

Hard information post-*origination* and definitely post-*securitization* about the borrowers' real-time financial plight and property conditions are very difficult for lenders to obtain due to privacy laws in place. The number of times the borrowers call themselves on the one hand, and monitoring by the legal representatives on the other, depends greatly on the individual borrower and her situation. The representative has to give the mini-miranda[7] in accordance with FDCPA[8] (Fair Debt Collection Practices Act) and confirm certain account specifics to ensure it is the borrower to whom they are speaking. Then the representative will try to determine if the borrower wants to keep the house or liquidate, the reason for default as well as occupancy of the property, if not already known. Hence, I explore soft information to our advantage from servicer comments and gain unforeseen insights.

To give an idea of the economic significance, I provide the size of EBO deals from January-June of 2020 (see Figure 39) in each delinquency buckets for the largest 4 issuers for this Private

[7] At the beginning of a collection call, a debt collector must recite wording that has come to be called the mini-Miranda disclosure. It informs the consumer that the call is from a debt collector, that they are calling to collect a debt, and that any information revealed in the call will be used to collect that debt. The disclosure must also be included in written correspondence with consumers, such as a collection letter. This important provision is required by the Fair Debt Collection Practices Act to prevent debt collectors from calling under false pretenses and gaining information from consumers that can later be used against them. *Source:* https://www.consumer-action.org/helpdesk/articles/mini_miranda_disclosure

[8] https://www.federalreserve.gov/boarddocs/supmanual/cch/fairdebt.pdf

Equity firm, which has around 15% market share of the national EBO market in the United States. The sheer volume of delinquencies creates capacity constraints and necessitates the use of ML to operate nimbly with less manpower. The marginal cost of quantifying and ingesting soft information is much less than hiring more lawyers (who bill by the hour) and unnecessary conflict of interest between master and special servicer.

To extract soft information, I utilize the *legal abbreviations* that are internal codes for the servicer to keep track of the monitoring process and keeping the comments, either, when the borrowers are calling themselves to report material changes in their professional life (e.g., loss of job) or personal life (e.g., birth of a baby, death of a significant household member, divorce, etc.) or, when the servicers are really monitoring the borrower and asking detailed and decisive questions. The *intent* of the servicer can be captured by the proportion of the valuable comments to the total number of times a borrower is called after their mortgage becomes non-performing.

To evaluate the cumulative effect of soft information throughout the month, I define delinquency classes as loan outcomes at the end of each month and find the cumulative effect of soft information throughout the month. The soft information is obtained from the time of acquisition of these loans by the Private Equity firm. These loans come in the form of Early-Buyout deals, of which 60% loans get re-pooled to a new GNMA pool within 3-6 months, providing the value of quantified soft information is prediction accuracy. The remaining 40% remain in the portfolio of the PE firm which are monitored via the surveillance process vis-a-vis hard information and quantified soft information (see Figure 40). I keep two delinquency classes **PL** (Performing Loan) and **NPL** (Non-Performing Loan) and two liquidated (terminal classes/absorbing states) classes **PIF** (Paid-in-Full) and **ShortSale** to aggregate the transitions into meaningful buckets.

I conduct robustness checks for a larger and more *granular* set of delinquency classes **B120D** (Beyond 120 days), **BK** (Bankruptcy), **FC** (Foreclosure), **PIF** (Paid-in-Full), **REO** (Real-Estate owned), **ShrtSal** (Short Sale), **W0_30D** (Within 0 to 30 days), **W30_60D** (Within 30 to 60 days), **W60_90D** (Within 60 to 90 days), **W90_120D** (Within 90-120 days) and document the impact of quantified soft information on asset prices and their estimated volatilities.

I quantify soft information across 8-12 different aspects and severity of delinquency. This soft information has differential time-varying impact for different delinquency classes, which cannot be captured by frequently used measures of *sentiment* used previously in the literature. Sentiment, a measure of investor beliefs, has been viewed as a determinant of the variation in asset prices (Keynes 1937). The plausible role of sentiment is especially important for the housing market, where financial errors can have very costly as beliefs are unobservable and difficult to quantify (Soo 2018). The usual proxies of sentiment used in the stocks, mutual fund flows, dividend premiums, etc. are unfortunately unavailable for the housing market (Baker and Wurgler 2006). Over two-thirds of US households invest the majority of their portfolio in real estate via home ownership (Chan, Schneider, and Tracy 1999). Survey measures during the 2008 financial crisis are limited in capturing the spatial variation, as their scope is not national.

Soft information has been explored in the literature in several contexts. Agarwal and Ben-David 2018 find loan officers put emphasis on hard information in approval decisions and henced the loan-prospecting incentivizes overlooking of unfavorable soft information. Agarwal et al. 2011b investigate soft information from revealed risk tolerance from the borrower's choice of credit contract which revealing her risk level. Demers and Vega 2008 use text from management's quarterly earnings to extract soft information on managerial net optimism. They test both cases: informativeness of contemporaneous hard information and soft information per se. D'Aurizio, Oliviero, and Romano 2015 use soft information in bank lending decisions focusing on firm heterogeneity in corporate ownership structure.

Faure-Grimaud, Laffont, and Martimort 2003 provide evidence that soft information helps the principal even in the event of collusion of the supervisor and agent under asymmetric information. An et al. 2015 show that soft information is more evident in low-competition markets and among borrowers with low credit score, while adverse selection is more applicable to liquidity-constrained borrowers and loans originated by brokers. Elul 2011 find evidence using soft information that privately securitized loans perform worse than observably equivalent portfolio loans due to adverse selection. Campbell, Loumioti, and Wittenberg-Moerman 2019 explore whether non-agency-related costs and behavioral biases come in the way of the effective internalization of soft information. Ergungor and Moulton 2014 evaluate soft information on loan performance of low-income cohort and provide adequate explanation for observed performance differences between bank and nonbank lenders.

Machine learning is recently used in real estate literature. I study deep neural network (DNN) for disentangling liquidity constrained and strategic default in the first chapter. Ambrose et al. 2019 uses machine learning methodology to assess the completeness of PSAs for conduit CMBS deals. To the best of our knowledge, there has not been any machine learning application on proprietary servicer comments data. My paper uses cutting edge Natural Language Processing (NLP) techniques to alleviate the well-documented conflict between the master servicers and special servicers. Jiang et al. 2018 predict whether a borrower default is significantly concerning online peer-to-peer (P2P) lending.

The rest of the paper is structured as follows. Section 3.2 briefly discusses the data set that I use. I plan to attach much more details about the data in the online appendix later. Section 3.3 describes our model structure and estimation. I detail a lot of our analysis in Subsection 3.3.1 using the frequency-based Bag of Words Models, Subsection 3.3.2 using semi-supervised Word-to-Vector Models and Subsection 3.3.4 using Topic Models. I then provide evidence for the incremental value in market efficiency and resolution of the conflict of interest between Master and Special servicers from out-of-sample predictions for four delinquency states in Section 3.4 and then for more granular 10 delinquency states in Section 3.5. Section **??** provides some asset pricing implications of quantifying soft information. Section 4.4 provides significant evidence of the usefulness and ubiquity of the quantified soft information in relation to the COVID-19 pandemic and captures the consequent borrower financial hardship. Section 3.6 concludes.

3.2 Data

I have daily proprietary novel data on *servicer comments* from November, 2017 to December, 2019 from servicers of bulge bracket banks for 44,182 loans. Overall, I have 14,000,575 pooled observations when I stack all the comments for all the loans in our data. Once I clean the 'NULL' and other meaningless messages, I analyze 12,567,500 messages using cutting-edge Natural Language Processing (NLP) techniques in GPU/Clouds. I extract soft information about the borrowers's financial condition and life events & property condition from this test data and try to prove predictive power for borrower behavior beyond what is observed in the loan performance data.

The special servicer, who provided me this proprietary data, has a national presence and has a big market share among the servicers in residential mortgages in the United States.[9] A private equity firm purchases these loans in a joint-venture with the servicer and gets access to the master servicer call transcripts with the borrowers. The servicer legal representative uses his/her own abbreviations and phrases to enter the results of communication with the borrower. Hence, there is a lot of noise in the call transcript data. To quantify soft information from this unstructured data, I resort to novel NLP techniques. This paper is the first attempt to use these time-varying soft information about the borrowers in addition to loan performance variables. In fact, NLP helps me structure this data into a tabular format, so that it can be ingested by classification models along with other *hard* loan performance variables, detailed in Table 26.

Vanilla structural models cannot retrieve the narrative of the conversation between the master servicer and the borrower. There are several aspects of borrower financial health, e.g., unemployment status, furlough, impact of COVID-19, etc. which are highly uncertain in the current turbulent times. Similarly, life events, e.g., death, divorce, marriage, relocation due to job change, etc. which have historically not captured in the hard information due to privacy laws in place. A simple upward or downward market sentiment is hence unable to capture these nuances. I explain in great details the different cases of these soft information in Subsection 3.3.2. The primary purpose to significantly improve the prediction accuracy of the adverse delinquency states, e.g., foreclosure, real-estate owned, shortsale, bankruptcy, etc. This improves the mortgage market efficiency tremendously and mitigates the well-known conflict of interest between the master servicer and the special servicer. These soft information does internalize lenders' expectations about future

[9] The entire list of servicers is not public information. The GSEs may have a listing of their servicers, but that would be for a subset of the loans.

payments by the borrowers, however I only obtain the call transcripts in this research, which is the servicer's summarization of the dialogue. I intend to delve deep into the incentive compatibilities in a future paper using the audio conversation between the borrower and the servicer.

To be consistent with prior literature, I include residential mortgage loan-specific attributes from our proprietary data source and yield curve, mortgage-treasury rate spread, 10 year risk-free rate volatility, region dummy, seasonal/quarter dummy, among others. I include local macro-variables to capture the channel of potential contagion in which defaults in the same state might amplify the barely surviving loans. Such feedback mechanisms are well-documented by Agarwal, Ambrose, and Yildirim 2015, Anenberg and Kung 2014, Harding, Rosenblatt, and Yao 2009, Lin, Rosenblatt, and Yao 2009, Towe and Lawley 2013, and among others.

3.3 Methodology and Techniques

I use several cutting-edge Natural Language Processing (NLP) techniques to quantify soft information. Bag of Words is the first approach I try to visualize the clusters (groups of similar words) of keywords. This does not alleviate the sparsity issue [10] when the words are converted to vectors. Once I have the word clusters, I can use the word-to-vector technique to convert words to easily computable vectors, which the computer can understand. Finally, I define topics and themes from these words to capture the sentiment of these vectors. I intend to have a continuous measure of sentiment and can create an index from the sentiment measure. This decision is based on the number of distinct topics that emanate from the topic model. I can use the sentiment index as a control variable along with loan performance attributes to gain superior prediction of future borrower behavior.

I exclude the servicer comments with gibberish (no English meaning) and **focus** on a set of targeted keywords that have been historically categorized by the Private Equity firm who provided me the data. I need these words to compare the context of lemmatized tokens from the servicer

[10] The text is vectorized into counts i.e. each sentence is represented by a vector with the dimensions of size of the vocabulary of the corpus (say $|V|$), and the value of each dimension being the number of occurrences of that word in the sentence. This enables a tabular representation of the document (commonly known as the term-document matrix), with the rows representing sentences and the columns the vectorized word counts.

comments data via a supervised word-to-vector algorithm. Of course, this limits our analysis to sentiment related to these historical keywords, but they are very comprehensive, listed in Table 50.

3.3.1 Bag of Words

Bag of Words is an NLP technique that extracts features and meaning from documents. Numerous linguistic techniques assume independence, implying that the words are contextually insignificant and order-agnostic.

While the sequence of appearance of the words themselves aren't as important as their existence (meaning), counting the individual occurrence allows for a determination of the degree of importance or significance. Some of the well-known parsing techniques that fall under the bag-of-words methodology include: 1) targeted phrases 2) word lists 3) Naive Bayes, and 4) thematic document structure. Targeted phrases (Loughran, McDonald, and Yun 2009) seek to identify particular words or phrases that are highly significant with respect to the context of the document. Word lists, as the name suggests, refers to a compilation of words that possess a certain sentiment (for example, positive or negative). The relative frequencies of words that possess a certain sentiment direct the sentiment of the overall document. Naive Bayes is a well-established supervised learning algorithm to classify sentences into sentiments or classes based on Bayes' rule. Antweiler and Frank 2004, Das and Chen 2007.

While these models are simple, efficient, and robust, they have some drawbacks. Since words are treated as stand-alone objects, their ordering is completely ignored, and in turn, so are their semantic relationships with each other. As the size of the corpus increases, it is highly likely the vocabulary ($|V|$) scales up and along with it, the size of the sparse term-document matrix. This large-scale sparsity can cause computational challenges in further processing. These shortcomings necessitate more computationally efficient methods that capture the relationships between the words.

Hence, I follow the steps below to clean the servicer comments:

1. **Splitting:** I use deconstruction (see Whitelaw et al. 2009) to remove to conjoined words. As

a lot of the sentences are conjoined, the keywords are mixed with other words, e.g., *cramdownacknowledge* or *nocramdown*. Without split, one loses the keyword, from the English vocabulary perspective. So, a probabilistic split (package) is run that splits *nocramdown* into *no* and *cramdown*. The misspelled words like *agnecy* will unfortunately be split into a random string of characters like *agn* and *ecy*, because on its own it is not meaningful as an English word. *ins* gets split into 'in' and 's', but *ins* means insurance in our data. So, *ins*, *fc* (foreclosure), *bk* (bankruptcy), *reo* (real-estate owned) are kept as is.

2. **Lemmatization:** Lemmatization just brings the word to its base form. Since a lot of words are not meaningful English words, lemmatization will leave them as is. To clarify how difficult it is to lemmatize the data at hand, I provide one example, where there are random letters in the middle, e.g., *"DSPN=209 09/26/2016 Fire AT/NEAR FIRE NO UTLS;LISTENED TO PREV TAGED C WITH 09/26/2016 Fire AGNECY;AGT ONLY TRIED TO FIND POL BY NAME;SENT FAX REQT 09/26/2016 Fire FOR POL INFO"* becomes *"dspn = 209 09/26/2016 fire fire no utl listen prev tag c 09/26/2016 fire agn ecy agt only try find pol name send fax req 09/26/2016 fire pol info"* after lemmatization, removing prepositions and stopwords.

3. **Vocabulary:** After a careful combination of the above steps, I remove the non-English words, except a few from our list of keywords, from all the servicer comments. This helps remove the unnecessary characters which provide no information in this research question. As an example, the above lemmatize sentence becomes: *"fire fire no listen fire only try find name send fax fire info"*.

There is a tradeoff between splitting and lemmatization, e.g., without splitting the sentences are not meaningful. With the split, one can remove non-English words like 'in' and 's', but one would lose the meaning of the word 'in' and hence the context in the sentence. This is primarily because the legal comments (abbreviations) are all written in shorthand (unfortunately, this is the industry standard among servicer legal representatives) and combined in a jumbled-up form. The following Dendogram (hierarchical agglomerative clustering Murtagh and Legendre 2014) in Figure 41 captures the hierarchical structure among the keywords and renders a clear indication of the heterogeneity in keywords' thematic similarity.

3.3.2 Word to vector

A more sophisticated technique compared to Bag of Words involves converting words (or phrases) themselves into their vector representations, known as word (or phrase) embeddings. The main goal of vectorization is to capture the similarity between words using their vector representations, which is not existent in simple models like bag of words that treat words as atomic units. The embeddings capture semantic meaning when trained on text corpus, using word2vec model. Skip-Gram model learns word embeddings in semi-supervised setting, since there are no direct labels associated with the words, but the neighboring words (the context word in a sentence) provides a proxy for the labels.

These distributed word representations [11] have been studied in depth in the past (Rumelhart, Hinton, and Williams 1986). To overcome the curse of dimensionality suffered by older models, neural networks were proposed to represent words as high-dimensional real valued vectors (Bengio et al. 2003; Mikolov et al. 2010). The goal of using neural networks is to potentially determine a probability distribution for the next word or phrase, based on the previous words (context). The choice of neural networks was found to be encouraging as the vectors thus generated were known to capture many syntactic and semantic relationships in a meaningful manner - for ex: *vector("King")* - *vector("Man")* + *Vector("Woman")* yields vector that is nearly identical to *vector("Queen")* (Mikolov, Yih, and Zweig 2013).

First, I use RAKE (Rapid Automatic Keyword Extraction), which uses stopwords and phrase-

[11] The two main neural network architectures that are used for distributed word representations are Feedforward Neural Net Language Models (NNLM) and Recurrent Neural Net Language Models (RNNLM). While NNLMs are highly effective in overcoming the curse of dimensionality suffered by statistical language modeling, they have two drawbacks: they are computationally complex due to the presence of a projection layer (resulting in dense matrix multiplications); in addition to the input, hidden and output layers, and the number of preceding words (context) have to be fixed beforehand. To overcome this, Mikolov et al. 2010 proposed a RNNLM which does not have a projection layer, reducing computational complexity. Furthermore, the hidden layer is linked to itself, allowing for some short-term memory using information from the past and thus removing the need for an ad-hoc context length.

delimiters to detect the most relevant words/token in a comment. For the sentiment classification part, a word2vec model converts each lemmatized word from the corpus into a high dimensional vector. Since I know the vector representation of each word in the servicer comments, I have to ”combine” these vectors together and get a new one that represents the comment as a whole. Since averaging the word vectors together is equal-weighted, I compute a weighted average where each weight gives the importance of the word with respect to the corpus. The tf-idf score (term frequency-inverse document frequency) acts as a weighting scheme that extracts the relevant words in a comment.

$$tfidf(t,c) = tf(t,c).idf(t) \tag{3.1}$$

where tf(t,c) is the term frequency of the token t in the comment c. Also, the $idf(t)$ is the inverse document frequency of the term t, computed by this formula:

$$idf(t) = log(1 + \frac{1 + n_c}{1 + df(c,t)}) \tag{3.2}$$

where n_c is the number of comments and df(c,t) is the number of comments containing token t.

The idea is to find clusters of similar words to our pre-specified keywords (chosen from qualitative judgement and historical perspective by the Private Equity firm who provided methe data). To expand the list of keywords, I find similar words to our keywords in the following way. The idea is to convert these similar words to vector and quantify higher-dimensional soft information (beyond sentiment).

3.3.3 Visualization of clusters of similar words

To achieve this objective, I first plot the main adverse delinquency states, namely, ”delinquent”, ”bankruptcy”, ”foreclosure”, ”reo”, ”short sale”, using T-SNE, which is just a two-dimensional visualization of clusters of similar words, with the axes scaled appropriately to fit important similar words. [12], in Figure 42. The axes do not have any units or physical significance and is chosen automatically by the T-SNE algorithm to appropriately fit the important similar words in one diagram. If I start from the south-east corner of Figure 42, I find the blue cluster of words similar to ”delinquent”. These words could signify any number of days of delinquency from 30 days to 120 days of late payment. Next to the ”delinquent” cluster, I find the ”reo” cluster in peach color. The word ”fcl” defines a boundary of ”reo” delinquency state with the ”foreclosure” state, since most of the time, a Foreclosure process precedes an REO proceeding. To the north of peach color ”reo” cluster, I find the red color ”short sale” cluster, which is a different kind of outcome for severely delinquent loan. The one red spot for ”forebear” signifies a specific kind of ”short sale” where the lender shows forbearance and takes a loss and the borrower gets relief from the put option by surrendering the property in non-recourse states. The light green color for the ”foreclosure” cluster is at the center and is a logical state for severely delinquent loans. The ”foreclosure” state could precede bankruptcy proceedings and generally precedes short-sale or REO. The violet color ”bankruptcy” cluster is a transient delinquency state. It can happen when the borrower voluntarily files for Bankruptcy proceedings in Chapter 13 and tries to negotiate the terms of the loan with the lender. The bankruptcy proceedings from Chapter 7 are more of a natural continuation of the foreclosure proceedings. Hence, there is a boundary between ”foreclosure” and ”bankruptcy” clusters, where I can see the word ”bankruptcy” in the ”foreclosure” cluster and similarly the word ”foreclosure” in the ”bankruptcy”.

Next, I investigate and report the T-SNE results for several key events that happen during the lifecycle of a loan and extract intricate relationships and interactions with the main 5 delinquency states described above. I add each category on top of the main 5 delinquency states to visualize how each category results in one of these 5 delinquency states.

As is evident in Figure 43, the ”legal” cluster is mostly intertwined with the ”foreclosure”

[12] T-SNE is a machine learning graphing tool known as a stochastic nearest neighbor embedding van der Maaten and Hinton 2008) which helps me visualize clusters in two-dimensional space. In a topic model, servicer comments have a mixture of themes, which can be a narrative snapshot of the state of the loan during it’s lifecycle. The t-SNE plots help digest the high-dimensionality and complexity of servicer comments, illustrating their intuitive underlying structure as recurrent co-location of topics.

cluster. There are two types of legal proceedings, one related to the title-issue from improper or missing promissory note for the residential mortgage. The other ”legal” sub-cluster refers to the court proceedings related to the foreclosure process.

Figure 44 describes very crucial life events for the borrower, which act as triggers for worsening delinquency state. Some of the key life-changing events are ”death”, ”divorce” and ”unemployment”. On the south-east and north-west corners of the Figure 44, I find two sub-clusters of ”unemployment” in red. The north-west sub-cluster of ”unemployment” has an extended boundary line with the blue color ”delinquent” cluster. This can be due to sudden loss of job for a household depending on employment income or deteriorating conditions for small and medium business owners. The south-east red sub-cluster for ”unemployment” indicates those borrowers who self-report their employment status with the lender and they are able to negotiate the terms of the loan, e.g., mortgage rate, duration, etc. and hence avoid the more severe delinquency states. On the north-east corner of the Figure 44, I find the clusters for ”death” and ”divorce”, which have an extended boundary, which indicates re-marriage due to death of the significant other. I also find some intersection of the south-east sub-cluster of ”unemployment” for borrowers who negotiate the terms of the loan with the lender. It is highly likely that the same borrowers also head into divorce proceedings with their significant other because of the tight financial condition due to loss of employment. One peach dot of ”divorce” is between ”short sale” and ”reo” clusters at the center of the Figure 44, as a sudden flight to those severe delinquency states, can result from divorce proceedings. As usual, on the south-west corner, I see some overlap between the ”bankruptcy” and ”foreclosure” clusters.

In the south of the Figure 45, I see that the ”military” cluster in red color has a boundary with almost all the key delinquency states. This is because military personnel are looked at more leniently by lenders and these loans can enter these delinquency states and sometimes come back to a current/non-delinquent state. Also, I have VA loans in our data and hence veterans, who are viewed leniently by the lenders and are backed by HUD programs. Here, the lenders go through a claims process in case the veterans or military personnel default.

In Figure 46, I see the red cluster of borrowers who have experienced natural disasters. I have removed the cases of natural disasters where the borrowers have fire and/or flood insurance and other forms of insurance. These loans sometimes become delinquent due to a few months of missed payments in case the borrower gets the insurance money and repairs the home. If the home is severely damaged, then REO is a natural outcome for these homes, as is indicated by the intersection with the light green ”reo” cluster.

Now, I focus on the cases where delinquency behavior is observed due to ”Occupancy”-related circumstances in Figure 47. We try to capture all ”Occupancy”-related circumstances with the keywords ”abandon” and ”vacant”. The peach color cluster for ”abandon” at the center of the Figure 47 obviously has some intersection with the red cluster ”vacant” as they fall under the same umbrella of ”Occupancy”-related circumstances. I see a lot of duplexes being abandoned due to the cost burden from maintenance. A house could be abandoned due to a whole host of factors, e.g., the house has become uninhabitable or is irreparable, a trespasser is found more likely near an abandoned house, due to punitive legal proceedings, a house-owner can abandon the house. On peach dot in the ”bankruptcy” cluster signifies voluntary bankruptcy filed by certain borrowers who may have sinister motives to avoid the debt-service obligation. It is more likely that such a voluntary applicant of bankruptcy can very well abandon the property. One peach dot in the ”delinquent” cluster could be an indication of future voluntary bankruptcy petition. The ”vacant” cluster is much broader. Of course, the ”vacant” cluster has some intersection with the ”delinquent” cluster as ”occupancy” (decision to occupy or live in) of the house where the borrower is delinquent in the debt-service is a questionable choice of the borrower household. I see some red dots in the center of the Figure 47 having an intersection with the light green ”reo” cluster as it is the natural choice for the lender for ”vacant” houses.

Figure 48 refers to keywords ”condemn”, ”damage”, ”hazard/loss” which are related to the *property condition*. Condemnation is an outcome due to property condition. All of these keywords have intersection with the ”reo” cluster as this is the natural outcome for houses with Property Condition related to ”condemn”, ”damage”, ”hazard/loss”.

Figure 49 lists keywords ”contest”, ”lien strip”, ”title issue”, ”repurchase” related to the Title of the promissory note related to the mortgage and ownership of the residential property. The peach color ”lien strip” has heavy intersection with the violet ”bankruptcy” cluster, which could be an outcome of the involuntary bankruptcy proceedings following foreclosure petition by the lender. One light blue dot also defines ”foreclosure” as the channel for ”lien strip” leading to the bankruptcy process. The ”title issue” and ”contest” are intertwined due to obvious reasons. The ”contest” cluster is closer to ”foreclosure” and hence ”contest” can be used as a leading indicator for foreclosure proceedings. The red color ”repurchase” cluster has a broad intersection with several clusters. Repurchase can first be can a natural outcome of ”lien strip”. A repurchase is also associated with an REO proceeding by the lender. One red dot in the light green ”short sale” cluster indicates ”repurchase” from short-sale.

3.3.4 Topic Models

Bybee et al. 2020 have proposed a novel approach of estimating the state of the entire economy via textual analysis of news from the Wall Street Journal. They use a topic model condensing news into interpretable themes and also quantifies the time-varying proportion of attention to each theme. Topic models are a popular dimension reduction technique from the fields of unsupervised machine learning and natural language processing. They have two essential elements. Just as principal component analysis condenses large data matrices into a comparatively low number of common factors, a topic model's first element reduces an inherently ultra-high dimension representation of a text corpus into a relatively low-dimensional set of common ”topics”. The formation of topics is unsupervised; they are estimated as clusters of terms that are most likely to co-occur in the same article. Those clusters are optimized so that relatively few clusters (many fewer than the number of distinct terms in the data set) preserve as much of the meaning in the original corpus as possible, by best explaining the variation in term usage across articles.

A topic model estimates the proportion of text relevant to each topic. This makes it possible to analyze the interaction between topics, categories, and delinquency classes. The topic model as a narrative (servicer comments) retrieval device which providers a nuanced verbal interpretation of the communication between the borrower and the master servicer. First, because servicer comments are recorded almost daily, I can estimate the delinquency state of a loan at a higher frequency than most loan-specific and macroeconomic series allow. This time-varying correlation of high frequency sentiment provides a narrative understanding of credit lifecycle of a loan.

Thematic structures are used to identify overarching themes in the document or classify the various themes in a document by relating words to latent semantic variables. Two of the most important styles in thematic structures are Latent Semantic Analysis. Hofmann 1999, Boukus and Rosenberg 2006, where singular-value-decomposition is used as a dimensionality reduction technique on the term-document matrix, and Latent Dirichlet Allocation, where a latent thematic structure can be inferred from the document.

I use the keywords identified from the servicer comments data via the list of keywords identified from experience. Historically, those experienced keywords have been categorized and mapped into *delinquency classes*, e.g., current, delinquent, short-sale, paid-in-full, bankruptcy, foreclosure. I use these newly identified keywords and perform unsupervised learning to form topics, which gives me a better understanding of the different *delinquency classes*. I implicitly identify the transitions among the previous *delinquency classes* and add more to the list. This helps me understand the life of a loan completely and what possible states a loan can be in.

First, I characterize the topical structure in servicer comments. Servicer comments can be lemmatized into easily interpretable topics with intuitive time-series patterns. A model with less than 200 topics is statistically optimal according to Bayes factor criteria. Models with fewer topics tend to mix themes into overly broad clusters, while more topics use more parameters and hence overfit. Almost all topics exhibit strong time-series persistence.

I first list the word clouds for each of the 12 topics that I identify from the lemmatized data in Figure 52. I also try with 8 topics to find some topics that contain more than one category (e.g., property condition and natural disaster). I even try 16 topics and find the same category is present in more than one topic. Hence I decide to keep 12 topics for further analysis. In Figure 53 I plot the relative weights of words in each topic and also their frequencies for topics 1-6 and a similar analysis for topics 7-12 in Figure 54. In Figure 55, I provide snapshots of each topic and the above relative weights and frequencies. More importantly, I can visualize the clustering of certain topics in Figure 55 and can identify themes based on these clustering of similar topics, which capture both cross-sectional and time-series correlations of the different categories. This step is crucial in terms of mapping these categories into delinquency states of the loans and is a clear evidence of the predictive power of the servicer comments from the soft information about borrower financial condition and property condition beyond the loan performance variables.

3.4 Model Results with four delinquency classes

In this section, I provide evidence that including predictive *soft* information using the servicer comments has incremental value over prediction just using loan performance variables. I keep two delinquency classes **PL** (Performing Loan) and **NPL** (Non-Performing Loan) and two liquidated (terminal classes/absorbing states) classes **PIF** (Paid-in-Full) and **ShortSale** to aggregate the transitions into meaningful buckets.

3.4.1 Model Results with Loan Data without Servicer Comments

I first provide the misclassification errors for the four classes **NPL** (Non-Performing Loan), **PIF** (Paid-in-Full), **PL** (Performing Loan), **ShortSale** from the In-Sample Confusion Matrices in Table 27 only using loan performance data. I then provide the first 47-49 rows of the Variable Importance Tables for DRF (in Table 28), Lasso (in Table 29) and DNN (in Table 30). The I provide the misclassification errors for the four classes **NPL** (Non-Performing Loan), **PIF** (Paid-in-Full), **PL** (Performing Loan), **ShortSale** from the Out-of-Sample Confusion Matrices in Table 31 only using loan performance data.

3.4.2 Model Results with Loan Data and Servicer Comments

Now I add the vectors created by the word2vec methodology, described in Section 3.3.2 on top of the loan performance variables in Subsection 3.4.1. I first provide the misclassification errors for the four classes **NPL** (Non-Performing Loan), **PIF** (Paid-in-Full), **PL** (Performing Loan), **ShortSale** from the In-Sample Confusion Matrices in Table 32 using loan performance data with servicer comments. I then provide the first 47-49 rows of the Variable Importance Tables for DRF (in Table 33), Lasso (in Table 34), GBM (in Table 35) and DNN (in Table 36). The I provide the misclassification errors for the four classes **NPL** (Non-Performing Loan), **PIF** (Paid-in-Full), **PL** (Performing Loan), **ShortSale** from the Out-of-Sample Confusion Matrices in Table 37 using loan performance data with servicer comments.

Clearly, there is improvement in misclassification error from Table 31 to Table 37 after adding the soft information from servicer comments. The misclassification error decreases from 3% to 2% for non-performing loans (reducing the cost in monitoring these loans), from 7% to 5% for loans that are paid-in-full (increasing the alpha for the investor), from 8% to 6% for loans that are on short sale (reducing the uncertainty in the plausible delay in pricing from the auction process in short sale). The economic significance of these improvements is huge and this helps the servicer undertake optimal actions, if they know the future delinquency status of the loans with better accuracy. Moreover, the variable importance tables (Table 28 with only hard information and Table 33 overlaying the soft information on the hard information) provide a clear indication about the variables which have the largest marginal contributions towards the default outcome.

3.5 Model Results with Granular Delinquency Classes

I conduct robustness checks for a larger and more *granular* set of delinquency classes **B120D** (Beyond 120 days), **BK** (Bankruptcy), **FC** (Foreclosure), **PIF** (Paid-in-Full), **REO** (Real-Estate owned), **ShrtSal** (Short Sale), **W0_30D** (Within 0 to 30 days), **W30_60D** (Within 30 to 60 days), **W60_90D** (Within 60 to 90 days), **W90_120D** (Within 90-120 days).

3.5.1 Granular Model Results with Loan Data without Servicer Comments

I list the *Misclassification Errors* from the In-Sample Confusions matrices of DRF, Lasso, GBM and DNN for the loan-performance data without the servicer comments data in Table 38. Then I list the first 47-50 variables in the Variable Importance tables for each of DRF (Table 39), Lasso (Table 40), GBM (Table 39) and DNN (Table 42) for the loan-performance data. Then, I list the *Misclassification Errors* from the Out-of-Sample Confusions matrices of DRF, Lasso, GBM and DNN for the loan-performance data without the servicer comments data in Table 43.

3.5.2 Granular Model Results with Loan Data and Servicer Comments

I list the *Misclassification Errors* from the In-Sample Confusions matrices of DRF, Lasso, GBM and DNN for the loan-performance as well as servicer comments data in Table 44. Then I list the first 47-50 variables in the Variable Importance tables for each of DRF (Table 45), Lasso (Table 46), GBM (Table 47) and DNN (Table 48) for the combined data. Then, I list the *Misclassification Errors* from the Out-of-Sample Confusions matrices of DRF, Lasso, GBM and DNN for the loan-performance as well as servicer comments data in Table 49.

B120D category is where the loans become *limbo* and reduction of misclassification error between Table 43 and Table 49 reduces the uncertainty regarding the future outcome of many of these loans. This impact is significant as these are the loans that the servicers worry about and they often resort to expert judgement while manually optimizing the future disposition of these loans. Bankrupt loans can be predicted with 100% accuracy. This can save a significant amount of time and dollar cost for the servicer as they will know which loans to modify and which loans to file Bankruptcy Chapter 7 (liquidation) instead of Chapter 13 (renegotiation). Reduction in misclassification errors in the W60_90D and W90_120D categories also help ascertain which loans have a chance to cure and which loans are on a sure path to adverse states.

3.6 Conclusion

Ascertaining the categories of mortgage loan delinquency and the transitions between them with reasonable accuracy is, to this date, a subjective task, leading to massive imprecision in determining the health of loans. The majority of empirical research has focused on loan-specific, borrower-specific, property-specific and macroeconomic indicators to approach this problem. I offer an alternative approach that retrieves the borrower-lender interaction vis-a-vis the soft information from the servicer comments. Our approach is motivated by the view that servicer comments provide the most accurate delinquency state of a loan. The information asymmetry between the master servicer and the special servicer has been well-documented. But, to the best of our knowledge, no attempts have been made to alleviate this issue. The flawed interpretations and the erroneous inference thereof of the special servicers can be mitigated using Machine Learning techniques. I estimate NLP models from the full text of master servicers of loans sold to a Private Equity Fund having a Real Estate portfolio. I measure which keywords are allocated to each cluster at each point in time, and then use these measurements as inputs into statistical models of loan performance. Clusters of similar words from the servicer comments closely coincide with categories that a legal firm (an affiliate of the Private Equity firm) has identified from over 15 years of experience. I show that clusters generated from master servicer comments contain substantial information about the future delinquency states above and beyond standard indicators. Our approach relies on the model to digest massive text data that are beyond human readability and flags keywords that are most similar to a specific category or delinquency state of interest.

CHAPTER 4
DISCUSSION

4.1 Chapter 1

The implications for each delinquency class are unique and can be accurately measure using the DNN methodology. The results are also highly robust during these unforeseen times of the COVID-19 pandemic.

4.1.1 COVID-19 Results without internalizing 2008 Crisis

After cleaning the data, we have 1,315,421 observations from Jan 2017 - Sep 2020. We first try training the DNN model with data from Jan 2017 - Feb 2020 and confirm that NOI is more important than LTV leading upto the COVID-19 pandemic in Table 5.5. However, consistent with our conjecture that Commercial Mortgage delinquency leads a financial (in this case, induced by a global health crisis), the out-of sample predictions are inaccurate. Then we train the DNN model from Jan 2017 - Nov 2019, as the first cases of COVID-19 were identified in Wuhan, China during December, 2019. We still find that NOI is more important than LTV in Table 5.6 and also provide evidence about other variables which become important in co-determining the default behavior. We find similar accuracy in predictions in Table 5.7 compared to Table 5.4. This solidifies our previous and we can indeed claim that DNN extracts the inherent structural relationship among the covariates and can robustly predict even during several financial crises.

4.1.2 Results by Industry during COVID-19

We track the evolution of the number of loans in different delinquency classes across time (Dec 2019 - Sep 2020) and across industries (Assisted, Healthcare, Hotel, Industrial, Multi-family, Office, Special) in Figure 5.11. We clearly see the hotel industry being massively displaced. We see enormous number of commercial mortgage loans in the hotel industry degrade from *W0_30D* to *W30_60D* (orange bar) from May, 2020. There is some degradation from *W30_60D* (orange bar) to *W60_90D* (ash-color bar). These loans become limbo from August 2020 as evident from the rise in dark-blue bar of *B120D* loans. The office space has also seen massive cashflow shortages from the lack of business activities and inability of payments therefrom for tenants in big cities. This is evidenced by constant green bar (Non-Performing Mature Balloon Loans) which are trying to roll-over the loan contract and some oare successful from the lax underwriting standards due to historical levels of low interest rate monetary policy.

4.1.3 Determinants of Each Delinquency Class

We zoom in to individual delinquency classes and find the determinants of each adverse delinquency class using DNN. In Table 5.9, we still find NOI higher than LTV for DNN trained on Jan 2017 - Nov 2019 and misclassification error is almost 0%. For Real-Estate-owned (REO) loans, the misclassification error is again 0% in Table 5.10, but NOI is no longer higher than LTV in Table 5.11 as there is no NOI when the lender takes back the property. For foreclosure, again the misclassification error is again 0% in Table 5.12 and similarly NOI is no longer higher than LTV in Table 5.13 as foreclosure proceedings are lengthy processes which start after Bankruptcy Chapter 13 and there is no renegotiation to be done and either the loan gets resolved in court or leads to the terminal state of Bankruptcy Chapter 7 (liquidation). As shown in Subsection 4.1.2, the behavior of W90_120D and B120D loans are determined by industry heterogeneity. There is slightly higher error (in Tables 5.15 and 5.17) for these two uncertain delinquency states as the COVID-19 is an ongoing pandemic and still unfolding.

4.2 Chapter 2

Although, I do not have data on individual borrower sophistication (I have it at zip code level), but the concentration of forbearance applications or lack thereof from smaller counties suggests borrowers availing forbearance from hearsay. Some counties who were not abreast with the details of the CARES Act, especially for the Non-Gov. borrowers. Hence, they did not avail forbearance or were dissuaded. For a longer term permanent remediation of this issue, the government should initiate a marketing campaign whenever they enact something as serious as the CARES Act. The marketing campaign, although a short term solution, can lead to more awareness uniformly across the cohort of borrowers in an entire community. Educating financially less fortunate borrowers about consistent payments and loan modification options available to them can lead to a permanent solution which can prevent them being left out in the future.

The government had very little time to react after the pandemic ensued around March 2020. The CARES Act was designed in a rush to provide immediate relief to the hardest hit portion of the population. But the implementation of the CARES Act was conducted through the temporary non-payment of mortgage obligations for borrowers in financial hardship. The servicer acted as a crucial financial intermediary in the implementation of the CARES Act. Hence, the Govt.-backed borrowers were approved forbearance as the servicers do not want to mess with Govt. policies. The rampant approval of forbearance for Govt.-backed borrowers heavily affected the servicers' cash flow. Hence, the servicers, in order to compensate for this loss and to deal with this timing of cashflow dilemma, used discretionary adverse selection on African Americans and other minority borrowers, because the CARES Act left a grey area in their documentation. This clearly puts into question whether the government should depend on a financial intermediary to implement the policies after they are enacted.

More broadly, the government should look at the entire mortgage and housing market while implementing these adhoc policies. The CARES Act was specifically designed for the Govt.-backed borrowers. However, the share of Govt.-backed loans originated at the county level needed to be slowly increased, so that borrowers will be inclined to avail these Govt.-backed loans and then a policy on Govt-backed loans would be successful. However, the blatant ignorance towards the Non-Gov loans would create more stress in the real economy overall.

4.3 Chapter 3

Asset pricing economists have a simplistic view of mortgages as contracts sharing risk between lenders and borrowers. Borrowers with existing mortgages do not face margin calls on their collateral if they are in financial distress, or if the standards for underwriting new mortgages become tighter.

4.3.1 Impact of Soft Information on Asset Price

For better asset pricing for mortgages, a better prediction of the delinquency status is the key ingredient. For broader delinquency classes PL, NPL, PIF, ShortSale, when I compare the Out-of-Sample confusion matrices in Table 31 (only loan data) with Table 37 (loan data with soft information from servicer comments), I see significant improvement in misclassification error for the delinquency classes NPL and ShortSale across all 4 models. ShortSale is a severely adverse delinquency status. The reduction of misclassification to 25% of the original 8% has significant asset pricing implications. The costly auction process during shortsale can be avoided leading to material reduction in deadweight loss of the lender. Also, an accurate shortsale price negates the spillover effect of the Short Sale process on the future price of the property. The improvement in accuracy for NPL loans lead to better monitoring and the similar reduction in misclassification error for PIF loans lead to an immediate alpha for the investors. The better attribution of default risk leads to a better asset pricing, both at the loan-level and at the property-level.

Also, DNN performs really bad in Table 31 (only loan data). This is because the hard information is highly correlated cross-sectionally and is mostly sticky across time. This is again a clear evidence that a model as sophisticated as DNN cannot differentiate between current and adverse loans. Hard information mostly cannot capture the time-sensitive soft information which I now include in Table 37. Here, one can see how well the DNN classifies the good and the adverse delinquency classes. The misclassification errors are close to 0%.

4.3.2 Impact of Soft Information on Asset Volatility

I use a subtle methodology to capture the impact of quantifying soft information on asset volatility. I disintegrate the PL class into W0_30D and W30_60D classes based on months of missed payments and NPL class into W60_90D, W90_120D, B120D, BK, FC, REO, to identify the transitions into granular classes and in essence, capture the volatility of the asset quality. For a 10-class classification, with highly unbalanced data, DNN no longer performs good. The results of GBM and DRF are extremely robust. When I compare Table 43 (only loan data) with Table 49 (loan data with soft information from servicer comments) for the granular classes, I find that misclassification error is significantly reduced for the adverse classes like W90_120D, B120D, BK, FC, REO, etc. This reduces the estimated volatility of the process of classification. Bankruptcy, foreclosure, REO are all costly processes and classifying these loans more accurately saves time and money for the lender and allows the borrower to avail loan modification and to avoid going to a point of no return and permanently jeopardizing her future creditworthiness.

4.4 Corona Virus and Ongoing Work

A closer look at some of the servicer comments during March-April 2020 provides insights from phrases relating to COVID-19 and the CARES Act. The occurrences naturally include words like ’COVID’, ’corona’, ’coronavirus’ and ’virus’, as well as words like ’forbearance’ and ’foreclosure moratorium’ in Figure 50. Another important topic I look out for is unemployment and/or a loss in pay, as indicated by the phrase ’curtailment of income’. While occurrences of the words COVID, coronavirus and virus are always in the right context, the word corona appears on multiple occasions warranting its selective removal. I also studied the interplay of words like COVID and forbearance or COVID and curtailment of income, to see if any forbearances or losses in income were originating from layoffs from COVID or other COVID related causes.

Firstly, words relating to the virus, forbearance, foreclosure moratoriums and curtailment of income as well as some of their cross products (eg: COVID + forbearance) in Figure 51 are good indicators to track the status of loans under the CARES Act. It could help me differentiate between strategic and nonstrategic defaulters by individually tracking loans that associate losses in pay or jobs to the virus or those requesting forbearances due to COVID. Secondly, the Act was signed in the last week of March, implying that it will take some time for me to see its effects taking place. The comments from this month across loan pools will be crucial in identifying trends in delinquency classes and future payments. Life events like death and divorce might play a significant role in the months to come. Thirdly, coupling the identified loans with the industry the borrower works in, their occupation (if present) and their age group may enable me to generate a consolidated risk metric for all potential borrowers enabling more effective tracking and delinquency class predictions that might aid the legal team.

CHAPTER 5
CONCLUSION

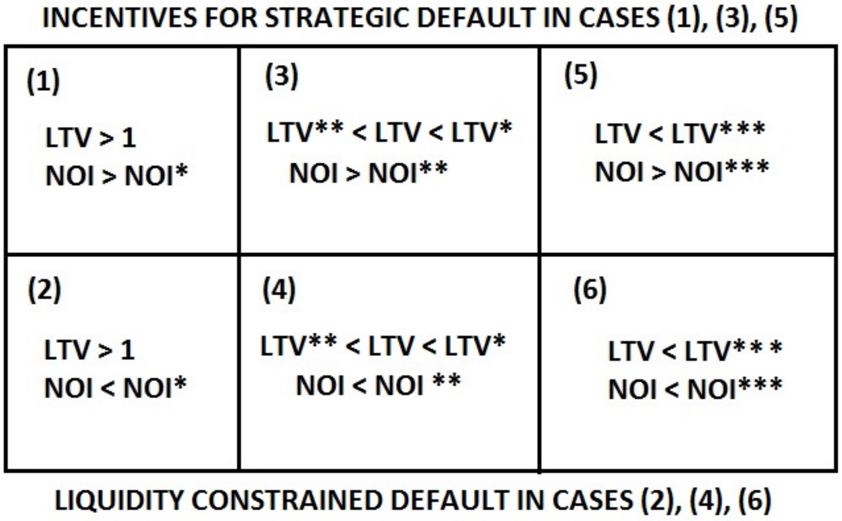
Machine Learning is still considered as a black-box in Finance and/or Economics profession, despite, innovations in leaps and bounds, in the last 10 years. Unprecedented computational resources, access to unique big data has enabled me to answer three fundamental questions in real estate finance, both commercial and residential. My thesis [1] aims to bridge the gap between the use of flexible, interpretable ML models without full-fledged theories and mainstream econometrics. I disentangle strategic default from liquidity-constrained default, which has been an open research agenda in mortgage (due to lack of data on material life events, e.g., death, divorce, job loss trigger, etc.) literature and econometrics for 30-40 years. The context of commercial mortgages enables me discern a whole host of strategic behaviors due to unique contractual features. Similarly, I use an administrative data on proprietary servicer call transcripts and retrieve unforeseen time-varying soft information about borrowers. I also use this data to track financial hardship real-time emanating from the current COVID-19 pandemic. I find overcrowding in terms of forbearance application by certain borrowers in Govt. programs (FHA/VA/USDA) who did not suffer from unemployment or curtailment of income. In my thesis as well as in my current working papers, I emphasize the power and flexibility of ML techniques. I intend to further theoretical contributions in the field of deep neural networks so that neural networks can used side-by-side mainstream econometrics. I have work in progress on the concepts of identification strategy, instrumental variable and causality in the realm of deep neural networks.

[1] I am grateful for AREUEA Dissertation Award among 7 PhD candidates chosen globally.

APPENDICES

5.1 Deep Learning for disentangling Liquidity-constrained and Strategic Default

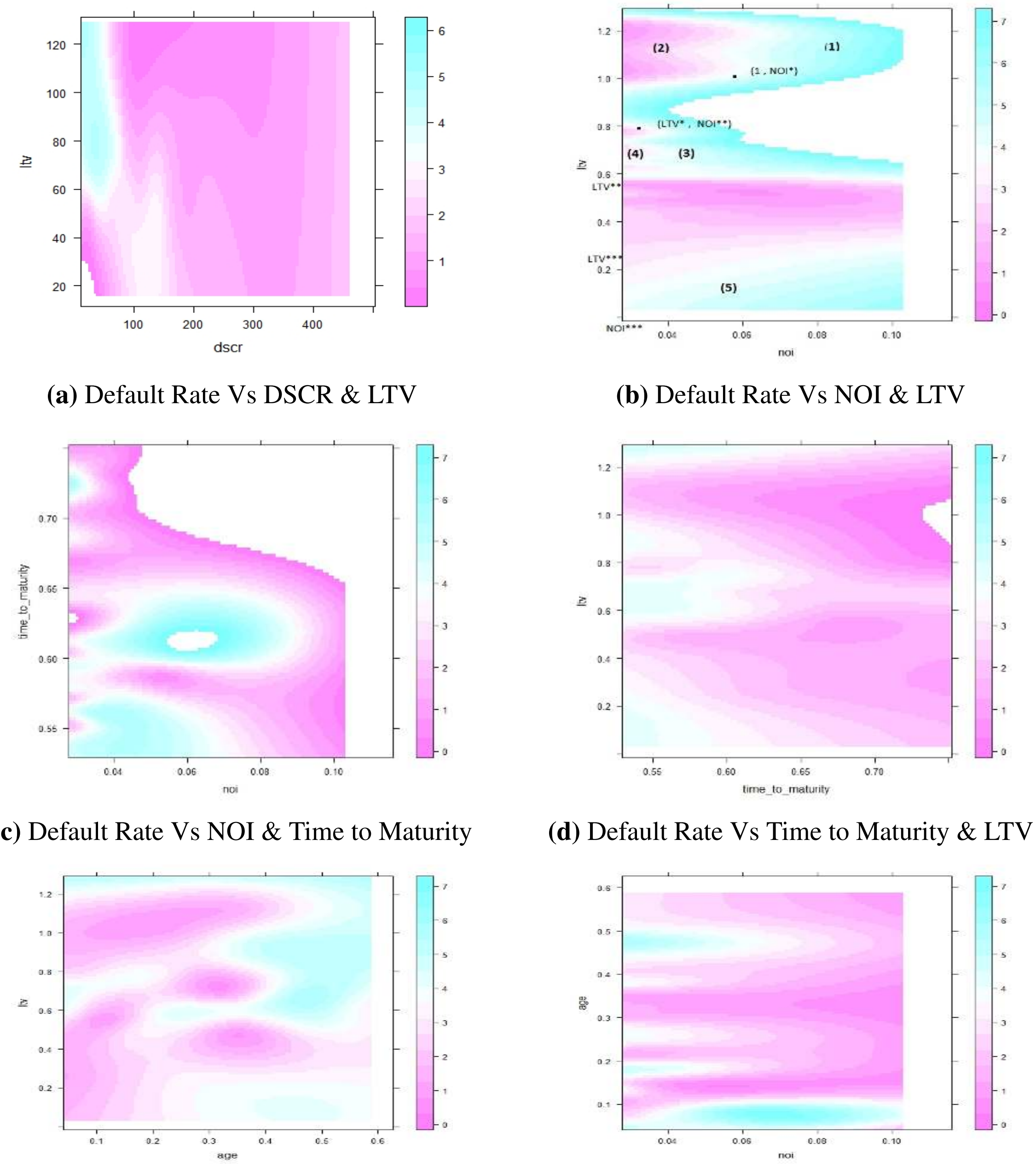
Figure 5.1: We list the possible combinations of LTV and NOI that can *disentangle* Liquidity-constrained Default and the incentives for Strategic Default behavior.



INCENTIVES FOR STRATEGIC DEFAULT IN CASES (1), (3), (5)

(1) LTV > 1 NOI > NOI*	(3) LTV** < LTV < LTV* NOI > NOI**	(5) LTV < LTV*** NOI > NOI***
(2) LTV > 1 NOI < NOI*	(4) LTV** < LTV < LTV* NOI < NOI **	(6) LTV < LTV** * NOI < NOI***

LIQUIDITY CONSTRAINED DEFAULT IN CASES (2), (4), (6)



(**a**) Default Rate Vs DSCR & LTV

(**b**) Default Rate Vs NOI & LTV

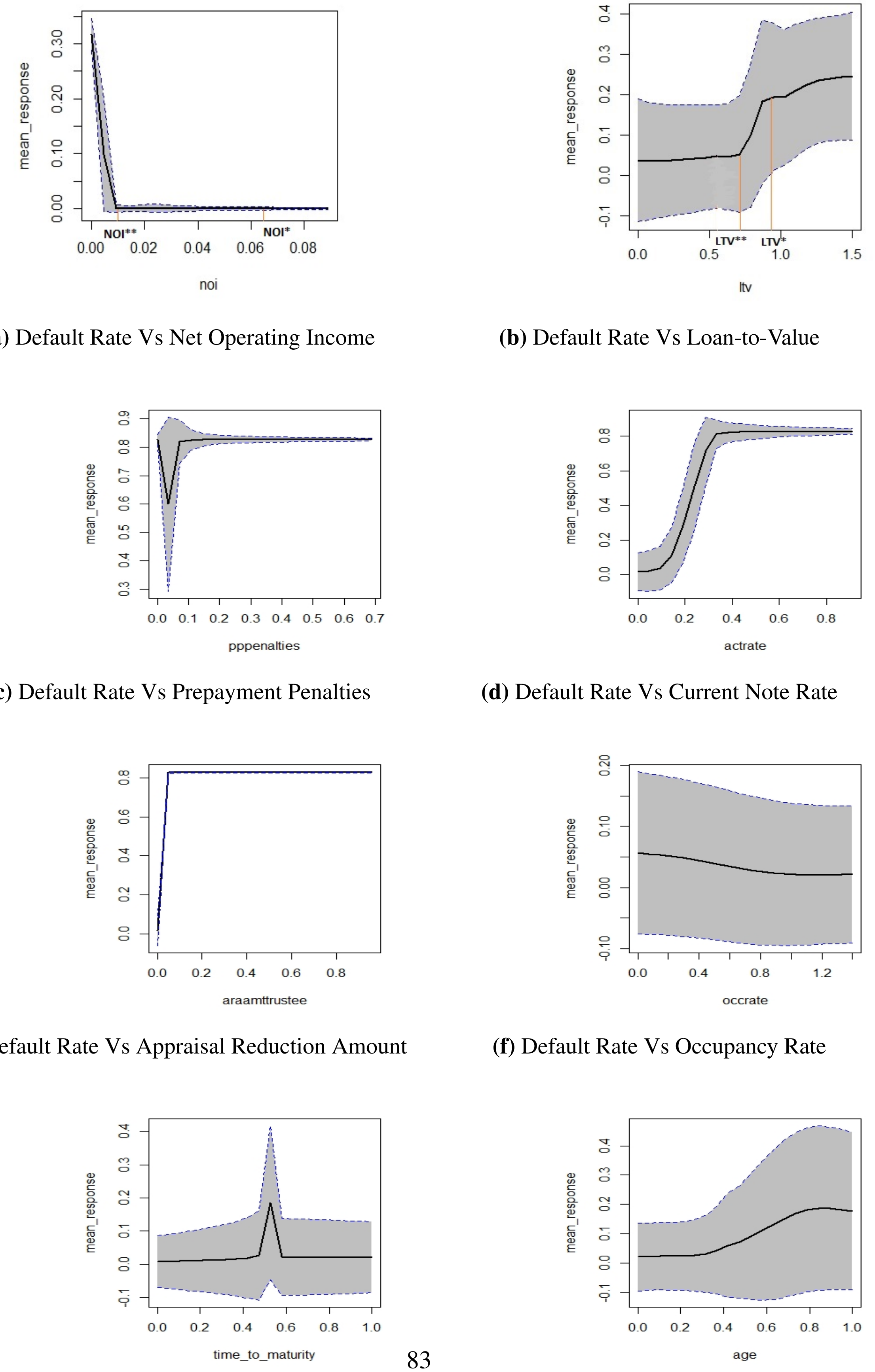
(**c**) Default Rate Vs NOI & Time to Maturity

(**d**) Default Rate Vs Time to Maturity & LTV

(**e**) Default Rate Vs Age & LTV

(**f**) Default Rate Vs NOI & Age

Figure 5.2: Bivariate Heatmaps are 2D projections of Default Rate surface over two co-vaoriates. Blue color signifies high default rate and pink color low default.



(a) Default Rate Vs Net Operating Income

(b) Default Rate Vs Loan-to-Value

(c) Default Rate Vs Prepayment Penalties

(d) Default Rate Vs Current Note Rate

(e) Default Rate Vs Appraisal Reduction Amount

(f) Default Rate Vs Occupancy Rate

(g) Default Rate Vs Time to Maturity

(h) Default Rate Vs Age of Loan

Figure 5.3: Partial Dependence Plots for Predicted Default Rate **??**

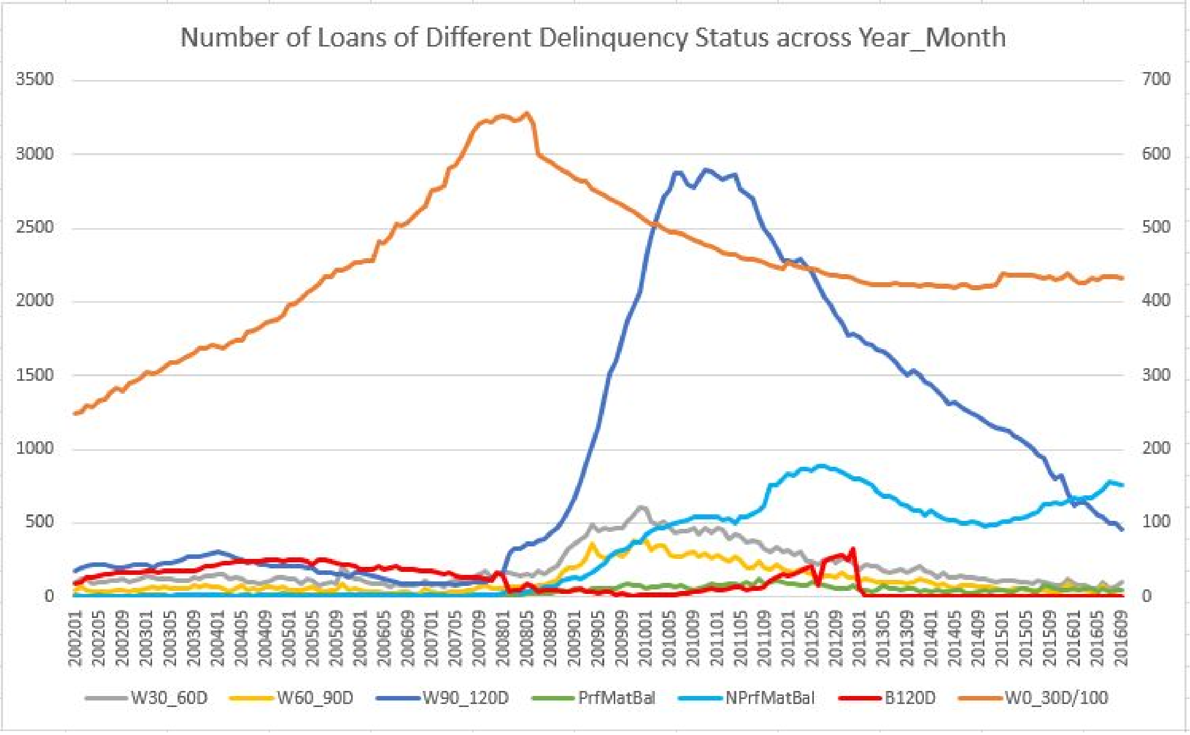


Figure 5.4: This diagram provides the evolution of delinquency buckets by year.

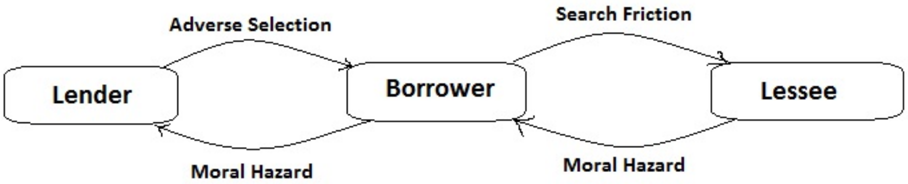
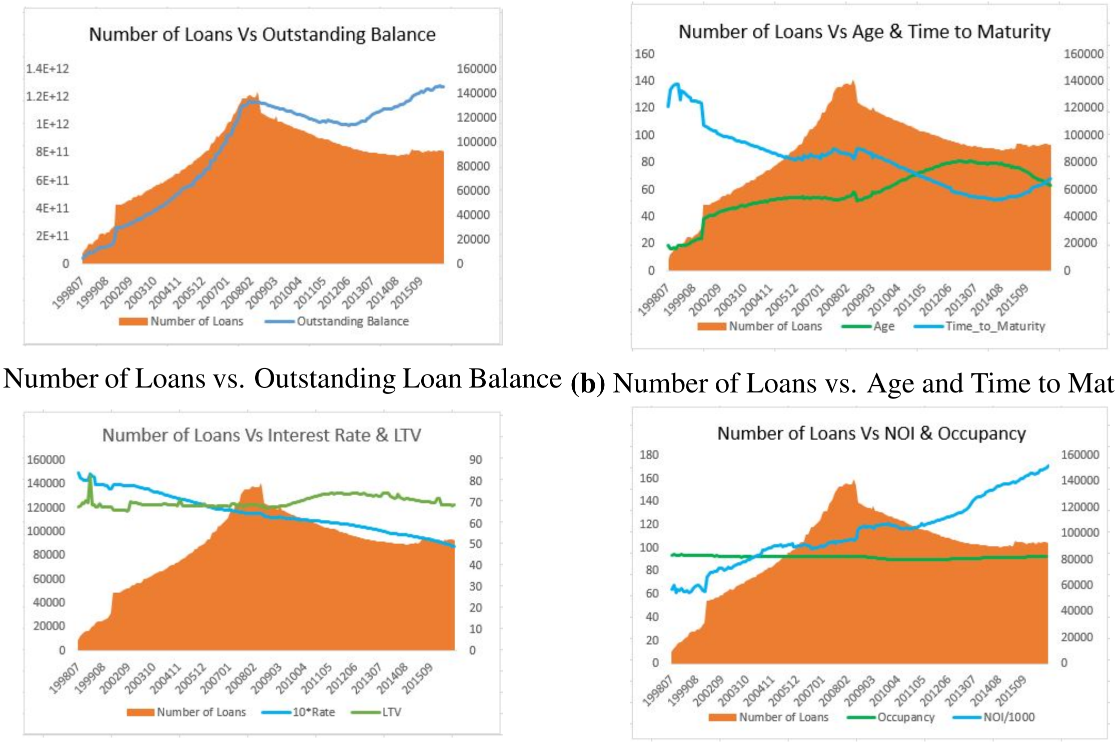


Figure 5.5: Adverse Selection and Moral Hazard



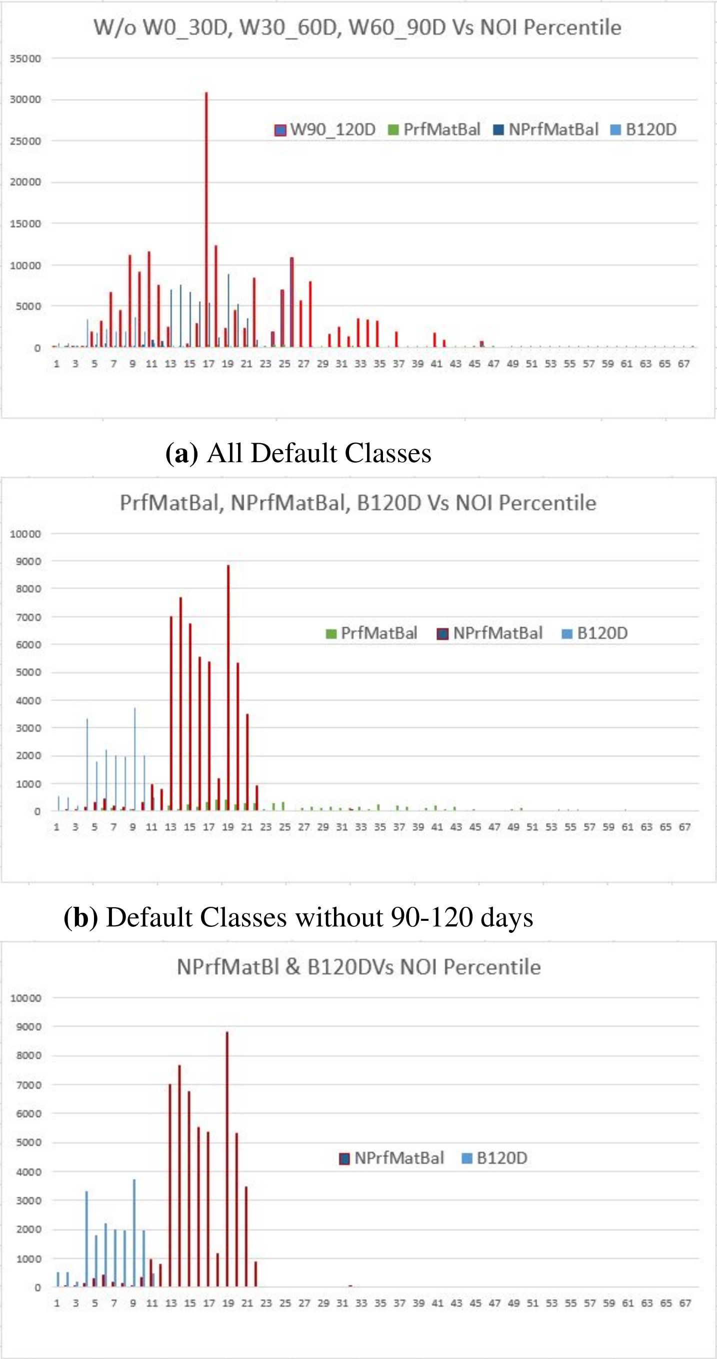
(a) Number of Loans vs. Outstanding Loan Balance

(b) Number of Loans vs. Age and Time to Maturity

(c) Number of Loans vs. Interest Rate and LTV

(d) Number of Loans vs. NOI and Occupancy

Figure 5.6: We provide evidence from the Trepp data in Figure 5.6a that from 2012, the number of loans have remained flat but the outstanding balance of loans have steadily increased until 2016.

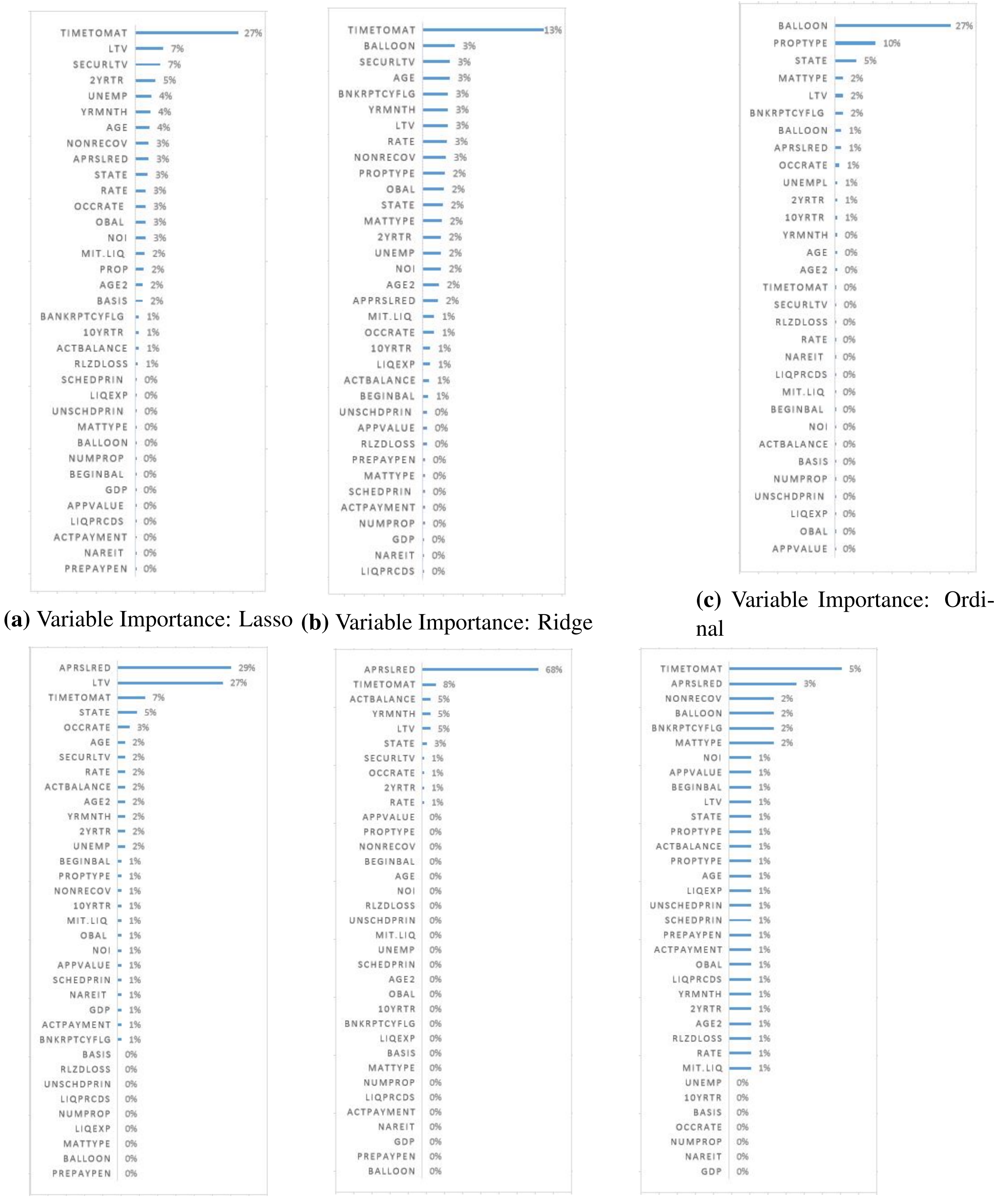


(a) All Default Classes

(b) Default Classes without 90-120 days

(c) Default Classes without 90-120 days & Performing

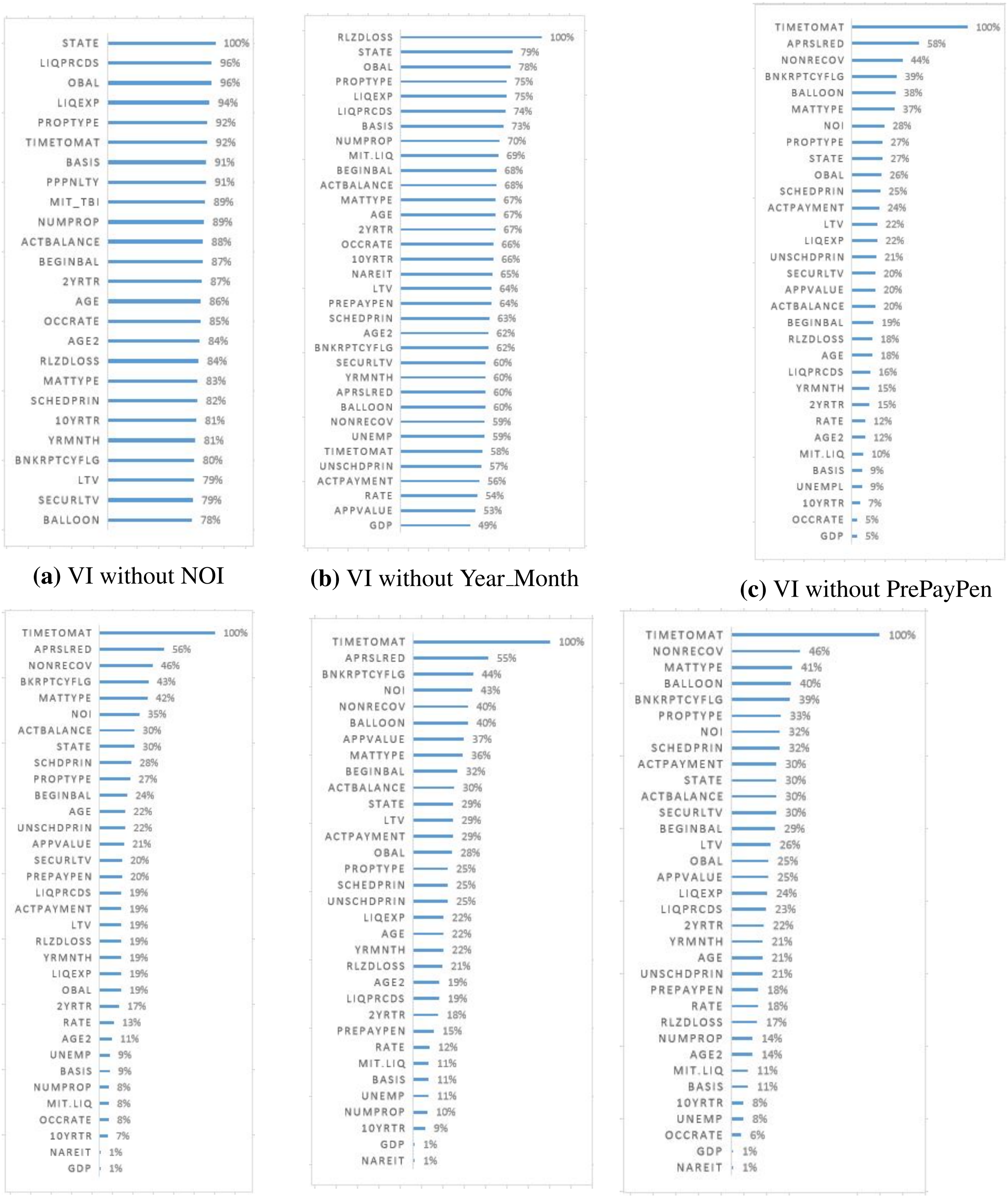
Figure 5.7: The delinquency states.



(a) Variable Importance: Lasso **(b)** Variable Importance: Ridge **(c)** Variable Importance: Ordinal

(d) Variable Importance: DRF **(e)** Variable Importance: GBM **(f)** Variable Importance: DNN

Figure 5.8: Variable Importance for 6 models, namely, Lasso, Ridge, Ordinal, Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), Deep Neural Network (DNN).



(a) VI without NOI

(b) VI without Year_Month

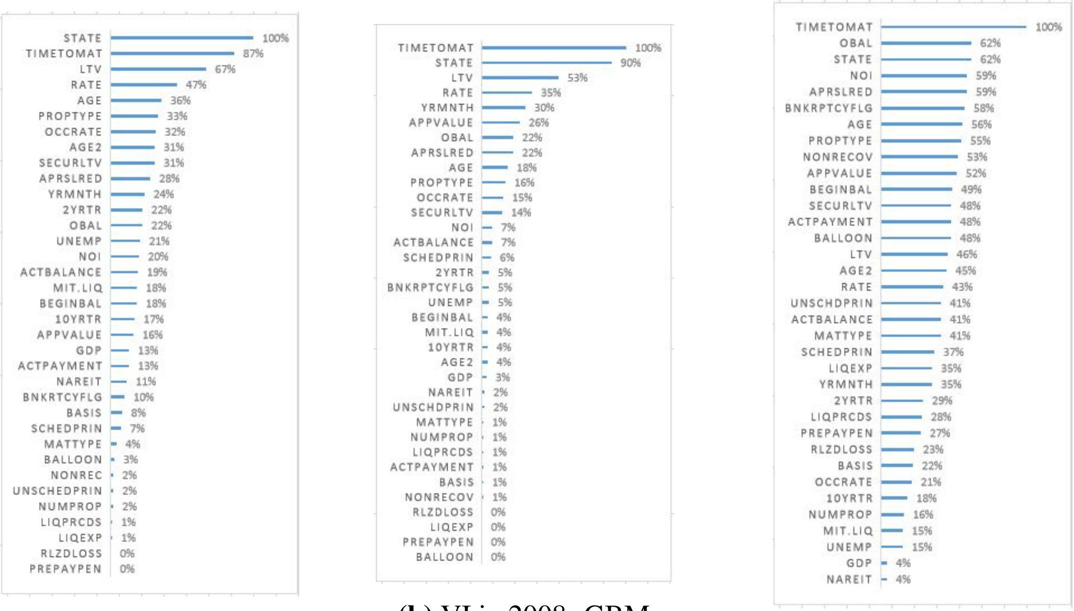
(c) VI without PrePayPen

(d) VI without Balloon Payment

(e) VI without Occupancy

(f) VI without Appraisal Reduc

Figure 5.9: These are Variable Importance (VI) charts, leaving one out.



(a) VI in 2008: DRF

(b) VI in 2008: GBM

(c) VI in 2008: DNN

Figure 5.10: Variable Importance is stress-tested during Financial Crisis across all non-parametric models.

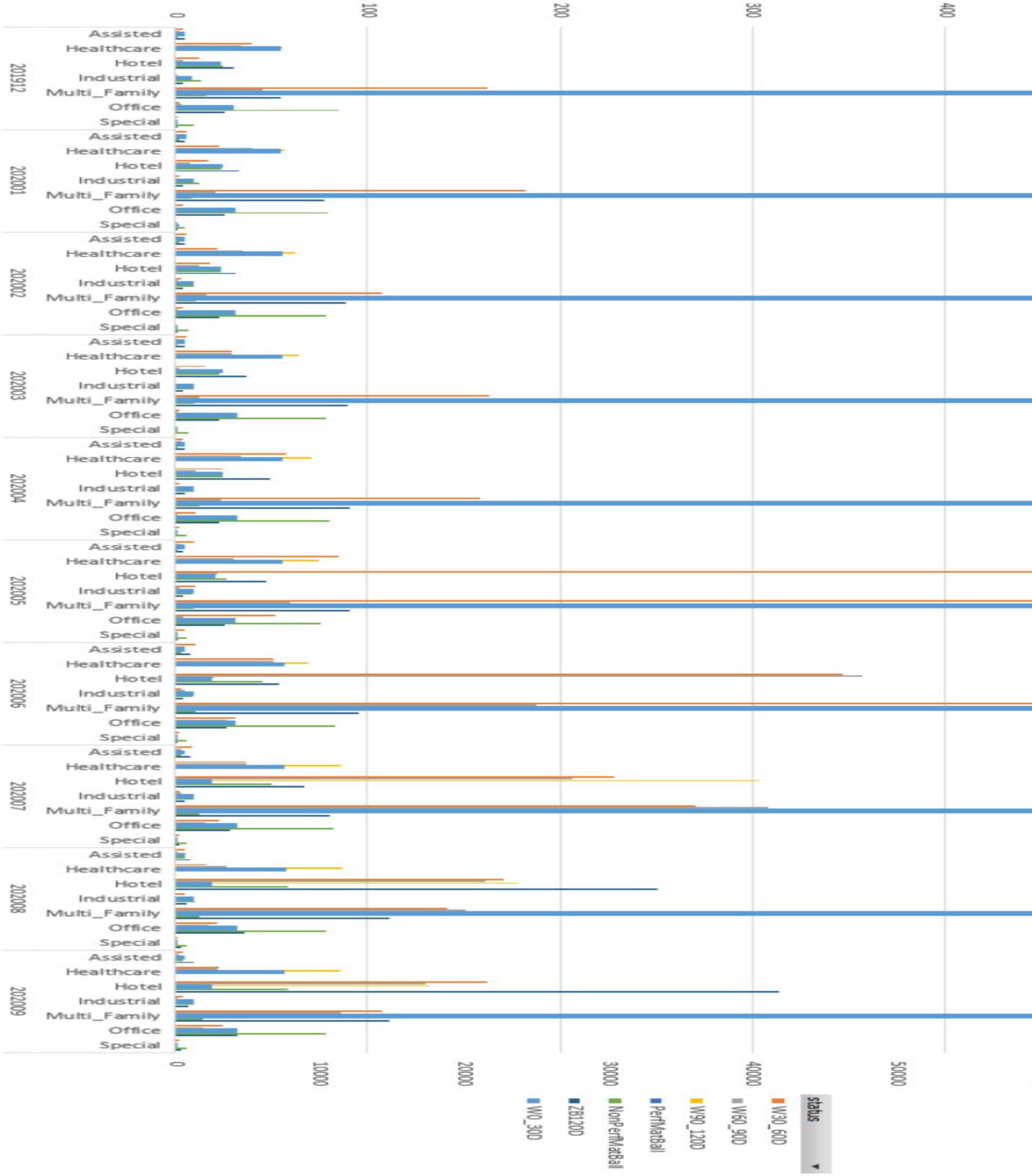


Figure 5.11: Impact of COVID-19 across industries

Table 5.1: The summary statistics for the cleaned data contains 9,617,333 observations of continuous variables.

Statistic	N	Min	Pctl(25)	Median	Mean	Pctl(75)	Max
beginbal	9,617,333	0	1,838,146	4,036,697	7,882,727	9,100,000	99,990,043
orig_bal	9,617,333	0	1.92M	4.10M	9.31M	9.14M	1.68B
rate	9,617,333	0	6	6	6	7	9
Sched_princip	9,617,333	0	1,060	4,235	15,665	9,838	430M
Unsched_prin	9,617,333	0	0	0	50,300	0	1.50B
balance_act	9,617,333	0	1,774,511	3,980,981	7,797,502	9,009,824	99,999,000
payment	9,617,333	0	15,000	28,600	65,900	57,800	675M
pppenalties	9,617,333	0	0	0	420	0	29,477,125
liqproceeds	9,617,333	0	0	0	17,900	0	2.56B
realizedloss	9,617,333	0	0	0	5,260	0	204M
liqexpense	9,617,333	0	0	0	3,200	0	1.06B
numprop	9,617,333	1	1	1	1	1	225
Appraisal_Reduc	9,617,333	0	0	0	132,000	0	391M
SecurLTV	9,617,333	0	63	71	67	76	150
Face	9,617,333	0	6	6	6	7	9
NOI	9,617,333	0	268,000	527,000	1.22M	1.10M	1.09B
LTV	9,617,333	0	63	71	69	77	150
AppValue	9,617,333	1,620	3.46M	6.80M	17.4M	14.5M	48.1B
OccRate	9,617,333	0.0	0.89	0.96	0.92	1.0	1.4
Basis	9,617,333	0.0	2.0	2.0	1.7	2.0	4.0
Unemp	9,617,333	0.019	0.049	0.061	0.066	0.080	0.154
GDP	9,617,333	0	008	016	013	048	0.0174
2YrTr	9,617,333	02	06	0.010	0.019	0.031	0.061
10YrTr	9,617,333	0.015	0.024	0.036	0.034	0.043	0.063
NAREIT	9,617,333	-0.45	-0.05	0.01	-0.01	0.04	0.38
MIT.Liq	9,617,333	0.017	0.022	0.036	0.025	0.105	0.184
time_to_maturity	9,617,333	0	35	66	80	100	765
age	9,617,333	0	31	60	63	90	409
age2	9,617,333	0	961	3600	5393	8100	167281

Table 5.2: Coefficients of Multinomial Logistic Regression: The marginal effect does not have a clear interpretation as evidenced by the co-efficients of Multinomial Logit.

	names	W0_30D	W30_60D	W60_90D	W90_120D	PrfMatBal	NPfMtBl	B120D
1	Intercept	-24.83	-230.36	-223.45	-319.34	591.88	-170.20	272.02
2	State	-0	0.69	0.30	0.99	-2.12	-0.76	-0.08
52	PropType	-0.39	0.65	-0.12	0.61	0.55	0.62	0.49
60	MatType	0.23	-4.82	-2.87	-3.79	7.32	-57.90	-6.62
61	Balloon	0.33	-5.39	-3.48	-4.05	5.79	-53.02	-5.83
62	NonRecov	-1.59	-0.94	-0.75	0.69	-3.13	0.53	0.43
63	BnkrptcyFlg	-0.15	0.02	0.19	1.10	-0.78	-1.88	-0.09
64	BeginBal	-3.39	-2.13	-3.27	1.56	-0.67	-8.03	5.99
65	Obal	-1.18	21.62	19.06	-41.82	18.39	-7.16	-115.35
66	**Rate**	**-17.31*****	29.77	29.74	34	**-27.34*****	8.05	32.65
	SE	**0.21**	0.35	0.35	0.40	**0.33**	0.10	0.39
67	SchedPrin	134.71	153.33	54.63	89.44	156.85	173.07	41.50
68	Payment	44.44	-361.30	-6.64	-37.43	24.03	-1.95	57.25
69	UnschdPrin	45.46	-133.66	70.57	-47.70	57.08	-71.93	-172.29
70	PrePayPen	11.65	-253.05	-103.59	-61.27	-44.45	-370.52	16.91
71	ActBalance	2.25	0.94	2.73	-2.03	-0.28	8.07	-4.28
72	Liqprcds	-75.90	76.39	35.06	131.46	8.39	179.17	165.38
73	Liqexp	448.71	-1139.31	-2653.88	-518.47	-124.37	-1035.34	72.08
74	RlzdLoss	69.41	26.56	-48.44	-42.75	39.97	-43.45	-58.66
75	NumProp	-2.03	3.72	4.96	16.78	5.49	4.55	-23.82
76	AprslRed	-101.36	8.64	11.74	66.68	-20.31	49.31	87.47
77	SecurLTV	5.65	5.77	5.39	3.49	3.20	3.68	-4.19
78	NOI	159.73	-7.38	-39.41	27.03	124.60	-77.96	-149.62
79	**LTV**	**-6.55*****	-3.10	-2.39	0.03	-3.56	-2.21	0.77
	SE	**1.11**	0.53	0.41	0.01	0.61	0.38	0.13
80	AppValue	-2.27	3.02	4.01	7.02	-201.11	11.14	9.66
81	OccRate	0.42	-3.10	-3.55	-3.51	-1.27	-2.28	-2.84
82	Basis	-7.14	-2.13	-1.89	0.28	19.61	48.04	5.77
84	Age	-6.55	-5.95	-4.29	-4.77	-2.70	6.02	8.03
85	Age2	16.16	0.62	-2.87	-1.52	12.32	-10.49	-8.52
86	**Unemp**	**-1.93*****	17.51	21.77	18.87	25.90	21.55	**-12.23*****
	SE	**0.05**	0.43	0.53	0.46	0.64	0.53	**0.30**
87	GDP	-0.84	0.75	0.27	3.30	8.21	2.35	-6.66
88	2YrTr	4.89	-7.05	-9.52	-34.94	-5.31	-50.57	-3.23
89	10YrTr	10.94	18.14	18.82	37.29	-20.60	7.96	15.26
90	NAREIT	0.10	0.13	0.32	0.20	0.02	0.41	0.17
91	MIT.Liq	0.29	-0.12	-0.35	-0.31	0.27	-0.34	3.71
92	TimeToMat	4.69	4.63	4.43	4.57	-101.20	-140.81	2.91

Table 5.3: Cross-Validation Training Errors of Models across Delinquency Classes with sample ranging from Jan 1998 - June 2012.

	Naive	**Mult**	**Lasso**	**Ridge**	**DRF**	**GBM**	**DNN**
W0_30D	0.01	0.01	0.01	0.01	0.01	0.01	0.0
W30_60D	**0.61**	1	1	1	**0.98**	1	1
W60_90D	**0.63**	1	1	1	**0.98**	1	1
W90_120D	0.36	0.48	0.48	0.47	**0.21**	**0.27**	**0.21**
PrfMatBal	1	1	1	1	**0.76**	**0.84**	**0.78**
NPrfMatBal	1	0.67	0.7	0.74	**0.16**	**0.12**	**0.12**
B120D	1	0.8	0.87	0.83	**0.22**	**0.21**	**0.2**
Totals	0.03	0.02	0.02	0.02	0.01	0.01	0.01

Table 5.4: Out-of-Sample Test Errors of Models across Delinquency Classes with sample ranging from Jul 2012 - June 2016.

	Naive	**Multi**	**Lasso**	**Ridge**	**DRF**	**GBM**	**DNN**
W0_30D	0.01	0.01	0.01	0.01	0.01	0.01	0.01
W30_60D	**0.89**	1	1	1	1	1	1
W60_90D	**0.59**	1	1	1	1	1	1
W90_120D	0.2	0.37	0.37	0.37	**0.14**	**0.15**	**0.17**
PrfMatBal	0.98	1	1	1	**0.97**	**0.96**	**0.95**
NPrfMatBal	1	0.42	0.42	0.47	**0.19**	**0.13**	**0.13**
B120D	1	1	1	1	**0.07**	**0.08**	**0.07**
Totals	0.98	0.03	0.03	0.03	0.03	0.03	0.03

Table 5.5: Variable Importance table for DNN (Deep Neural Network) across Delinquency Classes during COVID-19.

variable	relative_importance
state	1
appvalue	0.99
nonrecover	0.94
bankruptcyflag	0.90
securltv	0.90
proptype	0.89
2Yr_Tr	0.87
Unemployment	0.83
pmtbas	0.83
hasballoon	0.82
maturitytype	0.82
age	0.80
face	0.79
year_month	0.78
gdp	0.76
10Yr_Tr	0.73
actrate	0.73
noi	0.72
age2	0.70
occrate	0.68
securwac	0.68
actpmt	0.68
ltv	0.67
time_to_maturity	0.65

Table 5.6: Variable Importance table for DNN (Deep Neural Network) across Delinquency Classes during COVID-19.

variable	relative_importance
age2	1
age	0.93
state	0.90
time_to_maturity	0.86
proptype	0.78
face	0.68
maturitytype	0.64
nonrecover	0.60
hasballoon	0.60
actrate	0.60
securwac	0.59
actpmt	0.57
bankruptcyflag	0.56
pmtbas	0.56
appvalue	0.56
securltv	0.55
year_month	0.52
noi	0.51
2Yr_Tr	0.50
occrate	0.45
ltv	0.41
10Yr_Tr	0.39
Unemployment	0.27
gdp	0.07

Table 5.7: Out of sample Confusion Matrix for DNN (Deep Neural Network) across Delinquency Classes.

	W0_30D	**W30_60D**	**W60_90D**	**W90_120D**	**PrfMatBal**	**NPrfMatBal**	**B120D**	**Error**
W0_30D	565681	10	81	370	0	3588	1290	0.01
W30_60D	3174	0	5	4	0	31	35	1
W60_90D	1684	0	0	4	0	16	20	1
W90_120D	1396	0	0	11	0	47	39	0.99
PrfMatBal	493	0	0	4	0	148	6	1
NPrfMatBal	999	0	0	136	0	1566	43	0.43
B120D	2495	0	3	16	0	97	277	0.90
Totals	575922	10	89	545	0	5493	1710	0.03

Table 5.8: Out of sample Confusion Matrix Misclassification Error for Bankruptcy is close to 0% during COVID-19

	Non-BK	**BK**	Error
Non-BK	581697	13	0
BK	22	2037	0.01
Totals	581719	2050	0

Table 5.9: Variable Importance for Bankruptcy using DNN still have NOI higher than LTV

variable	relative_importance
bankruptcyflag	1
state	0.64
nonrecover	0.56
prop	0.54
hasballoon	0.51
maturitytype	0.51
appvalue	0.49
age	0.44
2Yr_Tr	0.43
securltv	0.42
noi	0.40
actrate	0.40
year_month	0.39
face	0.39
10Yr_Tr	0.39
Unemployment	0.38
actpmt	0.38
pmtbas	0.38
gdp	0.37
securwac	0.36
occrate	0.35
age2	0.35
ltv	0.33
time_to_maturity	0.29

Table 5.10: Out of sample Confusion Matrix Misclassification Error for REO is close to 0% during COVID-19

	Non-REO	REO	Error
Non-REO	364993	1084	0
REO	782	365042	0
Totals	365775	366126	0

Table 5.11: Variable Importance for REO using DNN no longer have NOI higher than LTV

variable	relative_importance
age	1
state	0.97
prop	0.89
age2	0.85
actpmt	0.79
ltv	0.76
securltv	0.75
securwac	0.73
hasballoon	0.72
time_to_maturity	0.72
bankruptcyflag	0.71
face	0.67
pmtbas	0.67
nonrecover	0.66
maturitytype	0.66
actrate	0.60
noi	0.59
appvalue	0.58
occrate	0.52
year_month	0.44
2Yr_Tr	0.40
Unemployment	0.36
10Yr_Tr	0.26
gdp	0.08

Table 5.12: Out of sample Confusion Matrix Misclassification Error for Foreclosure is close to 0% during COVID-19

	Non-FCL	FCL	Error
Non-FCL	364993	1084	0
FCL	782	365042	0
Totals	365775	366126	0

Table 5.13: Variable Importance for Foreclosure using DNN no longer have NOI higher than LTV

variable	relative_importance
age	1
state	0.97
prop	0.89
age2	0.85
actpmt	0.79
ltv	0.76
securltv	0.75
securwac	0.73
hasballoon	0.72
time_to_maturity	0.72
bankruptcyflag	0.71
face	0.67
pmtbas	0.67
nonrecover	0.66
maturitytype	0.66
actrate	0.60
noi	0.59
appvalue	0.58
nonrecover.N	0.54
occrate	0.52
year_month	0.44
2Yr_Tr	0.40
Unemployment	0.36
10Yr_Tr	0.26
gdp	0.08

Table 5.14: Variable Importance for W90_120D loans using DNN no longer have NOI higher than LTV

variable	relative_importance
state	1
age	0.86
proptype	0.75
bankruptcyflag	0.75
actrate	0.75
securltv	0.74
hasballoon	0.71
ltv	0.71
actpmt	0.69
pmtbas	0.68
securwac	0.68
face	0.64
nonrecover	0.64
noi	0.62
age2	0.61
appvalue	0.61
2Yr_Tr	0.60
maturitytype.	0.59
time_to_maturity	0.58
occrate	0.54
year_month	0.48
10Yr_Tr	0.33
Unemployment	0.31
gdp	0.16

Table 5.15: Out of sample Confusion Matrix Misclassification Error for W90_120D is high during COVID-19

	Non-W90_120D	W90_120D)	Error
Non-W90_120D	577640	4636	0.01
W90_120D	1365	128	0.91
Totals	579005	4764	0.01

Table 5.16: Variable Importance for B120D loans using DNN no longer have NOI higher than LTV

variable	relative_importance
age	1
state	0.97
prop	0.89
age2	0.85
actpmt	0.79
ltv	0.76
securltv	0.75
securwac	0.73
hasballoon	0.72
time_to_maturity	0.72
bankruptcyflag	0.71
face	0.67
pmtbas	0.67
nonrecover	0.66
maturitytype	0.66
actrate	0.60
noi	0.59
appvalue	0.58
occrate	0.52
year_month	0.44
2Yr_Tr	0.40
Unemployment	0.36
10Yr_Tr	0.26
gdp	0.08

Table 5.17: Misclassifical Error is high for B120D loans

	Non-B120D	B120D	Error
Non-B120D	579326	1555	0
B120D	2382	506	0.82
Totals	581708	2061	0.01

.1 Loan to Value and Net Operating Income

We take a deeper dive and investigate LTV in the following way:

$$LTV_t = \frac{AOB_{t-1} + DS_t + B_T}{MV_t - AR_t} \tag{1}$$

where AOB_t is the Outstanding Balance at time t-1 that is amortized , DS_t is the scheduled payment due for servicing the debt obligation at time t, B_T is the Balloon Payment due at maturity, MV_t is the Market value of the property/properties at time t (which varies significantly with respect to macroeconomic conditions and spatial/location context) for which the mortgage has been issued, AR_t is the Appraisal Reduction at time t.

AOB remains consistent, since, prepaymant penalty clauses discourage voluntary curtailment/full prepayment. SP obligations are not met both when the borrower is cash-constrained and also when the borrower chooses to strategically default. Proximity to balloon payment at maturity further complicates the endogenous behavior of the commercial borrowers towards maturity of the loan. The market value of a property is a function of the macro-economic factors like state GDP, Unemployment Rate, geographical location, 2 Year and 10 Year Treasury Rates. Until valuation is obtained, Appraisal Reduction Amount (ARA) may be calculated based on the scheduled principal balance or some other formula as defined in the servicing agreement.

NOI calculation involves the following key variables. Potential Rental Income assumes zero vacancy or could be based on a rental market analysis. Vacancy losses are realized when tenants vacate the property and/or tenants default on their lease obligations. Total Operating Expenses on an Investment Property could include ”Property Taxes, Rental Property Insurance, Property Management Fees, Maintenance and Repairs, Miscellaneous Expenses, etc. Debt service, depreciation, leasing commissions, tenant improvements, repairs to wear and tear, income taxes, and mortgage interest expenses are not included in the calculation of net operating income”. This is because NOI is property-specific devoid of other investor or borrower-specific expenses. NOI helps calculate Cap Rate (property’s potential rate of return), ROI, Debt Coverage Ratio, Cash Return on Investment. NOI provides an estimate a property’s ongoing operating revenue. NOI analysis can be manipulated from the choice to accelerate or defer certain expenses. The NOI of a property can change depending on the property management. Because other expenses are not considered in NOI, the real cash flow from a property may differ net other expenses. Further projected rents cannot be used to calculate NOI when rents differ from market rents.

.2 Multinomial Logit Model

In a Multinomial Logit Model, log-odds of each delinquency state with respect to the ”Current” state assumes a linear specification. The odds that a loan has a delinquency classes j as opposed to the baseline, depending only on individual loan-specific covariates is defined as:

$$\frac{Pr(Y_i = j|Z_i = z)}{Pr(Y_i = 0|Z_i = z)} = exp(Z^{'}\gamma_j) \tag{2}$$

the choice Y_i takes on non-negative, un-ordered integer values $Y_i \in \{0, 1, ..., J\}$. Multinomial logistic regression does not assume normality, linearity, or homoskedasticity; it has a well-behaved likelihood function, a special case of conditional logit. A more powerful alternative to multinomial logistic regression is discriminant function analysis which requires these assumptions are met. Multinomial logistic regression also assumes non-perfect separation.

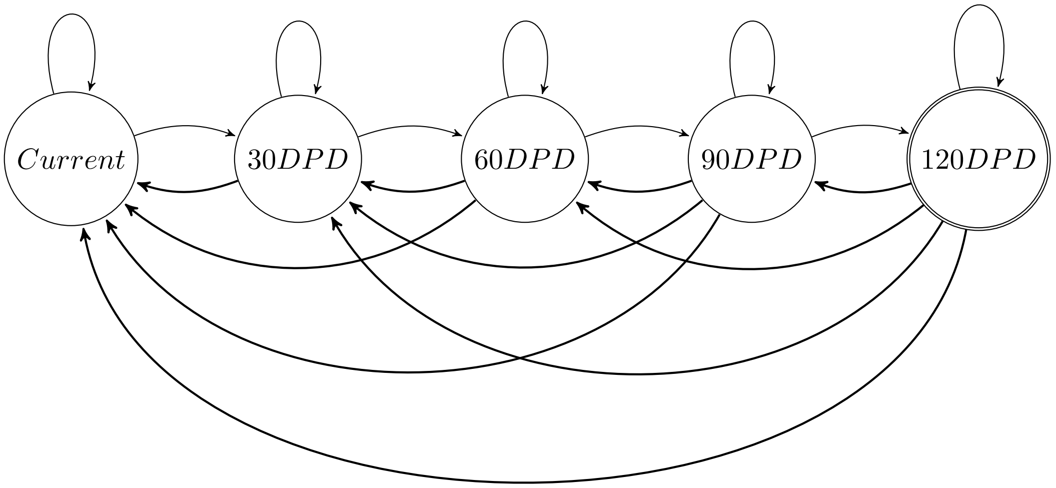
The Independence of Irrelevant Alternatives (IIA) assumption inherent in Multinomial Logit Model implies that adding or deleting alternative outcome categories does not affect the odds among the remaining outcomes.

$$Pr(Y_i = j|Y_i \in \{j, l\}) = \frac{Pr(Y_i = j)}{Pr(Y_i = j) + Pr(Y_i = l)} = \frac{exp(X_{ij}^{'}\gamma)}{exp(X_{ij}^{'}\gamma) + exp(X_{il}^{'}\gamma)} \tag{3}$$

This can be tested by the Hausman-McFadden test. There are alternative modeling methods, such as alternative-specific multinomial probit model, or nested logit model to relax the IIA assumption.

.2.1 Independence of Irrelevant Alternatives

Multinomial Logit assumes Independence of Irrelevant Alternatives (IIA). The following Finite State Automaton details all possible transitions so that the above arguments can be visualized.



Clearly, the borrower would like to stick with the first choice, as the second choice classifies him/her in the default category and is detrimental for her creditworthiness from a lender's perspective. Now suppose, one more choice for being in **30 days to 60 days of delinquency** is given to the borrower, s/he may choose to rather be in this new state instead of less than 30 days of delinquency and may **strategically** miss one payment if there is a great investment opportunity for him/her in that one month horizon. In fact, none of the models (except Naive Bayes) can distinguish these three classes (**W0_30D**, **W30_60D** & **W60_90D**) and considers all of them as **Current Loans** in Table 5.4.

The borrower can undertake this decision as she/she is already some days in delinquency and she/she wouldn't mind going to the next bucket until she/she falls in the bucket for **90 days to 120 days of delinquency**. In this situation, the borrower's creditworthiness doesn't change that much from a lender's perspective. hence, the odds for being in the **"less than 30 days delinquency"** to being in the classes of **90 days to 120 days of delinquency** will change drastically in the presence of this new choice of being in **30 days to 60 days of delinquency**. hence the IIA assumption is clearly violated.

Also the marginal effect of features towards classifying the response set does not have a clear interpretation in terms of sensitivity and directionality. We list the co-efficients of Multinomial Logit for the sake of completeness. (Table 5.2)

.3 Distributed Random Forest

Recursive partitioning, a critical data mining tool, shelps in exploring the stucture of a data set. This section provides a brief overview of CART modeling, conditional inference trees, and random forests.

Random Forests are developed by aggregating decision trees and can be used for both classification and regression. Each tree is a weak learner created from bootstrapping from subset of rows and columns. More trees will reduce the variance.It alleviates the issue of overfitting, can handle a large number of features. It shelps with feature selection based on importance. It is user-friendly with two parameters: number of trees (default 500) and variables randomly selected as candidates at each split, $\sqrt{ntree}$ for classification and $ntree/3$ for regression. "Out Of Bag Error" is estimated for each bootstrap iteration and related tree.

R's randomForest splits based on the Gini criterion and H2O trees are split based on reduction in Squared Error (even for classification). H2O also uses histograms for splitting and can handle splitting on categorical variables without dummy (or one-hot) encoding. Also, R's randomForest builds really deep trees, resulting in pure leaf nodes, leading to constant increments in prediction and ties and hence relatively lower AUC.The trees in H2O's random forest aren't quite as deep and therefore aren't as pure, allowing for predictions that have some more granularity to them and that can be better sorted for a better AUC score.

CART models an outcome y_i for an instance i as:

$$y_i = f(x_i) = \sum_{m=1}^{M} c_m I x_i \in R_m \tag{4}$$

where each observation x_i belongs to exactly one subset R_m, c_m is the mean of all training observations in R_m.

.4 Gradient Boosting Machine

GBM Friedman 2000 creates an ensemble Kuncheva 2003 of weak prediction models in stages and utilizes a differentiable loss function. Boosting trees does increase accuracy, but at the cost of speed and meaningful interpretability.

At each step m, $1 \leq m \leq M$ of gradient boosting, an estimator h_m is computed from the residuals of the previous model predictions. Friedman (2001) proposed regularization by shrinkage:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x) \tag{5}$$

where $h_m(x)$ represents a weak learner of fixed depth, γ_m is the step length and ν is the learning rate or the shrinkage factor. XGBoost Chen and Guestrin 2016 is a faster and more accurate implementation of the Gradient Boosting algorithm Chen, Lundberg, and Lee 2018.

.5 Deep Neural Network

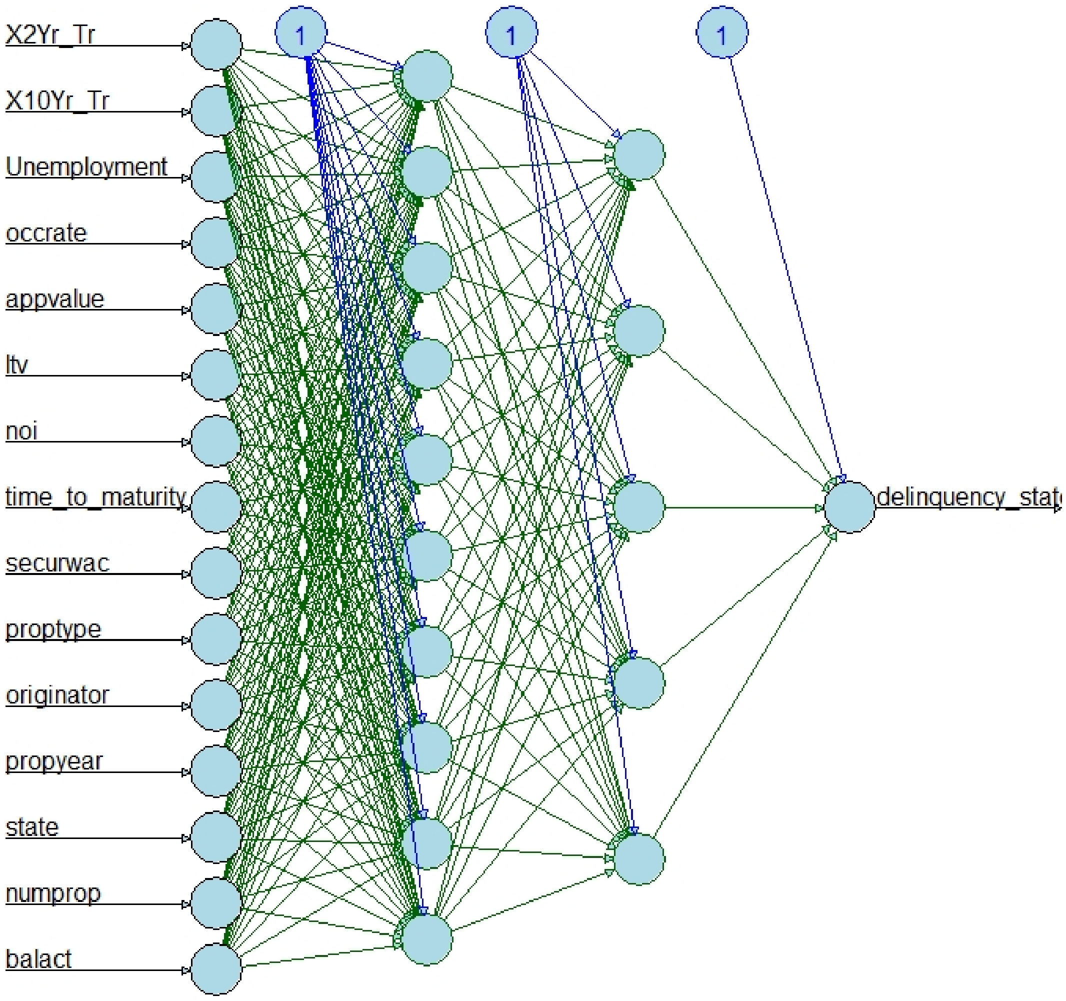


Figure 12: Deep Neural Network

.5.1 Deep Neural Network for CMBS

The purpose for a Deep Neural Network (DNN) is bourne out of the need to have transparency and accountability Albanesi and Vamossy 2019. By the very nature of the DNN, we do not have to add interaction terms in the specification of the model, especially in the case of high dimensional data. The sequential layers embody highly non-linear and non-trivial interaction among the variables and capture several latent fundamental features in the process. The causal interpretation of the covariates both in default Kvamme et al. 2018 and prepayment calculations have not been explored in details. The broader impact could be traced out by improved allocation of credit and aid in policy design (macroprudential, bankruptcy, foreclosure, etc.).

With the provision of enough hidden units, a neural network can mimic continuous functions on closed and bounded sets really well Hornik 1991, vis-a-vis the product and division of relevant features and their interactions. More layers, and not more units in each layer, learns atures of greater complexity. Deep neural networks 12, with three or more hidden layers, require exponentially fewer units than shallow networks or logistic regressions with basis functions; see Montufar et al. 2014 and Goodfellow, Bengio, and Courville 2016.

Under certain regularity conditions, the estimators are consistent and also asymptotically normal (see Hornik, Stinchcombe, and White 1989, Sussmann 1992 and Albertini and Sontag 1993 where they study the identifiability). One can regularize DNN using optimal hyper-parameter tuning via cross-validation and ”drop out” some sample values to reduce overfitting Srivastava et al. 2014.

Deep Neural Networks (DNN) **??** have an extensive set of current applications like: System identification and control (e.g., vehicle control, trajectory prediction, etc.), Game-playing and decision making (e.g., chess, poker, etc.), Pattern and sequence recognition (e.g., radar systems, face identification, signal classification, speech/image recognition, etc.), Medial diagnosis and finance (e.g., automated trading systems, cancer diagnosis, etc.).

.5.2 Hyperparameter Tuning and Grid Search

Hyper-parameter tuning with Random Grid Search (RGS) tests different combinations of hyper-parameters to find the optimal choice based on accuracy, without overfitting.

Hyperparameters can be divided into 2 categories:

- **Optimizer hyperparameters**
- **Model Specific hyperparameters**

Our model hyperparameters are: *score_training_samples* = 6125796, *epoch* = 60000, *hidden* = c(30,20,10,7), *hidden_dropout_ratios* = c(0.01, 0.01, 0.01,0.01), *momentum_start* = 0.5, *momentum_ramp* = 100, *momentum_stable* = 0.99, *missing_values_handling* = ”Skip”, *initial_weight_distribution* =”Uniform”, *nesterov_accelerated_gradient* = TRUE, *activation* = ”RectifierWithDropout”, *nfolds* = 10,*fold_assignment* = ”Stratified”, *keep_cross_validation_predictions* = FALSE, *variable_importances* = TRUE, *adaptive_rate* = FALSE, *l1* =1e-5, *l2* =1e-5, *export_weights_and_biases* = FALSE, *mini_batch_size* = 128, *loss* = ”CrossEntropy”, *distribution* = ”AUTO”, *balance_classes* = T, *max_after_balance_size* = 1, *rate* = 05, *rate_annealing* = 1e-06, *rate_decay* = 1, *stopping_metric* = ”MSE”, *seed* = 1122.

.5.3 Class Imbalance Problem

Most classifiers are unable to distinguish minor classes Kuncheva 2003 and are sheavily influenced by major classes, e.g., the conditional probability of minor classes are underestimated in a logistic regression King and Zeng 2001, Tree based classifiers, and KNN yield high recall but low sensitivity when the data set is extremely unbalanced Daelemans, Goethals, and Morik 2008. There are a plethora of techniques to balance the data, e.g., oversampling, under-sampling and Synthetic Minority Oversampling Technique (SMOTE) proposed by Chawla et al. 2002.

.6 Cost of Misaligned CARES Act: Overcrowding, Selective Verification and Unintended Racial Consequences

.6.1 CARES Act

Key words chosen manually

Keywords for identifying **employment** issues are: “unemployed”, ”out of work”, ”laid off”, ”furlough”, ”unable to achieve”, “collactivitie” and “suspended”, “insufficient” and “conditional liquidation”, “ADV” and “late charge”, “not available” and “payment”, “loss” and “request”, “lost job”, and “didn’t” and “work”.

Keywords for identifying **Inbound** communications are: ”Inbound”, ”IB”, ”reached out to” and not ”borrower”, ”received”, ”Borrower is writing”, ”was contacted”, ”borrower” and ”informed” or ”indicated”, ”request”, ”marital difficulties”, ”death of family member”, ”excessive obligations”, ”casualty loss”, ”payment dispute”, ”tenant not paying”, ”prior bk” and not ”OB” and not ”Outbound”.

Keywords for identifying **Outbound** communications are: ”OB”, ”Outbound”, ”COVID19 Forbearance Letter 712 Requested from vendor”, ”Asked”, ”Replied”, ”msg in dmm portal to da”, ”Called borrower”, ”LMStatus:”, ”No Contact”, ”email reply back”, ”Good Morning”, ”CMS encourages”, ”Carrington Mortgage Services authorizes”, ”next due”, ”response”, ”responded”, ”decline”, ”CMS representative”, ”will be asked”, ”CMS is committed”, ”was contacted”, ”LM Program”, ”payment dispute”, ”prior bk”, ”illness”, ”Offered Borrower” and not ”IB” and not ”Inbound”.

.6.2 Differential Borrower Behavior

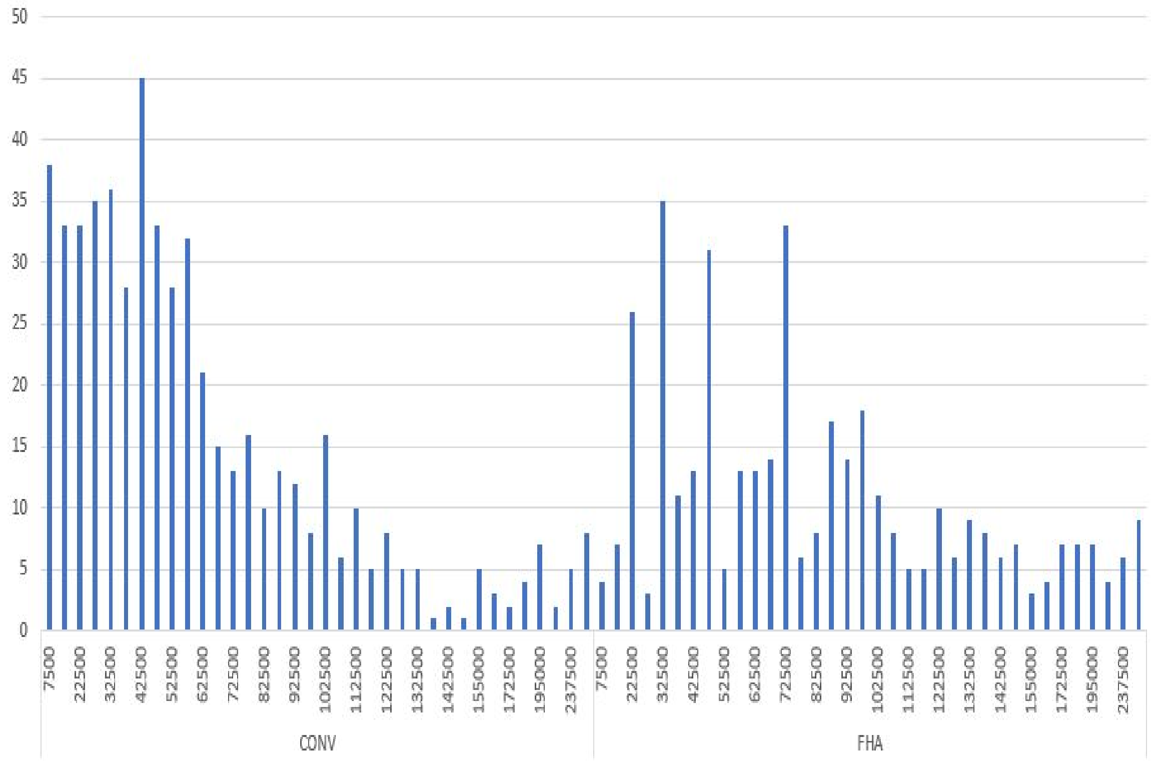


Figure 13: Forbearance applications and Loan Count across different median incomes in US Dollars.

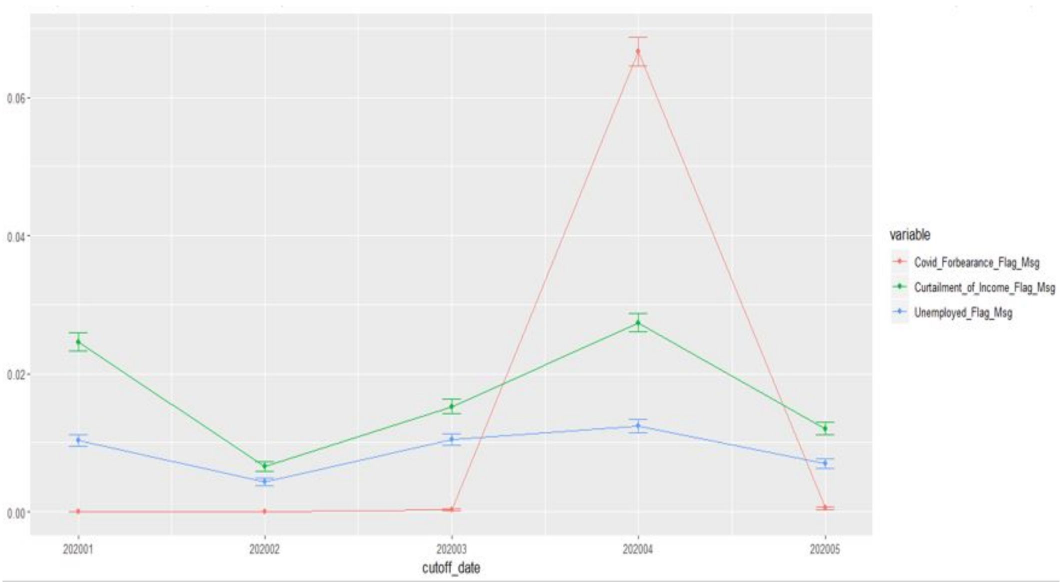


Figure 14: Performing Loans: Time Trend of COVID, Unemployed Curtailment of Income Flags.

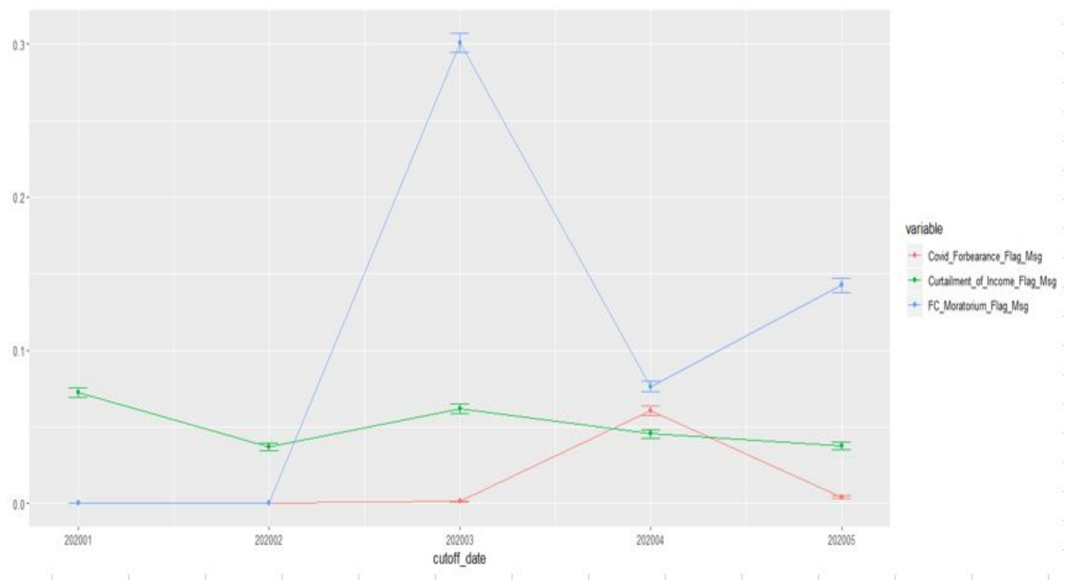


Figure 15: Non-Performing Loans: Time Trend of COVID, Unemployed Curtailment of Income Flags.

.6.3 Keywords from NLP on data

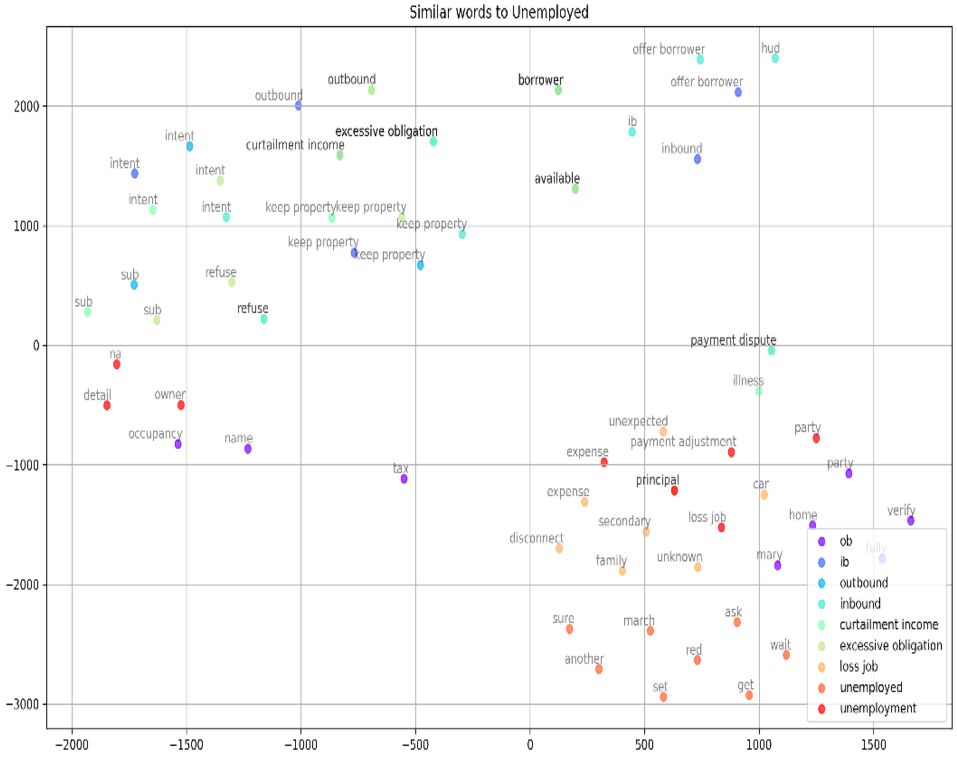


Figure 16: Words related to Unemployment, Inbound and Outbound.

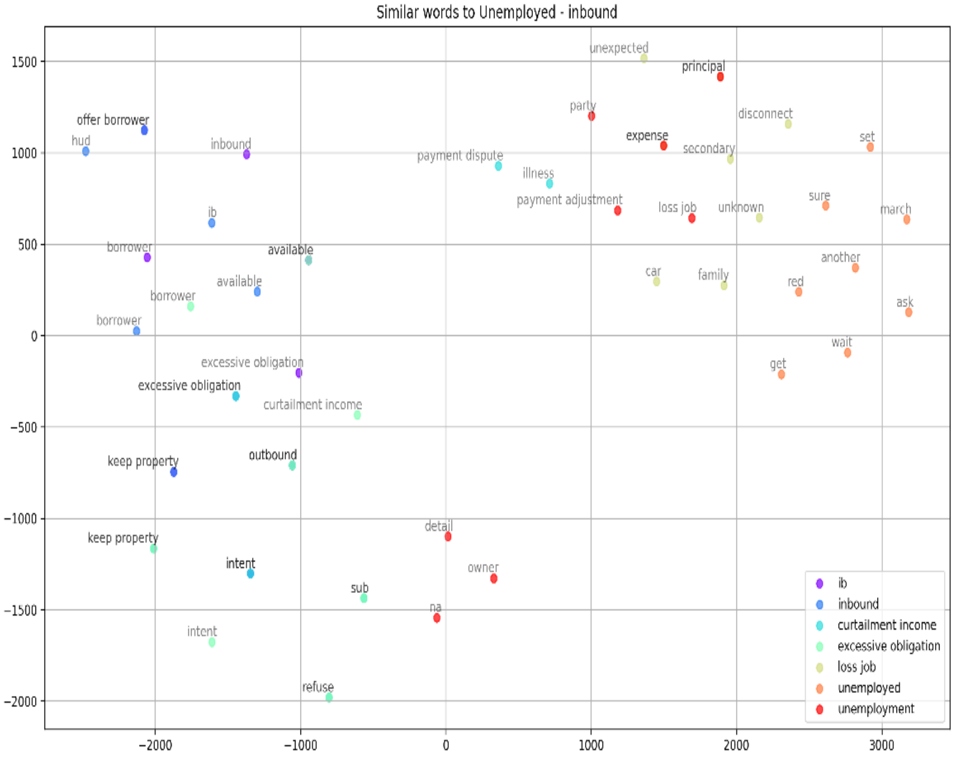


Figure 17: Words related to Unemployment and Inbound.

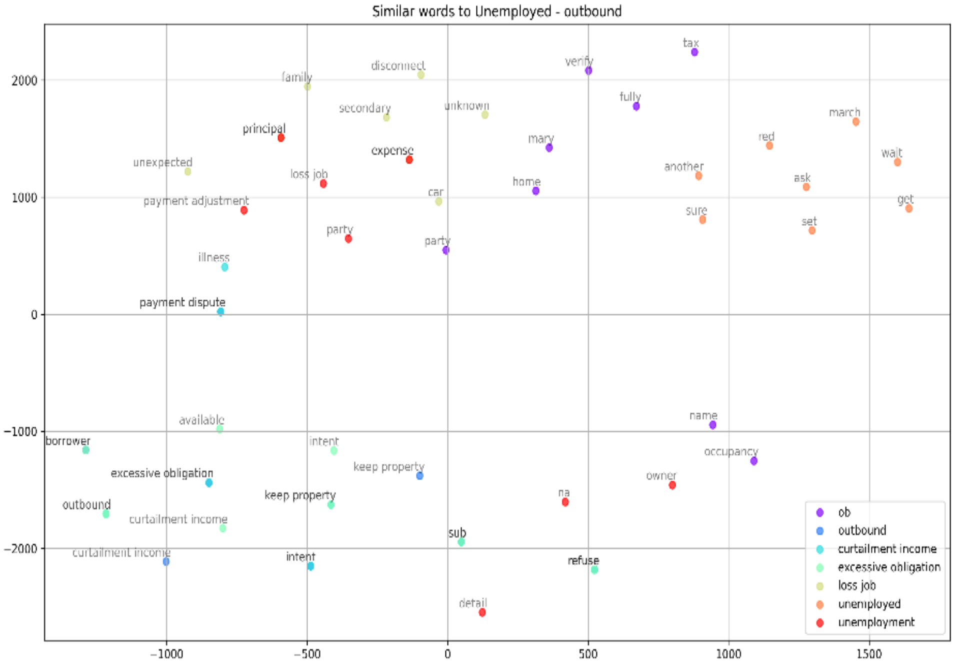


Figure 18: Words related to Unemployment and Outbound.

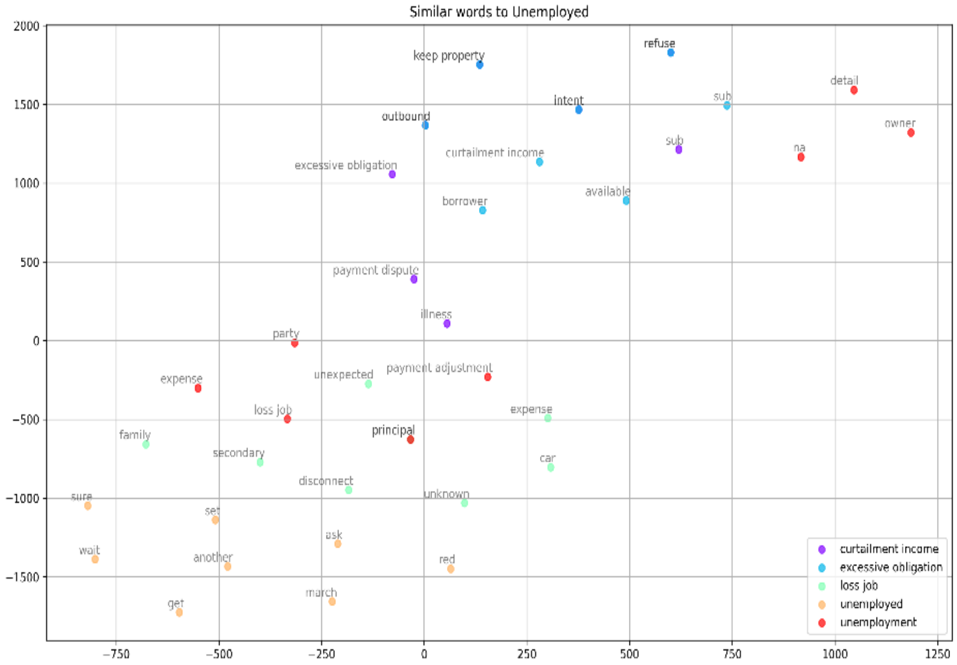


Figure 19: Words related to Unemployment only (without Inbound and Outbound).

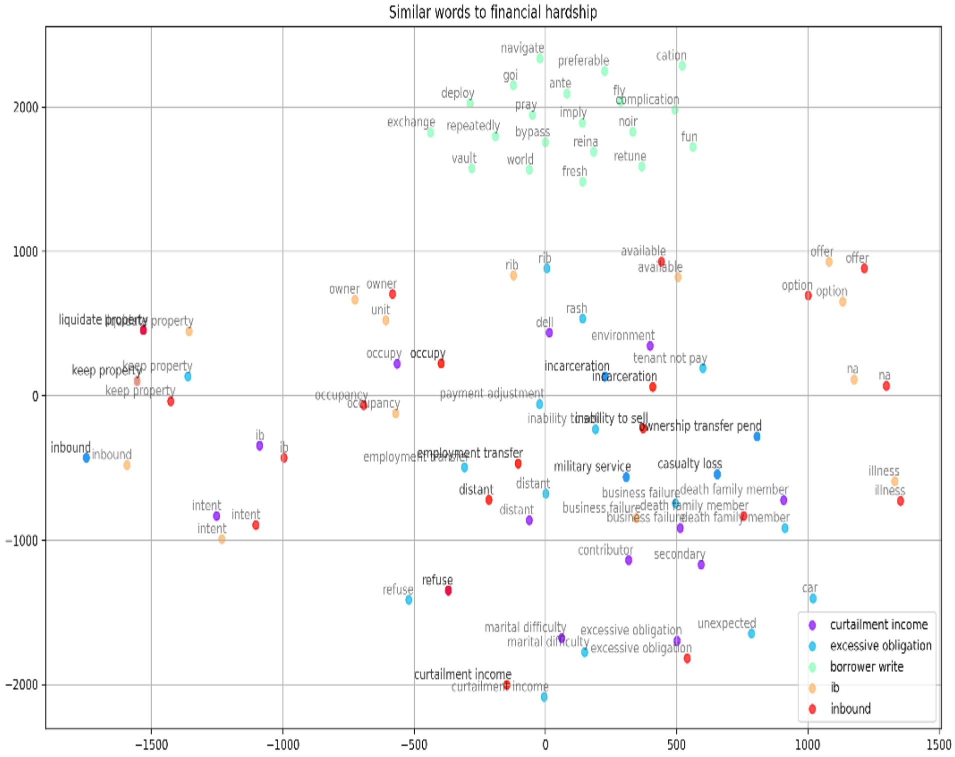


Figure 20: Inbound Communications and Financial Hardship.

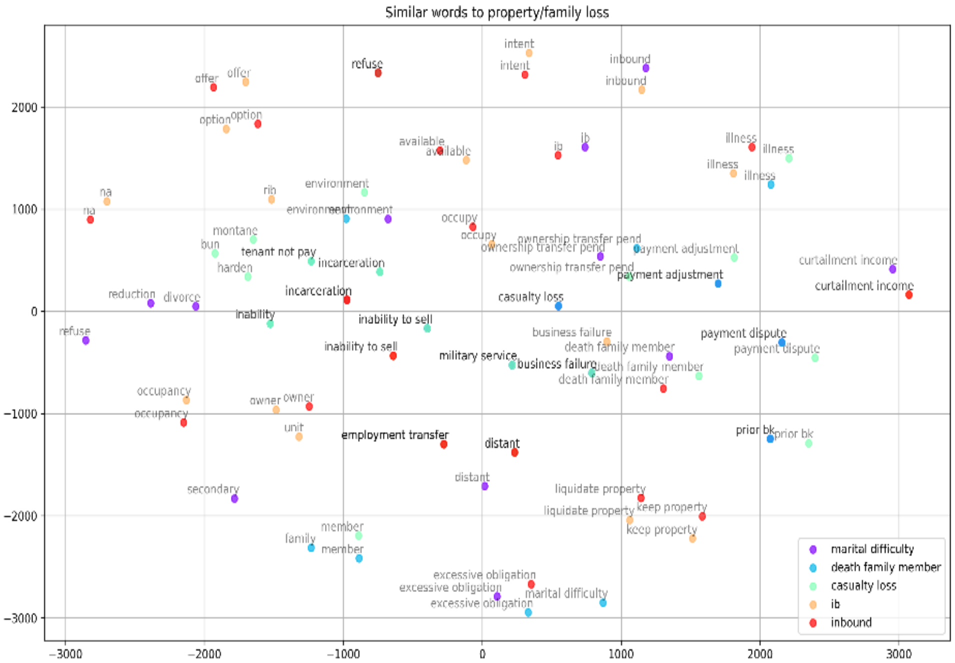


Figure 21: Inbound Communications, Family, Property and Loss.

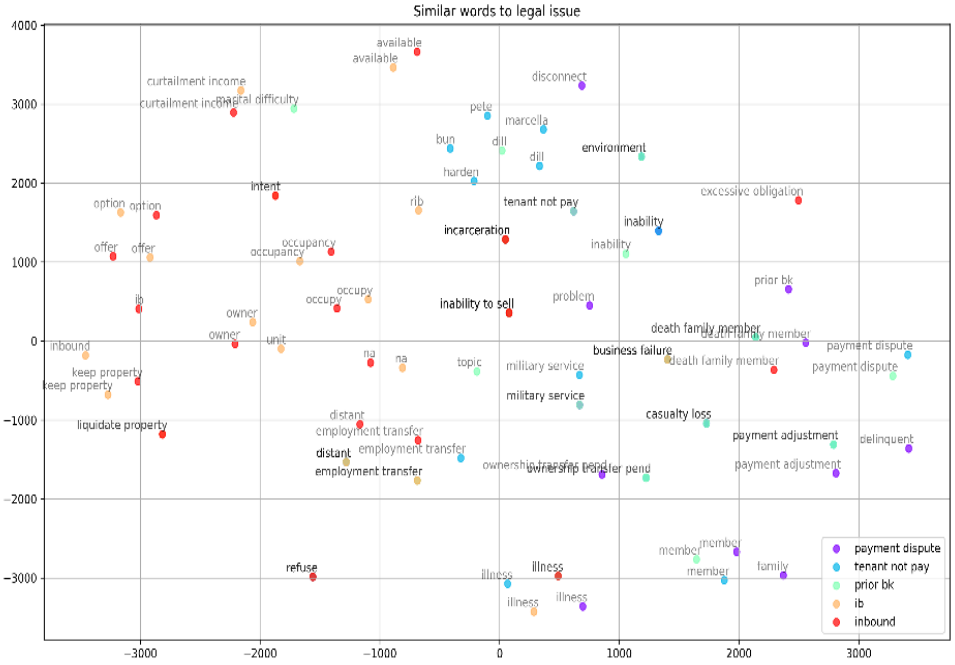


Figure 22: Inbound Communications and Legal Issues.

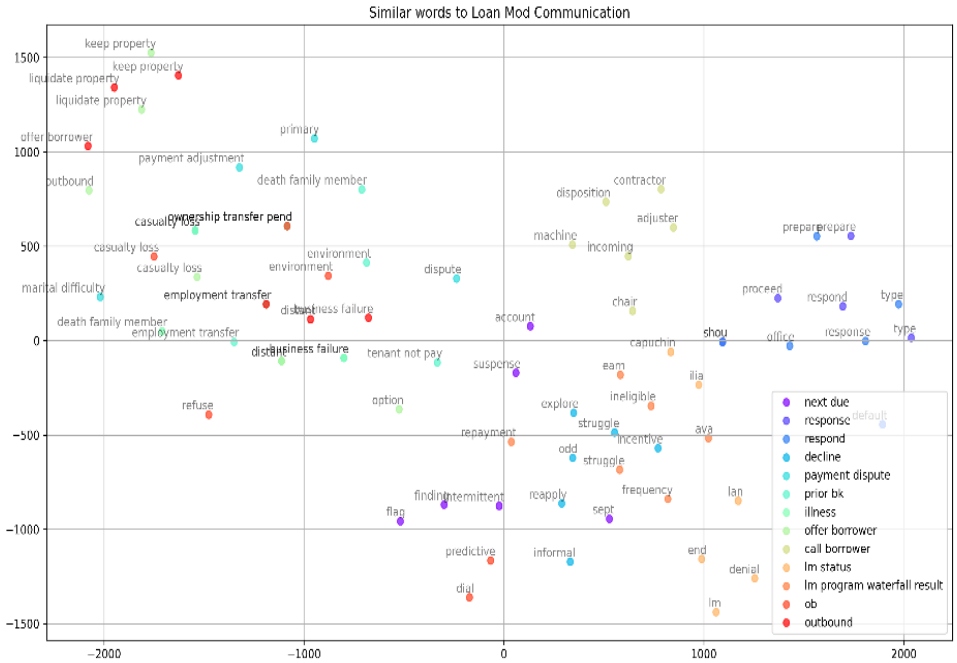


Figure 23: Outbound Communications and Loan Modification.

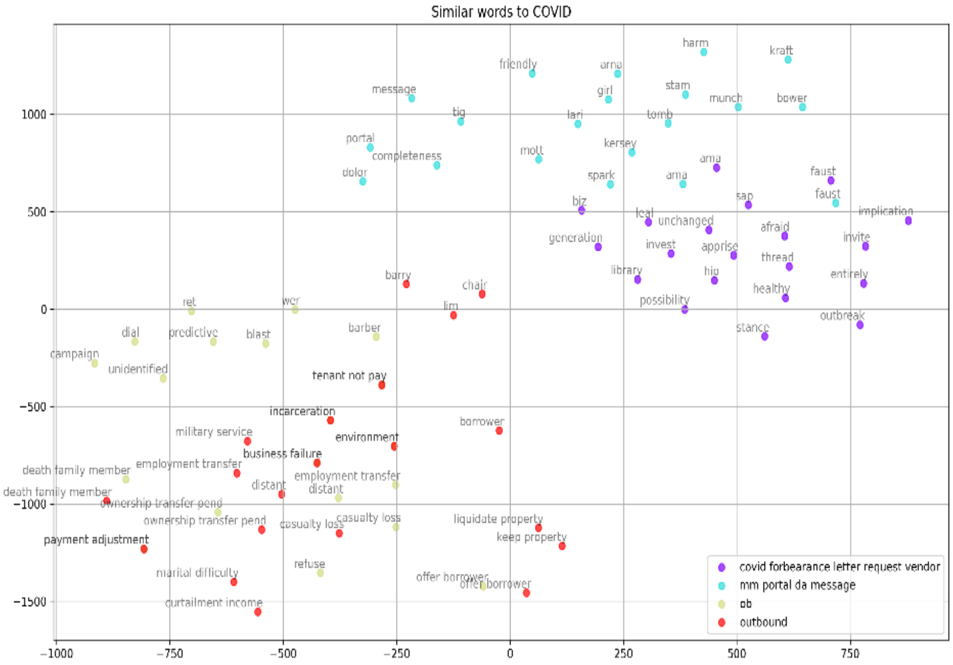


Figure 24: Outbound Communications and COVID.

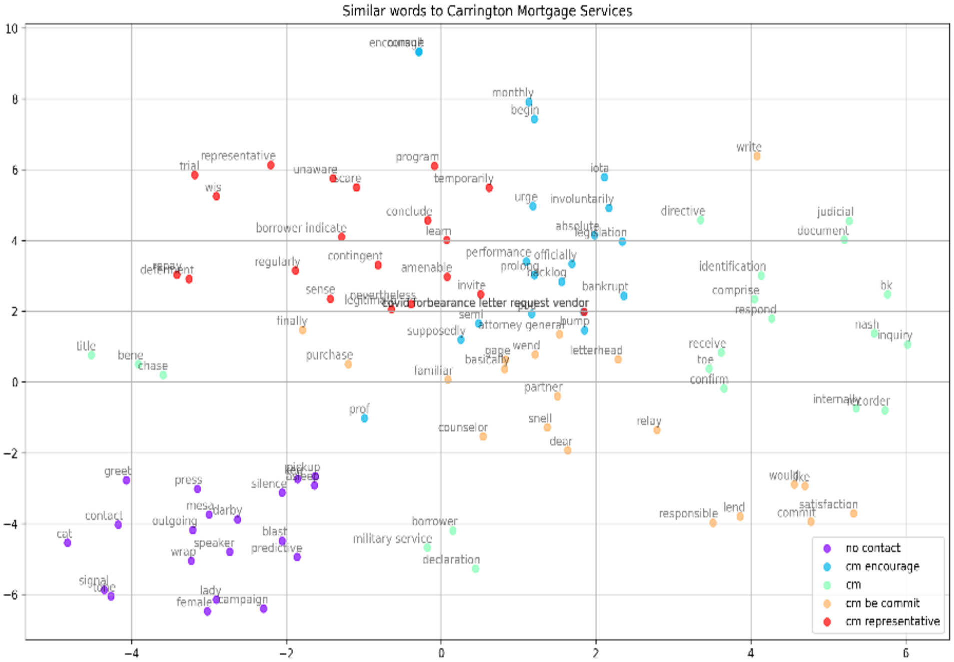


Figure 25: Outbound Communications related to Servicer.

.6.4 US Map Spatial distribution for key variables

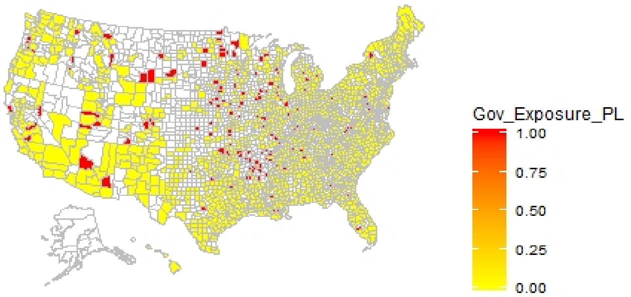


Figure 26: Percentage of Govt. loan Exposure by County in April 2020 among Performing Loans.

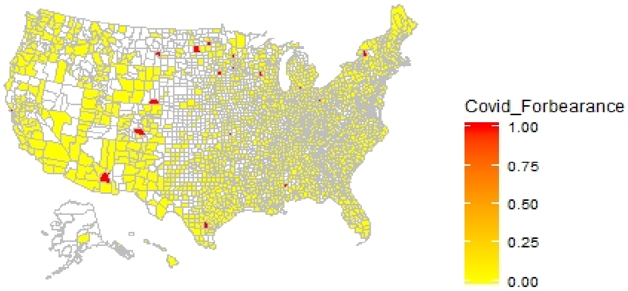


Figure 27: Percentage of COVID Forbearance Applications by County in April 2020.

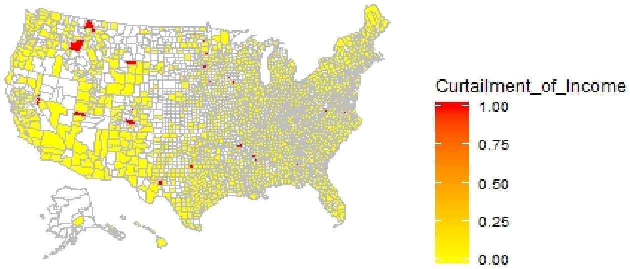


Figure 28: Percentage of Curtailment of Income by County in April 2020.

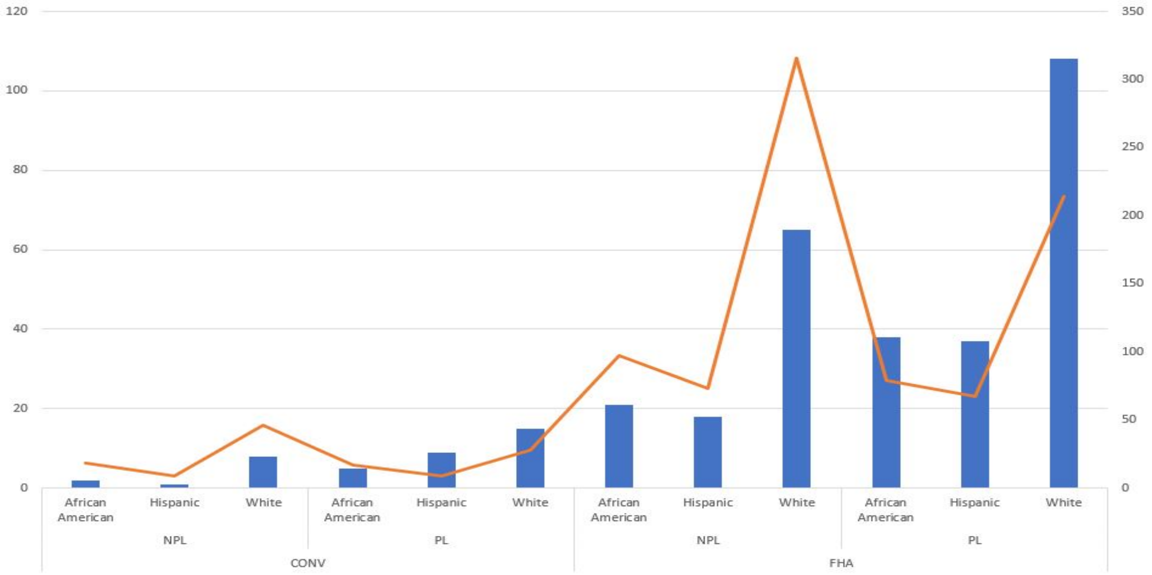
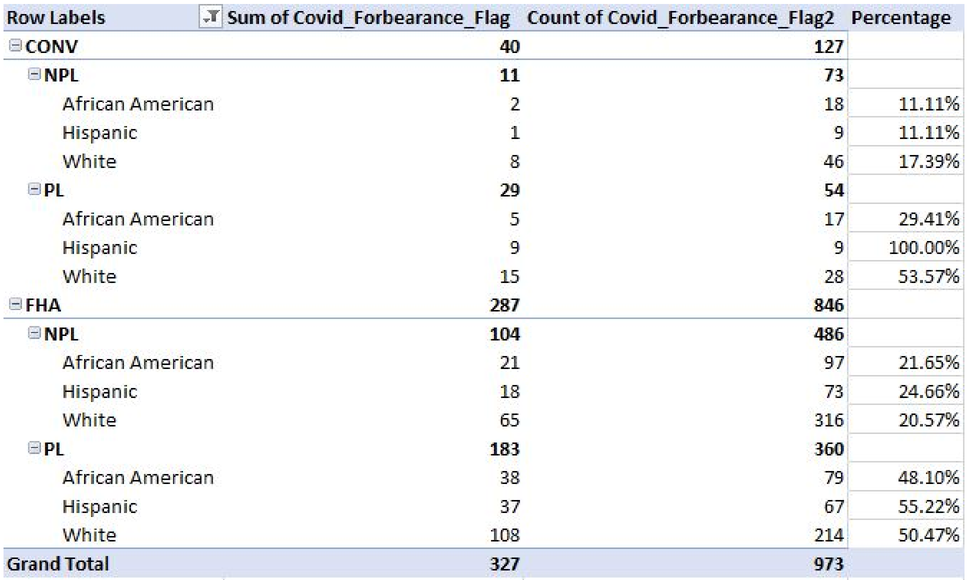


Figure 29: Inbound Forbearance Communications by Delinquency Status and Race in April 2020.



Row Labels	Sum of Covid_Forbearance_Flag	Count of Covid_Forbearance_Flag2	Percentage
⊟CONV	**40**	**127**	
⊟NPL	**11**	**73**	
African American	2	18	11.11%
Hispanic	1	9	11.11%
White	8	46	17.39%
⊟PL	**29**	**54**	
African American	5	17	29.41%
Hispanic	9	9	100.00%
White	15	28	53.57%
⊟FHA	**287**	**846**	
⊟NPL	**104**	**486**	
African American	21	97	21.65%
Hispanic	18	73	24.66%
White	65	316	20.57%
⊟PL	**183**	**360**	
African American	38	79	48.10%
Hispanic	37	67	55.22%
White	108	214	50.47%
Grand Total	**327**	**973**	

Figure 30: Inbound Forbearance Communications Table by Delinquency Status and Race in April 2020.

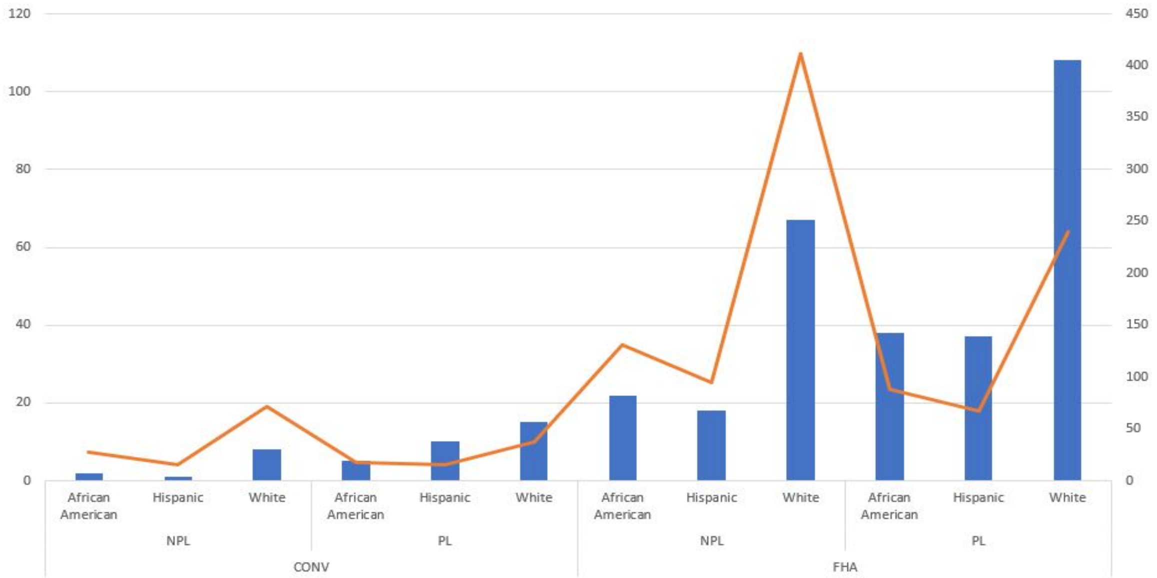
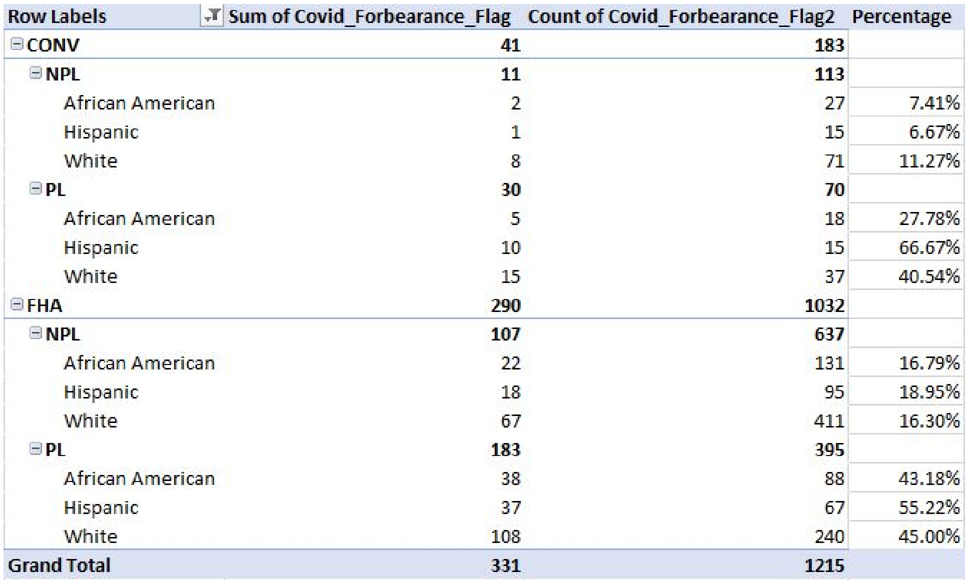
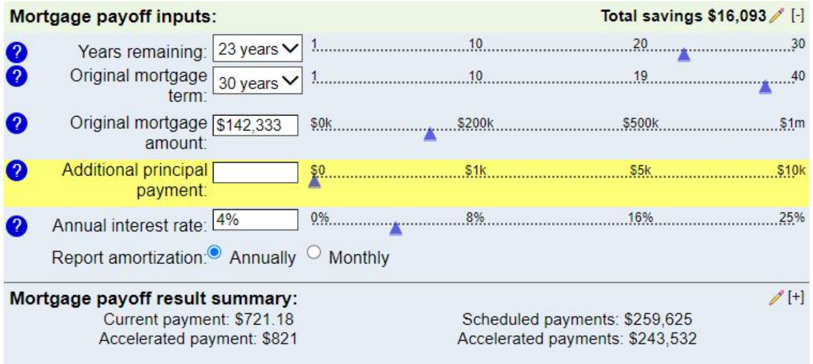


Figure 31: Outbound Forbearance Communications by Delinquency Status and Race in April 2020.



Row Labels	Sum of Covid_Forbearance_Flag	Count of Covid_Forbearance_Flag2	Percentage
⊟CONV	41	183	
⊟NPL	11	113	
African American	2	27	7.41%
Hispanic	1	15	6.67%
White	8	71	11.27%
⊟PL	30	70	
African American	5	18	27.78%
Hispanic	10	15	66.67%
White	15	37	40.54%
⊟FHA	290	1032	
⊟NPL	107	637	
African American	22	131	16.79%
Hispanic	18	95	18.95%
White	67	411	16.30%
⊟PL	183	395	
African American	38	88	43.18%
Hispanic	37	67	55.22%
White	108	240	45.00%
Grand Total	331	1215	

Figure 32: Outbound Forbearance Communications Table by Delinquency Status and Race in April 2020.



Mortgage payoff inputs: **Total savings $16,093** [-]

Years remaining: 23 years

Original mortgage term: 30 years

Original mortgage amount: $142,333

Additional principal payment:

Annual interest rate: 4%

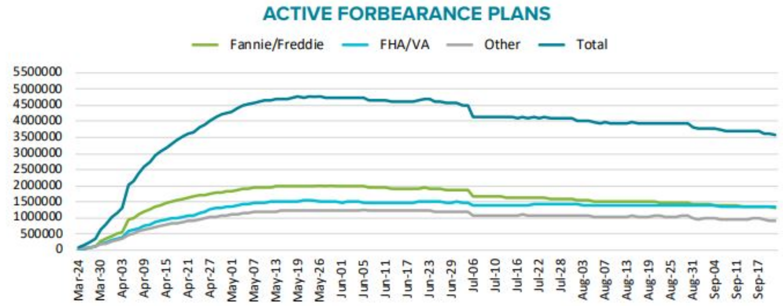
Report amortization: ● Annually ○ Monthly

Mortgage payoff result summary: [+]

Current payment: $721.18
Accelerated payment: $821

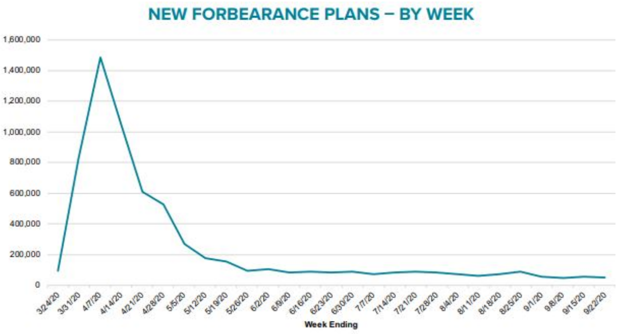
Scheduled payments: $259,625
Accelerated payments: $243,532

Figure 33: Average Mortgage Principal and Interest for Gov-backed loans.



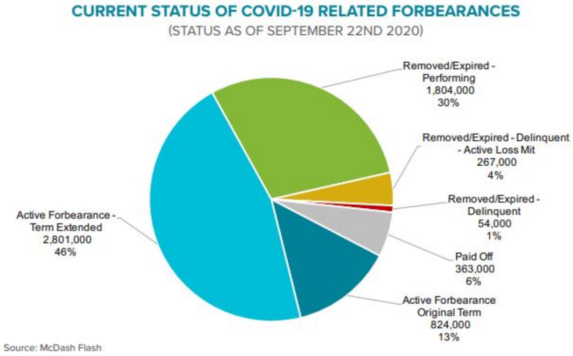
Source: McDash Flash

Figure 34: National Active Forbearance Plans.



Source: McDash Flash

Figure 35: National New Forbearance Plans - by week.



Source: McDash Flash

Figure 36: Status of COVID related Forbearance in August 2020.

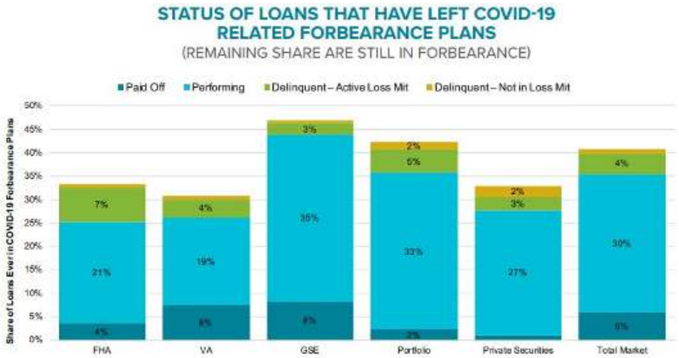


Figure 37: Status of Loans leaving COVID-19 related Forbearance Plans.

Table 18: Summary Statistics: Data for April 2020 performance report (all loans)

PANEL A: All Loans

PANEL A	count	mean	sd	median	min	max	range	skew	kurtosis
Covid_Forbearance_Flag	19159	6.74%	0.251	0	0	1	1	3.451	9.911
Unemployed_Flag	19159	3.48%	0.183	0	0	1	1	5.079	23.8
Prior_Unemployment_Flag	19159	0.64%	0.08	0	0	1	1	12.359	150.755
FC_Moratorium_Flag	19159	6.74%	0.251	0	0	1	1	3.451	9.911
Curtailment_of_Income_Flag	19159	4.86%	0.215	0	0	1	1	4.198	15.628
Inbound_Borrower_Flag	19159	22.31%	0.416	0	0	1	1	1.33	-0.23
Outbound_Borrower_Flag	19159	29.38%	0.455	0	0	1	1	0.906	-1.18
Dlq	19159	1.429	2.049	0	0	5	5	0.987	-0.837
original_balance	19159	96791.008	84115.159	70947.24	1972.51	747750	745777.49	2.353	8.299
original_appraisal	19159	119747.758	97459.965	90000	3122.24	978200	975077.76	2.821	11.993
original_fico	19159	612.567	67.522	613	372	847	475	0.244	-0.13
current_rate	19159	0.066	0.029	0.058	0	0.197	0.197	0.427	-1.026
orig_ltv	19159	80.504	23.657	88.585	2.704	139.349	136.645	-1.082	0.527
log(current_balance)	19159	10.801	1.08	10.898	0.01	13.584	13.574	-0.692	1.589
Gov	19159	0.348	0.476	0	0	1	1	0.64	-1.591
corporate_adv	19159	6.897	1.767	7.174	0.482	12.197	11.715	-1.34	3.213
rem_term	19159	190.817	112.797	202	6	526	520	0.061	-1.032
mod_flag	19159	0.354	0.478	0	0	1	1	0.611	-1.627
N	19159								

PANEL B: Government Loans

Covid_Forbearance_Flag	6660	11.13%	0.31	0	0	1	1	2.47	4.11
Unemployed_Flag	6660	5.23%	0.22	0	0	1	1	4.02	14.19
Prior_Unemployment_Flag	6660	1.07%	0.1	0	0	1	1	9.53	88.79
FC_Moratorium_Flag	6660	12.57%	0.33	0	0	1	1	2.26	3.1
Curtailment_of_Income_Flag	6660	7.76%	0.27	0	0	1	1	3.16	7.96
Inbound_Borrower_Flag	6660	32.43%	0.47	0	0	1	1	0.75	-1.44
Outbound_Borrower_Flag	6660	41.80%	0.49	0	0	1	1	0.33	-1.89
Dlq	6660	2.26	2.17	1	0	5	5	0.25	-1.73
original_balance	6660	142333.67	78248.12	126424	21825	730987	709162	1.78	5.35
original_appraisal	6660	149441.5	84418.46	131000	23000	890000	867000	1.96	6.66
original_fico	6660	617.38	65.13	628	376	813	437	-0.02	-0.21
current_rate	6660	0.04	0.01	0.04	0.01	0.1	0.09	1.76	4.47
orig_ltv	6660	95.92	7.55	98.11	9.97	130.77	120.79	-2.83	13.79
log(current_balance)	6660	11.51	0.71	11.59	0.01	13.41	13.4	-1.8	15.37
corporate_adv	6660	5.95	2.41	6.21	0.48	11.47	10.99	-0.58	-0.07
rem_term	6660	274.79	61.87	283	7	446	439	-1.44	2.72
mod_flag	6660	0.69	0.46	1	0	1	1	-0.83	-1.31
N	6660								

PANEL C: Non-Government Loans

Covid_Forbearance_Flag	12499	4.40%	0.21	0	0	1	1	4.45	17.77
Unemployed_Flag	12499	2.54%	0.16	0	0	1	1	6.03	34.33
Prior_Unemployment_Flag	12499	0.42%	0.06	0	0	1	1	15.4	235.33
FC_Moratorium_Flag	12499	3.63%	0.19	0	0	1	1	4.96	22.56
Curtailment_of_Income_Flag	12499	3.31%	0.18	0	0	1	1	5.22	25.22
Inbound_Borrower_Flag	12499	16.91%	0.37	0	0	1	1	1.77	1.12
Outbound_Borrower_Flag	12499	22.75%	0.42	0	0	1	1	1.3	-0.31
Dlq1	12499	0.99	1.83	0	0	5	5	1.56	0.65
original_balance	12499	72523.93	76741.5	50993.05	1972.51	747750	745777.49	3.63	17.55
original_appraisal	12499	103925.67	100213.71	74000	3122.24	978200	975077.76	3.48	16.07
original_fico	12499	610	68.63	605	372	847	475	0.38	-0.05
current_rate	12499	0.08	0.03	0.08	0	0.2	0.2	-0.29	-0.69
orig_ltv1	12499	72.29	25.17	77.54	2.7	139.35	136.65	-0.53	-0.28
log(current_balance)	12499	10.42	1.05	10.47	3	13.58	10.58	-0.38	1.48
corporate_adv	12499	7.4	0.98	7.37	5.02	12.2	7.18	0.74	0.9
rem_term	12499	146.07	108.18	125	6	526	520	0.91	0.31
mod_flag	12499	0.17	0.38	0	0	1	1	1.72	0.96
N	12499								

Table 19: Summary Statistics: Data for April 2020 performance report (granular)

	count	mean	sd	median	min	max	range	skew	kurtosis
PANEL A: Performing and Govt-backed loans									
Covid_Forbearance_Flag	3657	13.18%	0.34	0	0	1	1	2.18	2.74
Unemployed_Flag	3657	1.48%	0.12	0	0	1	1	8.04	62.7
Prior_Unemployment_Flag	3657	0.57%	0.08	0	0	1	1	13.08	169.05
FC_Moratorium_Flag	3657	0.16%	0.04	0	0	1	1	24.62	604.17
Curtailment_of_Income_Flag	3657	4.73%	0.21	0	0	1	1	4.26	16.18
Inbound_Borrower_Flag	3657	24.99%	0.43	0	0	1	1	1.15	-0.67
Outbound_Borrower_Flag	3657	27.89%	0.45	0	0	1	1	0.99	-1.03
Dlq4	3657	0.53	1.03	0	0	5	5	3.26	11.37
original_balance	3657	134613.71	74617.33	118907	21825	653015	631190	1.73	4.72
original_appraisal	3657	141636.78	81805.48	125000	23000	890000	867000	2.06	7.35
original_fico	3657	615.57	66.86	625	376	813	437	0.04	-0.27
current_rate	3657	0.04	0.01	0.04	0.02	0.1	0.08	1.93	6.45
orig_ltv	3657	95.89	7.71	98.16	9.97	124.39	114.42	-3.07	16.2
current_balance	3657	110525.8	70592.05	96412.37	522.02	669748.08	669226.06	1.61	4.53
corporate_adv	3657	603.63	2370.6	187.44	0.62	95760.35	95759.73	25.17	853.19
rem_term	3657	271.36	66.4	281	9	446	437	-1.44	2.36
mod_flag	3657	0.73	0.44	1	0	1	1	-1.05	-0.89
N	3657								
PANEL B: Performing and Non-Govt-backed loans									
Covid_Forbearance_Flag	10157	4.54%	0.21	0	0	1	1	4.37	17.08
Unemployed_Flag	10157	1.19%	0.11	0	0	1	1	9	78.94
Prior_Unemployment_Flag	10157	0.30%	0.05	0	0	1	1	18.32	333.5
FC_Moratorium_Flag	10157	0.01%	0.01	0	0	1	1	100.75	10150
Curtailment_of_Income_Flag	10157	2.64%	0.16	0	0	1	1	5.91	32.92
Inbound_Borrower_Flag	10157	13.22%	0.34	0	0	1	1	2.17	2.71
Outbound_Borrower_Flag	10157	14.43%	0.35	0	0	1	1	2.02	2.1
Dlq	10157	0.2	0.77	0	0	5	5	5.33	29.86
original_balance	10157	65724.2	61534.64	49605.26	2025.39	709600	707574.61	3.57	19.04
original_appraisal	10157	94719.04	83717.5	70000	3122.24	978200	975077.76	3.61	18.49
original_fico	10157	613.93	68.78	610	402	847	445	0.35	-0.04
current_rate	10157	0.08	0.03	0.09	0	0.2	0.2	-0.4	-0.52
orig_ltv	10157	72.22	24.81	77.05	2.7	139.35	136.65	-0.47	-0.34
current_balance	10157	48789.55	56157.64	32970.4	19.14	699407.76	699388.62	3.65	20.23
Gov	10157	0	0	0	0	0	0	NA	NA
corporate_adv	10157	1867.42	2898.92	1404.94	150	99951.25	99801.25	14.74	364.16
rem_term	10157	140.73	105.79	119	6	494	488	0.96	0.45
mod_flag	10157	0.16	0.36	0	0	1	1	1.88	1.52
N	10157								
PANEL C: Non-Performing and Govt-backed loans									
Covid_Forbearance_Flag	3003	8.62%	0.28	0	0	1	1	2.95	6.68
Unemployed_Flag	3003	9.79%	0.3	0	0	1	1	2.7	5.32
Prior_Unemployment_Flag	3003	1.67%	0.13	0	0	1	1	7.55	55.04
FC_Moratorium_Flag	3003	27.67%	0.45	0	0	1	1	1	-1.01
Curtailment_of_Income_Flag	3003	11.46%	0.32	0	0	1	1	2.42	3.85
Inbound_Borrower_Flag	3003	41.49%	0.49	0	0	1	1	0.35	-1.88
Outbound_Borrower_Flag	3003	58.74%	0.49	1	0	1	1	-0.35	-1.87
Dlq	3003	4.37	1.02	5	2	5	3	-1.46	0.71
original_balance	3003	151734.91	81484.54	135695	24600	730987	706387	1.81	5.79
original_appraisal	3003	158945.96	86563.52	140500	25000	800000	775000	1.88	6.13
original_fico	3003	619.58	62.89	631	392	810	418	-0.11	-0.12
current_rate	3003	0.04	0.01	0.04	0.01	0.09	0.08	1.57	2.71
orig_ltv	3003	95.95	7.35	98.05	38.16	130.77	92.61	-2.5	10.12
current_balance	3003	136656.28	81399.43	121333.65	0.01	665499.18	665499.17	1.68	4.87
corporate_adv	3003	4969.19	8068.78	2128.66	7.03	89295.04	89288.01	4.02	23.74
rem_term	3003	278.97	55.59	284	7	418	411	-1.33	2.76
mod_flag	3003	0.64	0.48	1	0	1	1	-0.59	-1.65
N	3003								
PANEL D: Non-Performing and Non-Govt-backed loans									
Covid_Forbearance_Flag	2342	3.80%	0.19	0	0	1	1	4.83	21.33
Unemployed_Flag	2342	8.41%	0.28	0	0	1	1	2.99	6.97
Prior_Unemployment_Flag	2342	0.94%	0.1	0	0	1	1	10.17	101.37
FC_Moratorium_Flag	2342	19.34%	0.4	0	0	1	1	1.55	0.41
Curtailment_of_Income_Flag	2342	6.23%	0.24	0	0	1	1	3.62	11.1
Inbound_Borrower_Flag	2342	32.92%	0.47	0	0	1	1	0.73	-1.47
Outbound_Borrower_Flag	2342	58.84%	0.49	1	0	1	1	-0.36	-1.87
Dlq	2342	4.39	1.05	5	2	5	3	-1.48	0.64
original_balance	2342	102013.65	118082.73	59776.37	1972.51	747750	745777.49	2.57	7.06
original_appraisal	2342	143853.8	145765.45	89950	13070	975000	961930	2.5	7.11
original_fico	2342	592.96	65.29	585	372	845	473	0.51	0.03
current_rate	2342	0.07	0.03	0.07	0	0.15	0.15	0.13	-1.04
orig_ltv	2342	72.62	26.66	79.42	4.3	138.5	134.2	-0.74	-0.13
current_balance	2342	90729.69	115243.56	49559.65	45.29	793649.71	793604.42	2.56	7.04
corporate_adv	2342	8367.93	14368.58	4529.73	170	198219.24	198049.24	7.02	70.28
rem_term	2342	169.24	115.18	158	6	526	520	0.68	-0.16
mod_flag	2342	0.24	0.43	0	0	1	1	1.2	-0.57
N	2342								

Table 20: Multivariate Analysis of Borrower Forbearance for all loans
This table reports the regression results for early forbearance.

	(1)	(2)	(3)	(4)
	All Loans		FHA/VA Loans	Non-FHA/VA Loans
Inbound Call	2.772***	2.939***	3.279***	2.627***
	(0.181)	(0.199)	(0.219)	(0.232)
Outbound Call	2.209***	3.094***	3.020***	3.145***
	(0.32)	(0.483)	(0.413)	(0.722)
AssetType_n	-1.088**	-0.169	-0.181	-0.654
	(0.414)	(0.501)	(0.617)	(0.83)
Government (FHA/VA) Loan	-0.714	0.107		
	(0.475)	(0.471)		
corporate_adv	-0.173	0.263	0.502	-6.723*
	(0.345)	(0.31)	(0.341)	(3.307)
Inbound Call X Government	0.453**	0.27		
	(0.146)	(0.166)		
Inbound Call X Performing Loan	0.827***	0.506**	0.197	1.065***
	(0.151)	(0.169)	(0.208)	(0.262)
Outbound Call X Government	0.691	-0.103		
	(0.438)	(0.429)		
Outbound Call X Performing Loan	2.134***	1.381**	1.379*	1.255
	(0.385)	(0.461)	(0.567)	(0.793)
AssetType_Gov	-0.164	-0.207		
	(0.155)	(0.147)		
current_balance	-9.172***	16.991***	19.264*	13.474*
	(0.787)	(4.604)	(9.022)	(5.814)
orig_ltv	0.001	0	-0.006	0.004
	(0.002)	(0.002)	(0.006)	(0.003)
original_fico	0.000059	0.000404	0.0003	0.0002
	(0.000467)	(0.000476)	(0.001)	(0.001)
current_rate	-9.754***	-3.181*	-23.120***	-1.362
	(1.782)	(1.62)	(6.442)	(1.775)
currentratetype_new	0.048	0.178	0.003	0.089
	(0.162)	(0.152)	(0.297)	(0.191)
Borrower Noted Unemployment	0.623***	0.690***	0.396	1.037***
	(0.174)	(0.174)	(0.232)	(0.258)
Borrower Noted Income Curtailment	2.144***	2.031***	2.007***	2.005***
	(0.152)	(0.161)	(0.218)	(0.233)
AIC	7992.973	7273.096	3408.143	2665.806
Log Likelihood	-3979.486			
Fixed Effect	None	County	County	County
Num. obs.	96,244	96,244	33,343	62,901

Table 21: Multivariate Analysis of Borrower Forbearance for Purchase and Cashout loans.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Purchase Mortgages				Cash-out Refinance Mortgages			
	All Loans		FHA/VA	Non-FHA/VA	All Loans		FHA/VA	Non-FHA/VA
Inbound Call	2.7711***	2.9722***	3.2793***	2.180***	2.6980***	2.7712***	3.139***	2.692***
	(0.287)	(0.3479)	(0.2187)	(0.571)	(0.2518)	(0.1268)	(0.443)	(0.264)
Outbound Call	1.6704***	2.1738***	3.0196***	19.228***	2.1714***	2.8869***	19.345***	2.918***
	(0.4038)	(0.5297)	(0.4126)	(0.347)	(0.3837)	(0.3593)	(0.209)	(0.726)
AssetType_n	-0.204	0.3569	-0.1808	16.335***	-1.8966**	-1.0973***	1.499**	-1.326
	(0.5658)	(0.589)	(0.617)	(1.171)	(0.6033)	(0.1293)	(0.499)	(0.884)
Government (FHA/VA) Loan	-1.7975***	-1.2754			-12.0984***	-11.4466***		
	(0.5349)	(0.6568)			(0.4401)	(0.2338)		
corporate_adv	0.1178	0.4746	0.502	-18.438	-2.3454	-1.1616	4.358*	-9.931**
	(0.3386)	(0.346)	(0.3413)	(10.488)	(1.4114)	(1.2119)	(1.947)	(3.559)
Inbound Call X Government	0.4134	0.1944			0.5063	0.0656		
	(0.2602)	(0.3016)			(0.4504)	(0.2692)		
Inbound Call X Performing Loan	0.5852**	0.3183	0.1966	0.797	1.2644***	0.9958***	0.079	1.173***
	(0.1908)	(0.2014)	(0.2081)	(0.7)	(0.2632)	(0.1064)	(0.378)	(0.292)
Outbound Call X Government	1.4391***	1.0081			12.0484***	11.5780***		
	(0.4326)	(0.5613)			(0.405)	(0.2338)		
Outbound Call X Performing Loan	1.4717**	1.0571*	1.3792*	-16.010***	2.6507***	1.8388***	0.062	1.846*
	(0.5004)	(0.5089)	(0.5674)	(0.82)	(0.5716)	(0.1263)	(0.499)	(0.845)
AssetType_Gov	-0.1993	-0.2795			-0.2702	-0.1664		
	(0.3389)	(0.301)			(0.4694)	(0.2856)		
current_balance	-7.2208***	14.1777*	19.2638*	-5.259	-7.8821***	16.6653**	75.064	16.701*
	(0.9811)	(7.2097)	(9.0217)	(15.407)	(1.5093)	(5.4343)	(63.105)	(6.884)
orig_ltv	-0.0041	-0.0032	-0.0059	0.013	0.0015	0.001	-0.041	0.002
	(0.0045)	(0.0051)	(0.0061)	(0.008)	(0.0022)	(0.0022)	(0.049)	(0.003)
original_fico	-0.0001	-0.0008	-0.0002	-0.005	0.0004	0.0009	-0.002	0
	(0.0006)	(0.0007)	(0.0007)	(0.003)	(0.0008)	(0.0007)	(0.004)	(0.001)
current_rate	-18.9145***	-14.0843***	-23.1202***	-17.760*	-6.3670***	-0.7243	-54.996	-0.968
	(4.9283)	(4.2271)	(6.4416)	(7.698)	(1.8336)	(1.7762)	(30.443)	(1.895)
currentratetype_new	0.1746	0.3908	0.003	0.64	0.1467	0.1225	0.0003	0.241
	(0.2483)	(0.2345)	(0.2966)	(0.489)	(0.2228)	(0.226)	(0.0005)	(0.229)
Borrower Noted Unemployment	0.3541	0.3803	0.3963	0.134	1.0615***	1.1448***	1.36	1.189***
	(0.2254)	(0.2234)	(0.2321)	(0.608)	(0.272)	(0.223)	(0.868)	(0.28)
Borrower Noted Income Curtailment	2.0287***	1.9053***	2.0072***	0.995	2.3170***	2.1810***	2.504***	2.193***
	(0.1959)	(0.2031)	(0.2182)	(0.637)	(0.2419)	(0.1611)	(0.521)	(0.256)
AIC	4687.395	3684.592	3125.64	223.096	3289.827	2411.503	91.332	2203.831
Log Likelihood	-2326.697				-1607.07			
Fixed Effects	None	County	County	County	None	County	County	County
Num. obs.	35750.000	35750.000	31273.000	4477.000	60494.000	60494.000	2070.000	58424.000

.7 Quantifying Soft Information, Mortgage Market Efficiency & Asset Pricing Implications

Table 26: Variables Names and Descriptions

Variable Name	**Description**
LoanID	Unique identifier of the loan in the Private Equity firm
cutoff_date	The date on which the observations on hard information are recorded
origination_date	The origination date of the mortgage loan by the original lender
first_pay_date	The first date on which the deb obligations are due for the borrower
maturity_date	The original maturity date of the loan, which may change due to loan modification and/or renegotiation
original_balance	Original Balance of the loan when issued by the lender
original_appraisal	The appraisal value of the property when the loan is originated
original_term	Original term of the loan
amort_term	Current amortization term of the loan, different from Original term if the loan term has been modified
original_fico	Original FICO score of the borrower at origination
original_rate	Original mortgage rate offered to the borrower at mortgage origination
original_pi	Original amortized amount of monthly principal and interest payments due
original_rate_type	Original type of mortgage rate: Fixed or ARM
balloon_flag	Y/N
io_flag	Y/N
io_term	Mortgage term where the loan is interest-only
loan_purpose	Cash Out, Purchase, Rate Term, Home Improvement
property_type	Condo Modular Home Multi Family PUD Single Family Townhouse
occupancy_type	Investment Property Owner Occupied Second Home
doc_type	Full Limited None Reduced Stated
current_balance	Current outstanding balance of the loan
value	Latest appraised value of the property
iptd	Interest payment due date
npdd	Next payment due date
current_rate	Current interest rate of the mortgage
current_pi	Current amortized amount of monthly principal and interest
bk_flag	Y/N
bk_file_date	The date when bankruptcy was last filed
fc_flag	Y/N
fc_start_date	The date when foreclosure was last filed
current_fico	Current FICO score of the borrower
current_fico_date	The date when the current FICO score was calibrated
zip	Zip code of the borrower residence
mod_flag	Y/N
mod_date	The date of the last loan modification
current_rate_type	Current type of rate: Fixed or ARM
orig_ltv	Original Loan-to-Value
lien_position	1/2/3
deferred_balance	Balance deferred to next year for tax-saving purposes
reo_flag	Y/N
jr_lien_balance	The balance left in the junior lien mortgage
pmts1	Payments in a month
pmts3	Payments in 3 months
pmts6	Payments in 6 months
pmts12	Payments in a year
corporate_adv	Corporate advances by the servicer
escrow_adv	Escrow advances by the serviser
state	State of the borrower residence
legal_grade	A, A-, A+, B, B-, B+, C, C-, C+, D, D-, D+, F, NI

mortgage_ins	Insurance related to mortgage obligaitons
fc_stage	F0 - No FC, F0 - Removed F1, F1 - File Referred, F2 F2 - First Legal Filed, F3 F3 - Judgment Entered, F5 - Sale Scheduled
address	Street address of borrower residence
city	Cit of borrower residence
loan_type	Conventional, FHA, HELOC, PMI, USDA, VA
number_of_units	Number of units in the property
bk_dq_delay	Delay between delinquency and filing of bankruptcy
bk_fc_delay	Delay between filing bankruptcy and filing foreclosure
loan_age	Age of the loan from origination
mnths_in_bk	Number of months a loan has been in bankruptcy
mnths_in_fc	Number of months a loan has been in forelosure
mnths_since_mod	Number of months a loan has been modified
rem_term	Remaining term of the loan, which can change if the loan term is modified
days_dq	Days of Delinqency
Status	Bankruptcy Current Foreclosure W0_30D W30_60D W60_90D W90_120D
Year_Month	The month in the year the observation is made about loan performance variables
AssetType	NPL, PL, REO, ShortSale
LoanStatus	Active - 120+ Days DQ, Active - 30 Days DQ, Active - 60 Days DQ, Active - 90 Days DQ, Active - BK, Active - Current, Active - FC, Active - REO, LIQ - Charge Off, LIQ - Third Party Sale, Pending Servicing Transfer
SaleStrat	REO, Foreclosure Sale Scheduled, Sale Ready, Pending Short Sale, Pending Deed-in-Lieu, Loss Mit Hold, Out for offers Modified Recently, Pending Third Party Sale, On Hold LoanSale Status, Foreclosure Judgment Entered,

Table 22: Multivariate Analysis of Borrower Forbearance and Race

In Table 22, I add an indicator whether a borrower in Black and interact the dummy variable with IB, OB, IB_PL and OB_PL.

	(1)	(2)	(3)	(4)	(5)	(6)
	All Loans		FHA/VA Loans		Non-FHA/VA Loans	
Inbound Call	2.2319^{***}	2.6727^{***}	3.2027^{***}	3.3648^{***}	1.8695^{***}	1.8817^{***}
	(0.2754)	(0.3798)	(0.2062)	(0.2582)	(0.5097)	(0.3850)
Outbound Call	0.2602	1.9040^{**}	2.1668^{***}	3.1023^{***}	1.4771^{***}	19.9902^{***}
	(0.3084)	(0.5886)	(0.2336)	(0.4711)	(0.4436)	(0.5709)
AssetType_n	-1.0236^{*}	0.4955	-1.5982^{**}	-0.3012	-0.3179	17.9754^{***}
	(0.4330)	(0.6446)	(0.5309)	(0.6930)	(0.7133)	(0.4405)
Gov	-3.1633^{***}	-1.7033^{*}				
	(0.4501)	(0.6719)				
Black	-0.9346	-0.3006	-0.3018	0.4258	-29.3945^{***}	-52.9484^{***}
	(0.6573)	(0.7690)	(0.6449)	(0.7629)	(0.9410)	(0.9052)
corporate_adv	-0.0615^{**}	-0.0241	-0.0545^{**}	-0.0198	-0.1633	-0.3023
	(0.0203)	(0.0199)	(0.0211)	(0.0204)	(0.1930)	(0.2364)
Inbound Call X Government	0.9447^{***}	0.5606				
	(0.2534)	(0.3231)				
Inbound Call X Performing Loan	0.4287	0.1033	0.3874	0.0072	0.9865	0.3951
	(0.2198)	(0.2355)	(0.2274)	(0.2514)	(0.6394)	(0.3921)
Inbound Call X Black	0.0797	-0.1952	-0.0172	-0.2187	0.4552	-0.4292
	(0.3563)	(0.4033)	(0.3755)	(0.4311)	(1.0909)	(1.2637)
Inbound Call X Performing Loan X Black	1.0892	1.1840	1.0405	1.0202	14.5899^{***}	37.1040
	(0.6471)	(0.6103)	(0.6460)	(0.6256)	(1.3884)	(5308.7008)
Outbound Call X Government	2.4230^{***}	1.3121^{*}				
	(0.4095)	(0.5874)				
Outbound Call X Performing Loan	2.2692^{***}	1.0948	2.6058^{***}	1.4769^{*}	1.3874^{*}	-17.3784^{***}
	(0.4377)	(0.5630)	(0.4756)	(0.6358)	(0.6651)	(0.4108)
Outbound Call X Black	0.9429	0.3988	0.3865	-0.3640	29.4304^{***}	52.3158^{***}
	(0.6668)	(0.7904)	(0.6277)	(0.7456)	(1.2576)	(0.9052)
Outbound Call X Performing Loan X Black	-1.2053^{*}	-1.1254^{*}	-1.1010	-0.8566	-15.7338^{***}	-18.2862
	(0.6125)	(0.5723)	(0.6089)	(0.5837)	(1.0205)	(5412.8656)
AssetType_Gov	-0.2585	-0.4327				
	(0.3734)	(0.3454)				
current_balance	-0.0920	0.1292	-0.1624^{**}	0.1734	-0.0205	-0.0927
	(0.0473)	(0.0889)	(0.0540)	(0.1080)	(0.1691)	(0.2751)
orig_ltv	-0.0150^{***}	-0.0056	-0.0234^{***}	-0.0084	-0.0119	0.0036
	(0.0043)	(0.0056)	(0.0051)	(0.0072)	(0.0076)	(0.0174)
original_fico	-0.0017^{**}	-0.0014	-0.0015^{*}	-0.0010	-0.0050^{*}	-0.0085^{*}
	(0.0006)	(0.0008)	(0.0007)	(0.0008)	(0.0022)	(0.0036)
current_rate	-24.7370^{***}	-12.1368^{*}	-43.1878^{***}	-22.6451^{***}	-11.5115	-14.6377
	(5.0955)	(5.0962)	(5.5675)	(6.8337)	(6.6101)	(8.6513)
currentratetype_new	0.0363	0.4382	-0.4367	0.2032	0.1956	0.9364
	(0.2606)	(0.2745)	(0.3944)	(0.3956)	(0.3744)	(0.5797)
Borrower Noted Unemployment	0.3068	0.3236	0.2180	0.3024	0.8731	0.2683
	(0.2468)	(0.2531)	(0.2637)	(0.2760)	(0.6727)	(0.5892)
Borrower Noted Income Curtailment	1.9243^{***}	1.8378^{***}	2.0206^{***}	1.9731^{***}	0.8551	0.8601
	(0.2044)	(0.2237)	(0.2220)	(0.2495)	(0.5318)	(0.5551)
AIC	4021.2615	3022.3850	3653.3871	2658.0711	398.9610	173.7232
Log Likelihood	-1988.6307		-1808.6936		-181.4805	
Fixed Effects	*None*	*County*	*None*	*County*	*None*	*County*
Num. obs.	28843	28843	25845	25845	2998	2998

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 23: Differences in Differences

	(1)	(2)	(3)	(4)	(5)	(6)
	All Loans		Inbound Communications		Outbound Communications	
After_March_2020_X_Gov	6.6243^{***}	6.6480^{***}	6.5795^{***}	6.2790^{***}	6.2565^{***}	6.3283^{***}
	(0.5883)	(0.9369)	(0.5832)	(0.6014)	(0.5816)	(0.6832)
Inbound Call	2.4607^{***}	3.2348^{***}	3.0057^{***}	3.8570^{***}		
	(0.2536)	(0.3597)	(0.2728)	(0.4061)		
Outbound Call	0.7644^{**}	2.5906^{***}			1.1442^{***}	3.2238^{***}
	(0.2735)	(0.5492)			(0.2668)	(0.5963)
AssetType_n	-1.3161^{**}	0.7332	-0.1727	0.6966	-1.4215^{***}	0.4652
	(0.4404)	(0.5443)	(0.2966)	(0.3644)	(0.3693)	(0.5514)
Gov	-8.2762^{***}	-6.6097^{***}	-6.6315^{***}	-5.5770^{***}	-7.6262^{***}	-6.5326^{***}
	(0.7661)	(1.2329)	(0.6573)	(0.7222)	(0.7148)	(0.9812)
corporate_adv	-0.1832^{***}	-0.1012^{***}	-0.1616^{***}	-0.0884^{***}	-0.2109^{***}	-0.1633^{***}
	(0.0277)	(0.0187)	(0.0248)	(0.0198)	(0.0227)	(0.0208)
Inbound Call X Government	0.8808^{***}	-0.0461	1.0352^{***}	0.2352		
	(0.2460)	(0.3158)	(0.2407)	(0.3278)		
Inbound Call X Performing Loan	0.8933^{***}	0.1936	1.2170^{***}	0.6775^{**}		
	(0.1928)	(0.2096)	(0.2222)	(0.2409)		
Outbound Call X Government	2.0149^{***}	0.6839			2.2600^{***}	1.0501
	(0.3733)	(0.5634)			(0.3679)	(0.5808)
Outbound Call X Performing Loan	2.1639^{***}	0.5864			3.0530^{***}	1.3807^{**}
	(0.4017)	(0.4656)			(0.3941)	(0.5035)
AssetType_Gov	-0.3351	-0.5857^{*}	-0.1123	-0.5491	-0.4950	-0.7052^{*}
	(0.3331)	(0.2948)	(0.2950)	(0.3154)	(0.3050)	(0.3047)
current_balance	-0.0704	0.0975	-0.0530	0.2068^{**}	0.0390	0.2699^{***}
	(0.0530)	(0.0730)	(0.0460)	(0.0719)	(0.0453)	(0.0751)
orig_ltv	-0.0126^{**}	0.0021	-0.0142^{***}	-0.0016	-0.0127^{**}	-0.0068
	(0.0048)	(0.0048)	(0.0043)	(0.0050)	(0.0039)	(0.0050)
original_fico	-0.0021^{**}	-0.0009	-0.0024^{***}	-0.0015^{*}	-0.0018^{**}	-0.0012
	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
current_rate	-18.6557^{***}	-10.7210^{**}	-21.8364^{***}	-9.2600^{*}	-20.9353^{***}	-9.3156^{*}
	(5.3494)	(4.0850)	(5.2717)	(4.3092)	(5.0597)	(4.3302)
currentratetype_new	-0.2190	0.2646	0.0120	0.2879	-0.0680	0.2741
	(0.2559)	(0.2304)	(0.2353)	(0.2315)	(0.2177)	(0.2257)
Borrower Noted Unemployment	0.5279^{*}	0.6688^{**}	0.8085^{**}	1.0197^{***}	-0.6248^{**}	-0.5150^{**}
	(0.2410)	(0.2323)	(0.2666)	(0.2733)	(0.1914)	(0.1845)
Borrower Noted Income Curtailment	2.0100^{***}	1.9669^{***}	2.9142^{***}	2.9683^{***}	-0.1915	-0.1544
	(0.1959)	(0.2055)	(0.2487)	(0.2634)	(0.1273)	(0.1301)
AIC	3245.3400	2528.2219	3954.6566	3029.6796	4354.0823	3269.9377
BIC	3398.0575		4081.9212		4481.3469	
Log Likelihood	-1604.6700		-1962.3283		-2162.0412	
Fixed Effects	$None$	$County$	$None$	$County$	$None$	$County$
Num. obs.	35750	35750	35750	35750	35750	35750

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 24: Differences in Differences, Gov for IB and NonGov for OB

	(1)	(2)	(3)	(4)	(5)	(6)
	All Loans		Inbound Communications		Outbound Communications	
After_March_2020_X_Gov	6.6243^{***}	6.6480^{***}	6.5795^{***}	6.2790^{***}		
	(0.5883)	(0.9369)	(0.5832)	(0.6014)		
After_March_2020_X_NonGov					4.7099^{***}	4.5125^{***}
					(0.9804)	(1.0035)
Inbound Call	2.4607^{***}	3.2348^{***}	3.0057^{***}	3.8570^{***}		
	(0.2536)	(0.3597)	(0.2728)	(0.4061)		
Outbound Call	0.7644^{**}	2.5906^{***}			2.6076^{***}	3.8210^{***}
	(0.2735)	(0.5492)			(0.1788)	(0.4421)
AssetType_n	-1.3161^{**}	0.7332	-0.1727	0.6966	-1.1151^{*}	0.0564
	(0.4404)	(0.5443)	(0.2966)	(0.3644)	(0.4413)	(0.5347)
Gov	-8.2762^{***}	-6.6097^{***}	-6.6315^{***}	-5.5770^{***}		
	(0.7661)	(1.2329)	(0.6573)	(0.7222)		
NonGov					-3.5564^{**}	-2.4699^{*}
					(1.1291)	(1.1568)
corporate_adv	-0.1832^{***}	-0.1012^{***}	-0.1616^{***}	-0.0884^{***}	-0.1009^{***}	-0.0695^{***}
	(0.0277)	(0.0187)	(0.0248)	(0.0198)	(0.0163)	(0.0182)
Inbound Call X Government	0.8808^{***}	-0.0461	1.0352^{***}	0.2352		
	(0.2460)	(0.3158)	(0.2407)	(0.3278)		
Inbound Call X Performing Loan	0.8933^{***}	0.1936	1.2170^{***}	0.6775^{**}		
	(0.1928)	(0.2096)	(0.2222)	(0.2409)		
Outbound Call X Government	2.0149^{***}	0.6839				
	(0.3733)	(0.5634)				
Outbound Call X Non-Government					-0.5191	-1.2545^{*}
					(0.5188)	(0.5718)
Outbound Call X Performing Loan	2.1639^{***}	0.5864			3.0772^{***}	1.8117^{***}
	(0.4017)	(0.4656)			(0.3239)	(0.4992)
AssetType_Gov	-0.3351	-0.5857^{*}	-0.1123	-0.5491	-0.6606	-0.4843
	(0.3331)	(0.2948)	(0.2950)	(0.3154)	(0.3557)	(0.2914)
current_balance	-0.0704	0.0975	-0.0530	0.2068^{**}	-0.0704	0.2702^{***}
	(0.0530)	(0.0730)	(0.0460)	(0.0719)	(0.0399)	(0.0735)
orig_ltv	-0.0126^{**}	0.0021	-0.0142^{***}	-0.0016	-0.0183^{***}	-0.0069
	(0.0048)	(0.0048)	(0.0043)	(0.0050)	(0.0035)	(0.0050)
original_fico	-0.0021^{**}	-0.0009	-0.0024^{***}	-0.0015^{*}	-0.0022^{***}	-0.0015^{*}
	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0005)	(0.0006)
current_rate	-18.6557^{***}	-10.7210^{**}	-21.8364^{***}	-9.2600^{*}	-29.0106^{***}	-14.5034^{**}
	(5.3494)	(4.0850)	(5.2717)	(4.3092)	(5.2108)	(4.6068)
currentratetype_new	-0.2190	0.2646	0.0120	0.2879	-0.0420	0.3452
	(0.2559)	(0.2304)	(0.2353)	(0.2315)	(0.2006)	(0.2303)
Borrower Noted Unemployment	0.5279^{*}	0.6688^{**}	0.8085^{**}	1.0197^{***}	-0.7835^{***}	-0.7555^{***}
	(0.2410)	(0.2323)	(0.2666)	(0.2733)	(0.1741)	(0.1777)
Borrower Noted Income Curtailment	2.0100^{***}	1.9669^{***}	2.9142^{***}	2.9683^{***}	-0.2202	-0.2116
	(0.1959)	(0.2055)	(0.2487)	(0.2634)	(0.1163)	(0.1227)
AIC	3245.3400	2528.2219	3954.6566	3029.6796	5683.7921	4454.4821
BIC	3398.0575		4081.9212		5811.0567	
Log Likelihood	-1604.6700		-1962.3283		-2826.8960	
Fixed Effects	$None$	$County$	$None$	$County$	$None$	$County$
Num. obs.	35750	35750	35750	35750	35750	35750

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 25: Differences in Differences, removing 2 or less observations per county

	(1)	(2)	(3)	(4)	(5)	(6)
	All Loans		Inbound Communications		Outbound Communications	
Interaction	6.6137^{***}	6.0908^{***}	6.5615^{***}	6.1486^{***}	6.2421^{***}	6.0306^{***}
	(0.5889)	(0.6775)	(0.5836)	(0.5832)	(0.5819)	(0.5895)
Inbound Call	2.4870^{***}	3.2611^{***}	3.0545^{***}	3.8719^{***}		
	(0.2544)	(0.3632)	(0.2731)	(0.4094)		
Outbound Call	0.8978^{**}	2.7987^{***}			1.2758^{***}	3.3759^{***}
	(0.2761)	(0.6108)			(0.2694)	(0.6615)
AssetType_n	-1.1251^{**}	0.9906	-0.1293	0.7092	-1.2550^{***}	0.6351
	(0.4225)	(0.6250)	(0.2973)	(0.3661)	(0.3579)	(0.6374)
Gov	-8.4231^{***}	-6.1146^{***}	-6.6397^{***}	-5.4512^{***}	-7.7384^{***}	-6.2794^{***}
	(0.7716)	(1.0131)	(0.6591)	(0.7063)	(0.7239)	(0.9009)
corporate_adv	-0.1796^{***}	-0.0982^{***}	-0.1565^{***}	-0.0863^{***}	-0.2108^{***}	-0.1617^{***}
	(0.0282)	(0.0188)	(0.0252)	(0.0199)	(0.0230)	(0.0209)
Inbound Call X Government	0.8981^{***}	-0.0417	1.0470^{***}	0.2370		
	(0.2480)	(0.3166)	(0.2426)	(0.3284)		
Inbound Call X Performing Loan	0.8637^{***}	0.1867	1.1701^{***}	0.6665^{**}		
	(0.1941)	(0.2117)	(0.2212)	(0.2436)		
Outbound Call X Government	2.1255^{***}	0.7326			2.3564^{***}	1.0845
	(0.3725)	(0.5739)			(0.3712)	(0.5900)
Outbound Call X Performing Loan	1.9753^{***}	0.3248			2.8769^{***}	1.2012^{*}
	(0.3899)	(0.5570)			(0.3858)	(0.6010)
AssetType_Gov	-0.2769	-0.5593	-0.0672	-0.5356	-0.4618	-0.6863^{*}
	(0.3335)	(0.2954)	(0.2968)	(0.3159)	(0.3057)	(0.3059)
current_balance	-0.0755	0.0985	-0.0522	0.2081^{**}	0.0343	0.2693^{***}
	(0.0549)	(0.0733)	(0.0469)	(0.0721)	(0.0465)	(0.0753)
orig_ltv	-0.0129^{**}	0.0022	-0.0146^{***}	-0.0015	-0.0130^{***}	-0.0068
	(0.0048)	(0.0048)	(0.0043)	(0.0050)	(0.0039)	(0.0050)
original_fico	-0.0023^{***}	-0.0009	-0.0025^{***}	-0.0015^{*}	-0.0020^{**}	-0.0012
	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
current_rate	-18.5651^{***}	-10.7232^{**}	-21.7854^{***}	-9.2273^{*}	-20.2961^{***}	-9.2813^{*}
	(5.4535)	(4.0943)	(5.3461)	(4.3101)	(5.1844)	(4.3247)
currentratetype_new	-0.1930	0.2655	0.0388	0.2873	-0.0412	0.2747
	(0.2594)	(0.2307)	(0.2378)	(0.2316)	(0.2216)	(0.2257)
Borrower Noted Unemployment	0.5562^{*}	0.6916^{**}	0.8448^{**}	1.0553^{***}	-0.6066^{**}	-0.5051^{**}
	(0.2434)	(0.2329)	(0.2697)	(0.2734)	(0.1926)	(0.1840)
Borrower Noted Income Curtailment	2.0476^{***}	1.9879^{***}	2.9676^{***}	2.9624^{***}	-0.1738	-0.1503
	(0.1988)	(0.2086)	(0.2510)	(0.2664)	(0.1283)	(0.1299)
AIC	3149.8709	2519.4062	3858.9352	3020.5540	4241.1855	3263.9657
BIC	3302.1351		3985.8221		4368.0724	
Log Likelihood	-1556.9354		-1914.4676		-2105.5928	
Fixed Effects	*None*	*County*	*None*	*County*	*None*	*County*
Num. obs.	34861	34861	34861	34861	34861	34861

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 27: In-Sample Confusion Matrices for PL, NPL, PIF, ShortSale using Loan Data

	DRF	Lasso	GBM	DNN
NPL	0.03	1	0.02	1
PIF	0.04	1	0.08	1
PL	0	0	0	0
ShortSale	0.01	1	0.04	1

Table 28: Variable Importance using DRF for PL, NPL, PIF, ShortSale with Loan Data.

	variable	relative_importance	scaled_importance	percentage
1	**intrst_payment_due**	1334526.88	1	0.13
2	**next_pay_duedate**	1193423.12	0.89	0.11
3	**loan_type**	1008397.88	0.76	0.10
4	**fc_stage**	971392.25	0.73	0.09
5	**fc_flag**	548978.38	0.41	0.05
6	**state**	483539.28	0.36	0.05
7	**rem_term**	473346.88	0.35	0.05
8	**cutoff_date**	415253.38	0.31	0.04
9	**legal_grade**	394640.56	0.30	0.04
10	**bk_flag**	343188.91	0.26	0.03
11	**loan_purpose**	292739.31	0.22	0.03
12	**fc_start_date**	282192.66	0.21	0.03
13	**jr_lien_balance**	189844.38	0.14	0.02
14	**current_rate**	189190.47	0.14	0.02
15	**reo_flag**	188070.17	0.14	0.02
16	**current_fico_date**	180391.91	0.14	0.02
17	**current_balance**	152752.38	0.11	0.01
18	**original_fico**	151196.14	0.11	0.01
19	**loan_age**	146700.42	0.11	0.01
20	**mths_in_fc**	146056.75	0.11	0.01
21	**current_fico**	131127.70	0.10	0.01
22	**orig_ltv**	105868.62	0.08	0.01
23	**bk_dq_delay**	97435.91	0.07	0.01
24	**value**	75367.45	0.06	0.01
25	**bk_file_date**	70044.45	0.05	0.01
26	**original_rate**	67839.54	0.05	0.01
27	**original_appraisal**	66309.92	0.05	0.01
28	**original_balance**	65674.52	0.05	0.01
29	**property_type**	62172.73	0.05	0.01
30	**current_pi**	60741.72	0.05	0.01
31	**mod_date**	59569.18	0.04	0.01
32	**original_pi**	59540.11	0.04	0.01
33	**doc_type**	59156.05	0.04	0.01
...	...	...	...	...
49	balloon_flag	3466.04	0	0

Table 29: Variable Importance using Lasso for PL, NPL, PIF, ShortSale with Loan Data.

	variable	relative_importance	scaled_importance	percentage
1	loan_type.NULL	0.86	1	0.45
2	cutoff_date	0.66	0.76	0.34
3	legal_grade.NI	0.09	0.11	0.05
4	rem_term	0.08	0.09	0.04
5	original_fico	0.08	0.09	0.04
6	jr_lien_balance	0.05	0.06	0.03
7	loan_type.CONV	0.05	0.06	0.03
8	loan_age	0.02	0.02	0.01
9	mortgage_ins	0.02	0.02	0.01
10	current_rate	0.02	0.02	0.01
11	pmts1	0	0	0
12	current_fico	0	0	0
13	value	0	0	0
14	intrst_payment_due	0	0	0
15	next_pay_duedate	0	0	0

Table 30: Variable Importance using DNN for PL, NPL, PIF, ShortSale with Loan Data.

	variable	relative_importance	scaled_importance	percentage
1	**pmts1**	1	1	0.02
2	**pmts6**	0.71	0.71	0.01
3	**pmts3**	0.70	0.70	0.01
4	**pmts12**	0.66	0.66	0.01
5	**state.DE**	0.49	0.49	0.01
6	**legal_grade.NI**	0.46	0.46	0.01
7	**state.OK**	0.45	0.45	0.01
8	**state.AR**	0.45	0.45	0.01
9	**current_rate_type.FIX**	0.45	0.45	0.01
10	**legal_grade.A**	0.45	0.45	0.01
11	**state.HI**	0.44	0.44	0.01
12	**doc_type.None**	0.44	0.44	0.01
13	**io_flag..N**	0.44	0.44	0.01
14	**state.FL**	0.43	0.43	0.01
15	**state.NC**	0.43	0.43	0.01
16	**state.NV**	0.43	0.43	0.01
17	**occupancy_type.Investment Property**	0.43	0.43	0.01
18	**property_type.PUD**	0.42	0.42	0.01
19	**occupancy_type.Owner Occupied**	0.42	0.42	0.01
20	**state.AZ**	0.42	0.42	0.01
21	**original_rate_type.FIX**	0.42	0.42	0.01
22	**doc_type.Full**	0.42	0.42	0.01
23	**loan_type.HELOC**	0.42	0.42	0.01
24	**doc_type.Stated**	0.42	0.42	0.01
25	**state.VA**	0.41	0.41	0.01
26	occupancy_type.NULL	0.41	0.41	0.01
27	bk_file_date	0.41	0.41	0.01
28	mortgage_ins	0.41	0.41	0.01
29	state.PA	0.41	0.41	0.01
30	state.KS	0.41	0.41	0.01
31	bk_fc_delay	0.41	0.41	0.01
32	state.WA	0.41	0.41	0.01
33	balloon_flag.Y	0.41	0.41	0.01
34	state.TN	0.40	0.40	0.01
35	bk_flag.N	0.40	0.40	0.01
36	mod_flag.Y	0.40	0.40	0.01
37	state.OR	0.40	0.40	0.01
38	state.AK	0.40	0.40	0.01
...	...	...	...	...
158	state.WY	0.27	0.27	0

Table 31: Out-of-Sample Confusion Matrices for PL, NPL, PIF, ShortSale with Loan Data
I provide the misclassification errors for the four classes **NPL** (Non-Performing Loan), **PIF** (Paid-in-Full), **PL** (Performing Loan), **ShortSale** from the Out-of-Sample Confusion Matrices in Table 31 only using loan performance data.

	DRF	Lasso	GBM	DNN
NPL	0.03	1	0.02	1
PIF	0.07	1	0.09	1
PL	0	0	0	0
ShortSale	0.08	1	0.08	1

Table 32: In-Sample Confusion Matrix for PL, NPL, PIF, ShortSale, Servicer Comments, Loan Data (Loan left join tfidf): I first provide the misclassification errors for the four classes **NPL** (Non-Performing Loan), **PIF** (Paid-in-Full), **PL** (Performing Loan), **ShortSale** from the In-Sample Confusion Matrices in Table 32 using loan performance data with servicer comments.

	DRF	Lasso	GBM	DNN
NPL	0.05	0.27	0.03	0.07
PIF	0.06	1	0.08	0.08
PL	0	0	0.01	0.01
ShortSale	0.02	1	0.04	0.04

Table 33: Variable Importance using DRF for PL, NPL, PIF, ShortSale, Servicer Comments, Loan Data (Loan left join tfidf).

	variable	relative_importance	scaled_importance	percentage
1	**next_pay_duedate**	1512022.12	1	0.12
2	**intrst_payment_due**	1463559.88	0.97	0.11
3	**loan_type**	1255022.12	0.83	0.10
4	**fc_stage**	1234660.25	0.82	0.10
5	**fc_flag**	804520.25	0.53	0.06
6	**state**	637637.50	0.42	0.05
7	**rem_term**	619736.06	0.41	0.05
8	**legal_grade**	510212.69	0.34	0.04
9	**bk_flag**	392216.34	0.26	0.03
10	**fc_start_date**	329544.81	0.22	0.03
11	**original_fico**	290480.78	0.19	0.02
12	**jr_lien_balance**	261891.11	0.17	0.02
13	**current_rate**	260584.06	0.17	0.02
14	**loan_purpose**	258598.89	0.17	0.02
15	**current_fico_date**	246395.77	0.16	0.02
16	**reo_flag**	223863.62	0.15	0.02
17	**loan_age**	199507.52	0.13	0.02
18	**mths_in_fc**	187982.50	0.12	0.01
19	**current_balance**	173566.98	0.11	0.01
20	**current_fico**	159067.14	0.11	0.01
21	**orig_ltv**	132592.11	0.09	0.01
22	**bk_dq_delay**	126313.82	0.08	0.01
23	value	120004.49	0.08	0.01
24	original_rate	119893.46	0.08	0.01
25	current_pi	98654.03	0.07	0.01
26	original_appraisal	96831.14	0.06	0.01
27	original_balance	91224.95	0.06	0.01
28	original_pi	86224.37	0.06	0.01
29	bk_file_date	85650.62	0.06	0.01
30	doc_type	84533.49	0.06	0.01
31	mod_date	81525.81	0.05	0.01
32	mths_since_mod	69582.52	0.05	0.01
33	miss	67608.87	0.04	0.01
34	pmts12	65172.74	0.04	0.01
35	mths_in_bk	65073.76	0.04	0.01
36	property_type	63111.77	0.04	0
37	mortgage_ins	62803.14	0.04	0
...	...	...	...	...
84	activemilitary	0.01	0	0

Table 34: Variable Importance using Lasso for PL, NPL, PIF, ShortSale, Servicer Comments, Loan Data (Loan left join tfidf).

	variable	relative_importance	scaled_importance	percentage
1	bk_flag.Y	0.56	1	0.20
2	fc_flag.Y	0.55	0.99	0.20
3	fc_stage.NULL	0.28	0.50	0.10
4	loan_type.NULL	0.27	0.48	0.10
5	next_pay_duedate	0.13	0.23	0.05
6	intrst_payment_due	0.13	0.23	0.05
7	loan_type.CONV	0.11	0.19	0.04
8	original_fico	0.09	0.16	0.03
9	legal_grade.NI	0.09	0.16	0.03
10	current_fico	0.09	0.16	0.03
11	legal_grade.C+	0.08	0.14	0.03
12	jr_lien_balance	0.07	0.13	0.03
13	reo_flag.Y	0.07	0.13	0.03
14	bk_dq_delay	0.06	0.11	0.02
15	rem_term	0.04	0.06	0.01
16	mths_in_fc	0.03	0.06	0.01
17	mortgage_ins	0.03	0.06	0.01
18	miss	0.02	0.03	0.01
19	legal_grade.B	0.02	0.03	0.01
20	fc	0.01	0.02	0.01
21	loan_age	0.01	0.02	0
22	current_fico_date	0.01	0.02	0
23	reo	0.01	0.01	0
24	bk	0.01	0.01	0
25	current_rate	0	0.01	0

Table 35: Variable Importance using GBM for PL, NPL, PIF, ShortSale, Servicer Comments, Loan Data (Loan left join tfidf).

	variable	relative_importance	scaled_importance	percentage
1	**intrst_payment_due**	442743.75	1	0.31
2	**fc_stage**	248945.55	0.56	0.18
3	**loan_type**	143038.84	0.32	0.10
4	**jr_lien_balance**	142050.20	0.32	0.10
5	**rem_term**	121786.64	0.28	0.09
6	**bk_flag**	90131.88	0.20	0.06
7	**reo_flag**	48947.54	0.11	0.03
8	**state**	41900.22	0.09	0.03
9	**fc_flag**	37033.07	0.08	0.03
10	**next_pay_duedate**	12802.30	0.03	0.01
11	**legal_grade**	9607.75	0.02	0.01
12	**loan_purpose**	8994.52	0.02	0.01
13	**mths_since_mod**	6792.25	0.02	0
14	orig_ltv	4395.98	0.01	0
15	current_rate	4346.65	0.01	0
16	current_fico	4285.89	0.01	0
17	original_fico	3988.37	0.01	0
18	loan_age	3329.74	0.01	0
19	current_fico_date	2793.37	0.01	0
20	pmts1	2546.16	0.01	0
21	mortgage_ins	2544.43	0.01	0
22	mod_date	1812.83	0	0
23	original_pi	1719.88	0	0
24	current_balance	1702.25	0	0
25	doc_type	1679.66	0	0
26	original_appraisal	1621.78	0	0
27	original_rate	1542.15	0	0
28	occupancy_type	1537.92	0	0
29	original_balance	1461.04	0	0
30	property_type	1295.35	0	0
31	pmts12	1178.78	0	0
32	fire	1126.52	0	0
33	pmts3	1086.13	0	0
34	bk_dq_delay	981.81	0	0
35	value	959.16	0	0
36	pmts6	861.37	0	0
37	bk_file_date	650.64	0	0
38	io_flag.	602.64	0	0
...	...	...	...	...
63	insuranceclaim	0.18	0	0

Table 36: Variable Importance using DNN for PL, NPL, PIF, ShortSale, Servicer Comments, Loan Data (Loan left join tfidf).

	variable	relative_importance	scaled_importance	percentage
1	**bk_flag.Y**	1	1	0.01
2	**jr_lien_balance**	0.83	0.83	0.01
3	**intrst_payment_due**	0.82	0.82	0.01
4	**next_pay_duedate**	0.81	0.81	0.01
5	**loan_type.NULL**	0.75	0.75	0.01
6	**reo_flag.Y**	0.73	0.73	0.01
7	**fc_flag.Y**	0.72	0.72	0.01
8	**bk_flag.N**	0.59	0.59	0.01
9	**fc_flag.N**	0.56	0.56	0.01
10	**fc_stage.F0 - Removed**	0.55	0.55	0.01
11	**loan_type.CONV**	0.54	0.54	0.01
12	**loan_type.PMI**	0.53	0.53	0.01
13	**fc_stage.F1**	0.51	0.51	0.01
14	**state.DC**	0.51	0.51	0.01
15	**state.NC**	0.51	0.51	0.01
16	**mod_flag.Y**	0.51	0.51	0.01
17	**state.NH**	0.50	0.50	0.01
18	**loan_type.USDA**	0.49	0.49	0.01
19	**loan_purpose.Home Improvement**	0.48	0.48	0.01
20	**loan_type.VA**	0.47	0.47	0.01
21	**doc_type.NULL**	0.47	0.47	0.01
22	**io_flag..N**	0.47	0.47	0.01
23	**loan_type.FHA**	0.47	0.47	0.01
24	**deedlieu**	0.47	0.47	0.01
25	**occupancy_type.Second Home**	0.46	0.46	0.01
26	fc_stage.F1 - File Referred	0.46	0.46	0.01
27	fc_stage.F3	0.46	0.46	0.01
28	doc_type.Stated	0.46	0.46	0.01
29	state.VA	0.46	0.46	0.01
30	state.CO	0.46	0.46	0.01
31	state.ME	0.45	0.45	0.01
32	fc_stage.F2 - First Legal Filed	0.45	0.45	0.01
33	loan_type.HELOC	0.45	0.45	0.01
34	state.IN	0.45	0.45	0.01
35	balloon_flag.N	0.45	0.45	0.01
36	io_flag.Y	0.45	0.45	0.01
37	state.NE	0.45	0.45	0.01
38	condemn	0.45	0.45	0.01
39	state.AK	0.44	0.44	0.01
...	...	...	...	...
200	original_rate	0.11	0.11	0

Table 37: Out-of-Sample Misclassification Errors for Loan Data and Servicer Comments.

	DRF	Lasso	GBM	DNN
NPL	0.02	0.27	0.02	0.08
PIF	0.05	1	0.08	0.09
PL	0	0	0	0
ShortSale	0.06	1	0.02	0.04

Table 38: In-Sample Confusion Matrices for Granular Classes using Loan Data.

	DRF	Lasso	GBM	DNN
B120D	0.04	0.51	0.04	0.15
BK	0.01	0.02	0	0.02
FC	0.01	0.03	0	0.02
PIF	0.06	0.17	0.06	0.13
REO	0	0.04	0	0.01
ShrtSal	0	0.41	0.02	0.16
W0_30D	0.01	0.01	0.01	0.02
W30_60D	0.60	1	0.45	0.7
W60_90D	0.40	1	0.25	0.67
W90_120D	0.29	1	0.25	0.62

Table 39: Variable Importance using DRF for Granular Classes using Loan Data.

	variable	relative_importance	scaled_importance	percentage
1	**next_pay_duedate**	613035.06	1	0.09
2	**reo_flag**	605082.88	0.99	0.09
3	**fc_stage**	544768.25	0.89	0.08
4	**intrst_payment_due**	505585.44	0.82	0.07
5	**bk_flag**	449212.62	0.73	0.06
6	**state**	411984.72	0.67	0.06
7	**fc_flag**	367797.50	0.60	0.05
8	**loan_type**	349388.12	0.57	0.05
9	**bk_dq_delay**	233447.95	0.38	0.03
10	**legal_grade**	197225.72	0.32	0.03
11	**rem_term**	170722.06	0.28	0.02
12	**original_fico**	158756.31	0.26	0.02
13	**fc_start_date**	153712.53	0.25	0.02
14	**current_balance**	138749.39	0.23	0.02
15	**jr_lien_balance**	133924.64	0.22	0.02
16	**current_fico**	130835.15	0.21	0.02
17	**bk_file_date**	117151.84	0.19	0.02
18	**mths_in_bk**	115025.51	0.19	0.02
19	**current_fico_date**	113643.59	0.19	0.02
20	**loan_purpose**	110146.92	0.18	0.02
21	current_rate	105895.52	0.17	0.02
22	loan_age	102130.12	0.17	0.01
23	value	97924.15	0.16	0.01
24	orig_ltv	95359.88	0.16	0.01
25	mths_in_fc	88330.61	0.14	0.01
26	original_balance	81280.03	0.13	0.01
27	original_appraisal	79568.83	0.13	0.01
28	original_pi	79296.91	0.13	0.01
29	current_pi	79163.44	0.13	0.01
30	original_rate	76294.62	0.12	0.01
31	mths_since_mod	58970.69	0.10	0.01
32	mod_date	46266.05	0.08	0.01
33	pmts12	46205.67	0.08	0.01
34	pmts6	45511.46	0.07	0.01
35	doc_type	44127	0.07	0.01
36	mortgage_ins	42250.52	0.07	0.01
37	pmts3	40947.18	0.07	0.01
38	pmts1	35087.26	0.06	0.01
...	...	...	...	...
48	balloon_flag	3142.38	0.01	0

Table 40: Variable Importance using Lasso for Granular Classes using Loan Data.

	variable	relative_importance	scaled_importance	percentage
1	**bk_flag.Y**	11.49	1	0.10
2	**reo_flag.Y**	11.34	0.99	0.10
3	**loan_type.NULL**	11.15	0.97	0.10
4	**fc_stage.NULL**	6.62	0.58	0.06
5	**next_pay_duedate**	5.66	0.49	0.05
6	**intrst_payment_due**	5.66	0.49	0.05
7	**fc_flag.Y**	5.50	0.48	0.05
8	**jr_lien_balance**	3.88	0.34	0.03
9	**fc_flag.N**	3.50	0.30	0.03
10	**fc_stage.F1 - File Referred**	3.45	0.30	0.03
11	**loan_purpose.CASH OUT**	3.43	0.30	0.03
12	**original_fico**	2.80	0.24	0.02
13	**legal_grade.NI**	2.72	0.24	0.02
14	**mod_flag.Y**	2.55	0.22	0.02
15	**loan_type.CONV**	2.41	0.21	0.02
16	**mod_flag.N**	2.36	0.21	0.02
17	**fc_stage.F0 - Removed**	1.88	0.16	0.02
18	**loan_type.USDA**	1.75	0.15	0.02
19	**mths_since_mod**	1.58	0.14	0.01
20	**doc_type.None**	1.14	0.10	0.01
21	**mortgage_ins**	1.08	0.09	0.01
22	**fc_start_date**	0.93	0.08	0.01
23	rem_term	0.90	0.08	0.01
24	bk_dq_delay	0.89	0.08	0.01
25	current_fico_date	0.88	0.08	0.01
26	current_fico	0.87	0.08	0.01
27	legal_grade.C-	0.82	0.07	0.01
28	legal_grade.B+	0.77	0.07	0.01
29	orig_ltv	0.74	0.06	0.01
30	legal_grade.C+	0.74	0.06	0.01
31	legal_grade.C	0.73	0.06	0.01
32	lien_position	0.70	0.06	0.01
33	mod_date	0.64	0.06	0.01
34	bk_file_date	0.63	0.05	0.01
...	...	...	...	...
84	state.FL	0.01	0	0

Table 41: Variable Importance using GBM for Granular Classes using Loan Data.

	variable	relative_importance	scaled_importance	percentage
1	**intrst_payment_due**	143950.55	1	0.21
2	**bk_flag**	109717.62	0.76	0.16
3	**reo_flag**	89517.88	0.62	0.13
4	**fc_flag**	84128.67	0.58	0.12
5	**jr_lien_balance**	74688.28	0.52	0.11
6	**loan_type**	39176.86	0.27	0.06
7	**next_pay_duedate**	24229.26	0.17	0.04
8	**state**	23780.70	0.17	0.03
9	**fc_stage**	17169.37	0.12	0.02
10	**mortgage_ins**	10481.79	0.07	0.02
11	**rem_term**	9947.15	0.07	0.01
12	**loan_purpose**	9337.47	0.06	0.01
13	**pmts1**	7182.65	0.05	0.01
14	**legal_grade**	6116.16	0.04	0.01
15	**current_balance**	5038.95	0.04	0.01
16	**pmts6**	3144.23	0.02	0
17	current_fico	3118.81	0.02	0
18	mths_since_mod	2318.51	0.02	0
19	pmts3	2260.23	0.02	0
20	pmts12	2115.03	0.01	0
21	original_fico	1853.44	0.01	0
22	current_fico_date	1847.31	0.01	0
23	orig_ltv	1775.01	0.01	0
24	current_rate	1687.95	0.01	0
25	loan_age	1652.90	0.01	0
26	original_balance	1546.06	0.01	0
27	value	1387.33	0.01	0
28	doc_type	1236.83	0.01	0
29	original_pi	1224.34	0.01	0
30	original_rate	1185.96	0.01	0
31	original_appraisal	1067.47	0.01	0
32	mod_date	977.69	0.01	0
33	bk_dq_delay	797.64	0.01	0
34	bk_file_date	539.07	0	0
35	fc_start_date	528.87	0	0
36	mths_in_bk	490.36	0	0
...	...	...	...	...
48	io_flag	0	0	0

Table 42: Variable Importance using DNN for Granular Classes using Loan Data.

	variable	relative_importance	scaled_importance	percentage
1	**bk_flag.N**	1	1	0.02
2	**intrst_payment_due**	0.89	0.89	0.02
3	**reo_flag.Y**	0.86	0.86	0.02
4	**reo_flag.N**	0.84	0.84	0.01
5	**next_pay_duedate**	0.76	0.76	0.01
6	**bk_flag.Y**	0.73	0.73	0.01
7	**fc_flag.N**	0.73	0.73	0.01
8	**fc_flag.Y**	0.70	0.70	0.01
9	**jr_lien_balance**	0.66	0.66	0.01
10	**loan_type.NULL**	0.57	0.57	0.01
11	**fc_stage.F2 - First Legal Filed**	0.50	0.50	0.01
12	**io_flag.N**	0.49	0.49	0.01
13	**loan_type.CONV**	0.48	0.48	0.01
14	**loan_type.HELOC**	0.48	0.48	0.01
15	**state.CT**	0.48	0.48	0.01
16	**state.RI**	0.47	0.47	0.01
17	**fc_stage.F1 - File Referred**	0.46	0.46	0.01
18	**state.ND**	0.45	0.45	0.01
19	**fc_stage.F1**	0.45	0.45	0.01
20	**fc_stage.F3 - Judgment Entered**	0.45	0.45	0.01
21	occupancy_type.Owner Occupied	0.45	0.45	0.01
22	fc_stage.F3	0.44	0.44	0.01
23	loan_purpose.Home Improvement	0.44	0.44	0.01
24	fc_stage.F0 - No FC	0.44	0.44	0.01
25	fc_stage.F2	0.44	0.44	0.01
26	state.WV	0.43	0.43	0.01
27	legal_grade.D	0.43	0.43	0.01
28	state.WI	0.43	0.43	0.01
29	pmts3	0.42	0.42	0.01
30	io_flag..Y	0.42	0.42	0.01
31	state.IN	0.42	0.42	0.01
32	state.PR	0.42	0.42	0.01
33	state.MD	0.42	0.42	0.01
34	current_rate_type.ARM	0.41	0.41	0.01
35	mod_flag.N	0.41	0.41	0.01
36	legal_grade.D-	0.41	0.41	0.01
...	...	...	...	...
152	current_balance	0.14	0.14	0

Table 43: Out-of-Sample Confusion Matrices for Granular Classes using Loan Data.

	DRF	Lasso	GBM	DNN
B120D	0.08	0.51	0.06	0.15
BK	0.01	0.02	0.01	0.02
FC	0	0.02	0.01	0.03
PIF	0.06	0.17	0.07	0.13
REO	0.01	0.03	0	0.01
ShrtSal	0.08	0.42	0.05	0.2
W0_30D	0	0.01	0.01	0.02
W30_60D	0.74	1	0.51	0.71
W60_90D	0.62	1	0.51	0.68
W90_120D	0.49	1	0.36	0.62

Table 44: In-Sample Confusion Matrices for Granular Classes using Loan Data and Servicer Comments.

	DRF	Lasso	GBM	DNN
B120D	0.06	0.5	0.04	1
BK	0.01	0.02	0	1
FC	0.01	0.03	0.01	1
PIF	0.08	0.17	0.06	1
REO	0	0.03	0	1
ShrtSal	0.01	0	0.03	1
W0_30D	0	0.01	0.01	0
W30_60D	0.65	1	0.44	1
W60_90D	0.47	1	0.41	1
W90_120D	0.36	1	0.27	1

Table 45: Variable Importance using DRF for Granular Classes using Loan Data and Servicer Comments.

	variable	relative_importance	scaled_importance	percentage
1	**fc_stage**	465347.91	1	0.09
2	**reo_flag**	436792.97	0.94	0.08
3	**intrst_payment_due**	429502.91	0.92	0.08
4	**next_pay_duedate**	388325.16	0.83	0.07
5	**bk_flag**	335261.22	0.72	0.06
6	**state**	306916.88	0.66	0.06
7	**fc_flag**	298051.41	0.64	0.06
8	**loan_type**	253430.25	0.54	0.05
9	**bk_dq_delay**	183503.30	0.39	0.03
10	**legal_grade**	136493.88	0.29	0.03
11	**rem_term**	123691.41	0.27	0.02
12	**original_fico**	110465.33	0.24	0.02
13	**current_balance**	99567.46	0.21	0.02
14	**current_fico**	95718.16	0.21	0.02
15	**bk_file_date**	92567.76	0.20	0.02
16	**fc_start_date**	88275.17	0.19	0.02
17	**jr_lien_balance**	81285.50	0.17	0.02
18	**current_rate**	76209.28	0.16	0.01
19	**current_fico_date**	75858.84	0.16	0.01
20	**loan_age**	73657.98	0.16	0.01
21	**mths_in_bk**	72002.57	0.15	0.01
22	**mths_in_fc**	71213.42	0.15	0.01
23	value	69008	0.15	0.01
24	orig_ltv	68509.26	0.15	0.01
25	loan_purpose	67997.71	0.15	0.01
26	original_balance	59834.68	0.13	0.01
27	current_pi	58469.17	0.13	0.01
28	original_appraisal	58350.36	0.13	0.01
29	original_pi	57792.35	0.12	0.01
30	original_rate	57703.61	0.12	0.01
31	mths_since_mod	43540.52	0.09	0.01
32	mod_date	36252.40	0.08	0.01
33	doc_type	33028.20	0.07	0.01
34	pmts12	32731.72	0.07	0.01
35	pmts6	32389.83	0.07	0.01
36	mortgage_ins	30166.88	0.06	0.01
37	pmts3	28953.71	0.06	0.01
...	...	...	...	...
85	manufacturehouse	0.01	0	0

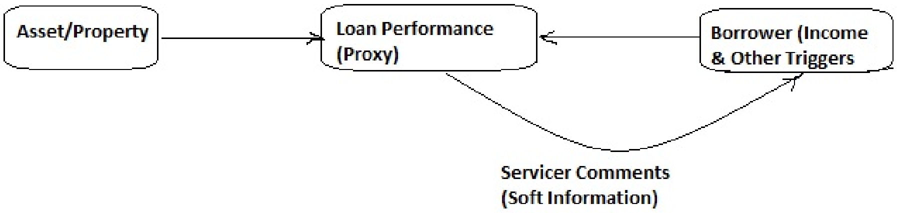
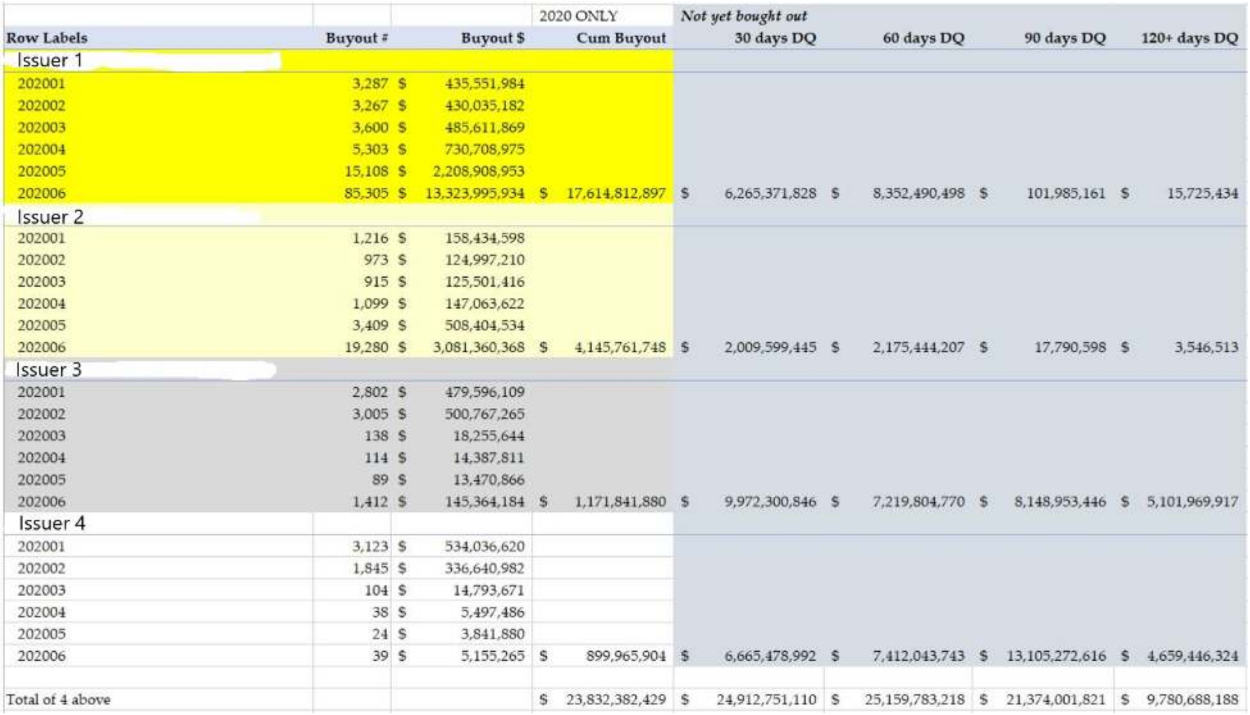

Asset/Property
Loan Performance (Proxy)
Borrower (Income & Other Triggers
Servicer Comments (Soft Information)

Figure 38: Servicer Comment as a channel.

Table 46: Variable Importance using Lasso for Granular Classes using Loan Data and Servicer Comments.

	variable	relative_importance	scaled_importance	percentage
1	**reo_flag.Y**	11.84	1	0.10
2	**loan_type.NULL**	11.16	0.94	0.10
3	**bk_flag.Y**	11.01	0.93	0.09
4	**fc_stage.NULL**	6.72	0.57	0.06
5	**next_pay_duedate**	5.66	0.48	0.05
6	**intrst_payment_due**	5.66	0.48	0.05
7	**fc_flag.Y**	5.63	0.48	0.05
8	**jr_lien_balance**	3.77	0.32	0.03
9	**loan_purpose.CASH OUT**	3.35	0.28	0.03
10	**fc_flag.N**	3.34	0.28	0.03
11	**fc_stage.F1 - File Referred**	3.30	0.28	0.03
12	**original_fico**	2.79	0.24	0.02
13	**legal_grade.NI**	2.62	0.22	0.02
14	**mod_flag.Y**	2.59	0.22	0.02
15	**mod_flag.N**	2.33	0.20	0.02
16	**loan_type.CONV**	2.11	0.18	0.02
17	**fc_stage.F0 - Removed**	1.83	0.15	0.02
18	**mths_since_mod**	1.47	0.12	0.01
19	**loan_type.USDA**	1.28	0.11	0.01
20	doc_type.None	1.14	0.10	0.01
21	mortgage_ins	1.06	0.09	0.01
22	rem_term	0.90	0.08	0.01
23	fc_start_date	0.89	0.08	0.01
24	current_fico_date	0.86	0.07	0.01
25	bk_dq_delay	0.86	0.07	0.01
26	current_fico	0.85	0.07	0.01
27	legal_grade.C-	0.78	0.07	0.01
28	legal_grade.B+	0.77	0.07	0.01
29	lien_position	0.75	0.06	0.01
30	orig_ltv	0.74	0.06	0.01
31	legal_grade.C+	0.73	0.06	0.01
32	legal_grade.C	0.71	0.06	0.01
33	reo	0.70	0.06	0.01
34	bk_file_date	0.68	0.06	0.01
35	fc	0.63	0.05	0.01
...	...	...	...	...
113	hurricanedamage	0.01	0	0



			2020 ONLY	*Not yet bought out*			
Row Labels	**Buyout #**	**Buyout $**	**Cum Buyout**	**30 days DQ**	**60 days DQ**	**90 days DQ**	**120+ days DQ**
Issuer 1 [redacted]							
202001	3,287	$ 435,551,984					
202002	3,267	$ 430,035,182					
202003	3,600	$ 485,611,869					
202004	5,303	$ 730,708,975					
202005	15,108	$ 2,208,908,953					
202006	85,305	$ 13,323,995,934	$ 17,614,812,897	$ 6,265,371,828	$ 8,352,490,498	$ 101,985,161	$ 15,725,434
Issuer 2 [redacted]							
202001	1,216	$ 158,434,598					
202002	973	$ 124,997,210					
202003	915	$ 125,501,416					
202004	1,099	$ 147,063,622					
202005	3,409	$ 508,404,534					
202006	19,280	$ 3,081,360,368	$ 4,145,761,748	$ 2,009,599,445	$ 2,175,444,207	$ 17,790,598	$ 3,546,513
Issuer 3 [redacted]							
202001	2,802	$ 479,596,109					
202002	3,005	$ 500,767,265					
202003	138	$ 18,255,644					
202004	114	$ 14,387,811					
202005	89	$ 13,470,866					
202006	1,412	$ 145,364,184	$ 1,171,841,880	$ 9,972,300,846	$ 7,219,804,770	$ 8,148,953,446	$ 5,101,969,917
Issuer 4							
202001	3,123	$ 534,036,620					
202002	1,845	$ 336,640,982					
202003	104	$ 14,793,671					
202004	38	$ 5,497,486					
202005	24	$ 3,841,880					
202006	39	$ 5,155,265	$ 899,965,904	$ 6,665,478,992	$ 7,412,043,743	$ 13,105,272,616	$ 4,659,446,324
Total of 4 above			$ 23,832,382,429	$ 24,912,751,110	$ 25,159,783,218	$ 21,374,001,821	$ 9,780,688,188

Figure 39: Opportunities in EBO in 2020

Table 47: Variable Importance using GBM for Granular Classes using Loan Data and Servicer Comments.

	variable	relative_importance	scaled_importance	percentage
1	**intrst_payment_due**	143319.83	1	0.21
2	**bk_flag**	109700.21	0.77	0.16
3	**reo_flag**	88941.77	0.62	0.13
4	**fc_flag**	84212.17	0.59	0.12
5	**jr_lien_balance**	74120.73	0.52	0.11
6	**loan_type**	39102.08	0.27	0.06
7	**next_pay_duedate**	24199.80	0.17	0.04
8	**state**	22955.12	0.16	0.03
9	**fc_stage**	17683.63	0.12	0.03
10	**mortgage_ins**	10940.99	0.08	0.02
11	**rem_term**	9655.52	0.07	0.01
12	**loan_purpose**	9092.03	0.06	0.01
13	**pmts1**	7104.02	0.05	0.01
14	**legal_grade**	5958.43	0.04	0.01
15	**current_balance**	4809.61	0.03	0.01
16	**current_fico**	2925.20	0.02	0
17	**pmts6**	2754.50	0.02	0
18	**mths_since_mod**	2202.12	0.02	0
19	**original_fico**	1960.34	0.01	0
20	current_fico_date	1940.18	0.01	0
21	pmts3	1844.07	0.01	0
22	orig_ltv	1829.55	0.01	0
23	pmts12	1825.18	0.01	0
24	current_rate	1674.95	0.01	0
25	loan_age	1674.73	0.01	0
26	mths_in_fc	1654.59	0.01	0
27	original_balance	1517.98	0.01	0
28	value	1291.51	0.01	0
29	doc_type	1283.80	0.01	0
30	fc	1279.58	0.01	0
31	original_rate	1222.97	0.01	0
32	mod_flag	1151.75	0.01	0
33	original_pi	1134.77	0.01	0
34	original_appraisal	1085.89	0.01	0
35	mod_date	987.81	0.01	0
36	fc_start_date	945.30	0.01	0
37	bk_dq_delay	724.05	0.01	0
38	lien_position	566.66	0	0
39	shortsale	496.73	0	0
40	bk_file_date	496.59	0	0
41	property_type	441.45	0	0
42	mths_in_bk	404.50	0	0
43	io_flag.	289.98	0	0
44	occupancy_type	209.18	0	0
45	fcl	195.96	0	0
46	foreclosure	194.11	0	0
47	bk	137.10	0	0
48	fire	136.27	0	0
49	balloon_flag	117.66	0	0
...	...	...	...	...
72	deedlieu	0.03	0	0

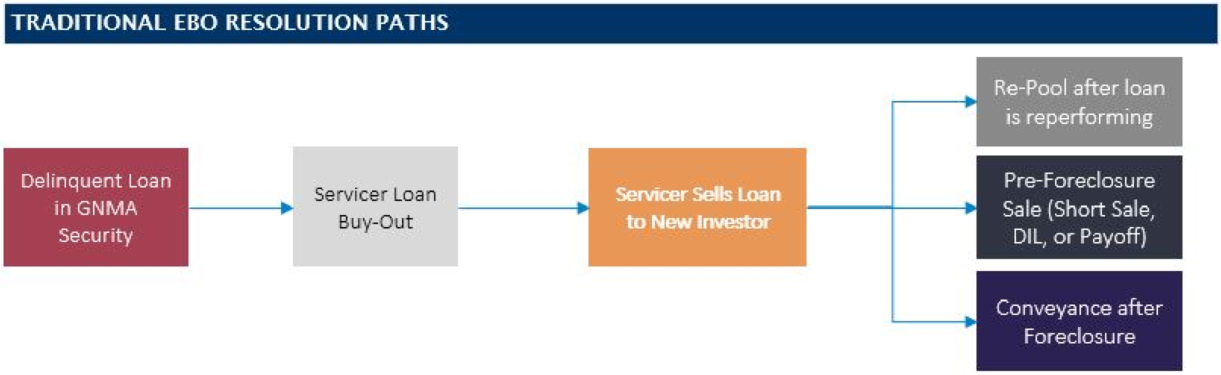

TRADITIONAL EBO RESOLUTION PATHS
Delinquent Loan in GNMA Security
Servicer Loan Buy-Out
Servicer Sells Loan to New Investor
Re-Pool after loan is reperforming
Pre-Foreclosure Sale (Short Sale, DIL, or Payoff)
Conveyance after Foreclosure

Figure 40: Resolution of Early-Buyout loans

Table 48: Variable Importance using DNN for Granular Classes using Loan Data and Servicer Comments.

	variable	relative_importance	scaled_importance	percentage
1	**divorce**	1	1	0.01
2	**rem_term**	0.99	0.99	0.01
3	**reo_flag.N**	0.98	0.98	0.01
4	**fc_stage.F3** - Judgment Entered	0.97	0.97	0.01
5	**pmts3**	0.94	0.94	0.01
6	**state.MA**	0.94	0.94	0.01
7	**bk_flag.N**	0.92	0.92	0.01
8	**state.AR**	0.91	0.91	0.01
9	**state.LA**	0.91	0.91	0.01
10	**bk_flag.Y**	0.90	0.90	0.01
11	**fraud**	0.90	0.90	0.01
12	**property_type.Condo**	0.90	0.90	0.01
13	**state.AL**	0.89	0.89	0.01
14	**current_rate**	0.89	0.89	0.01
15	**property_type.Modular Home**	0.89	0.89	0.01
16	**fc_stage.F0 - Removed**	0.89	0.89	0.01
17	**loan_purpose.Cash Out**	0.89	0.89	0.01
18	**lienstrip**	0.88	0.88	0.01
19	**doc_type.Reduced**	0.88	0.88	0.01
20	**loan_type.CONV**	0.88	0.88	0.01
21	state.OK	0.88	0.88	0.01
22	doc_type.Limited	0.88	0.88	0.01
23	legal_grade.D-	0.87	0.87	0.01
24	orig_ltv	0.87	0.87	0.01
25	state.PA	0.87	0.87	0.01
26	state.NM	0.87	0.87	0.01
27	legal_grade.D+	0.87	0.87	0.01
28	bankruptcy	0.87	0.87	0.01
29	state.CA	0.87	0.87	0.01
30	value	0.87	0.87	0.01
31	state.NJ	0.86	0.86	0.01
32	occupancy_type.Second Home	0.86	0.86	0.01
33	state.WY	0.86	0.86	0.01
34	damage	0.86	0.86	0.01
35	legal_grade.B+	0.86	0.86	0.01
...	...	...	...	...
196	fcl	0.46	0.46	0

Table 49: Out-of-Sample Confusion Matrices for Granular Classes using Loan Data and Servicer Comments.

	DRF	Lasso	GBM	DNN
B120D	0.06	0.45	0.04	1
BK	0	0.02	0	1
FC	0	0.01	0	1
PIF	0.04	0.16	0.06	1
REO	0.01	0.02	0	1
ShrtSal	0.08	0.42	0.05	1
W0_30D	0	0.01	0.01	0
W30_60D	0.72	1	0.49	1
W60_90D	0.57	1	0.43	1
W90_120D	0.43	1	0.27	1

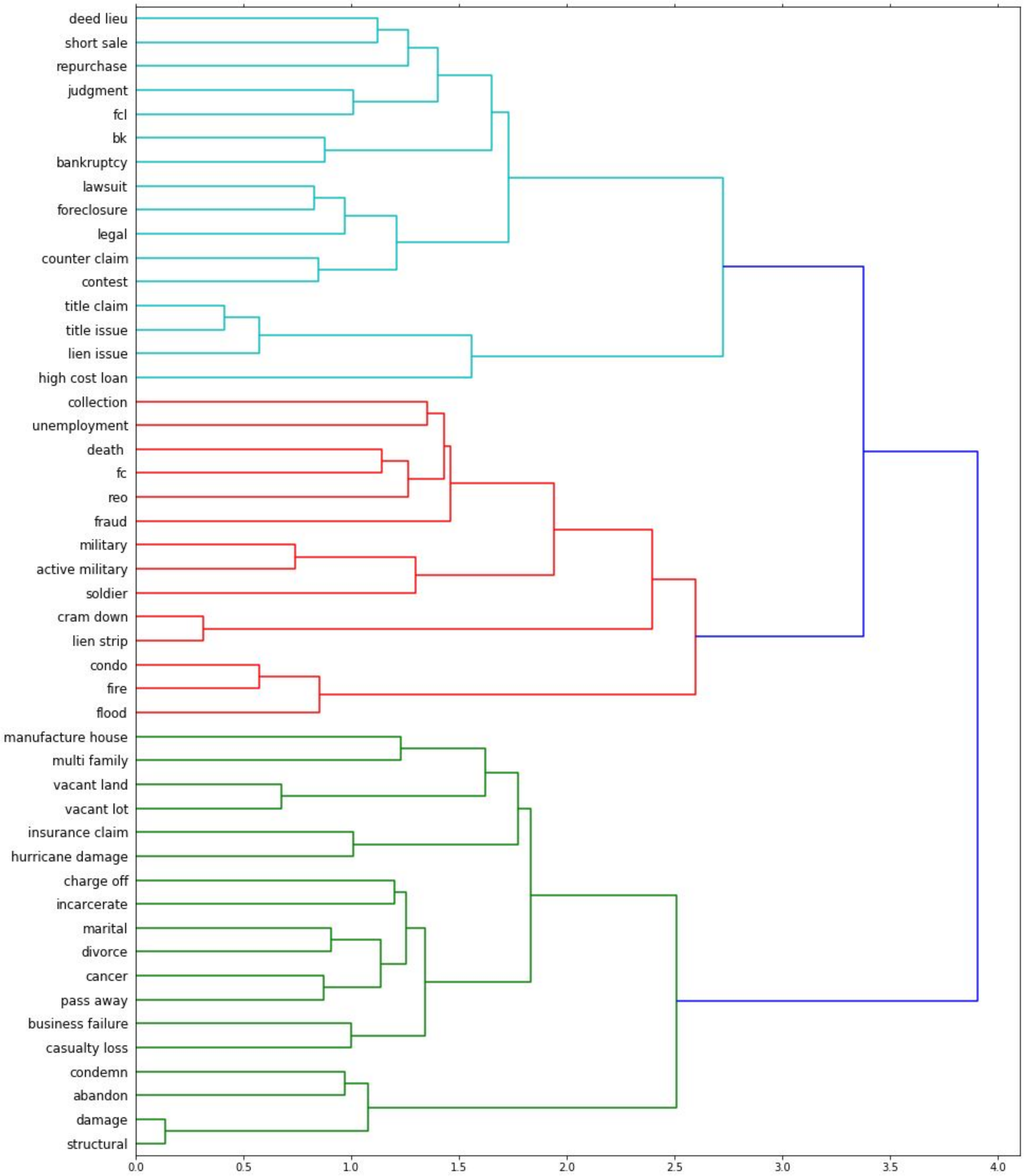


Figure 41: Dendogram of Keywords.

Table 50: Categories and Corresponding Words

Category	Corresponding Words
NATURAL DISASTER	SANDY, IRENE, HURRICANE, TORNADO, HAZARD POL, HAZARD/FLOOD INSURANCE, HAZARD LINE, HAZARD DEC PAGE, HAZARD MONTHLY STATEMENT, HAZARD MONTHLY AMOUNT, HAZARD DECLARATION, HAZARD DOC , OBRT FIRE, STATE FARM FIRE, FIRE INSURANCE, FIRE POLICY, NON-FLOOD, FLOOD PANEL, FLOOD CERT, ESCROW SET UP AND FLOOD FOR MODIFICATION, FLOOD AUDIT, FEMA, SINK HOLE.
MILITARY	ATTORNEY GENERAL, DEPLOY, SCRA, MILITARY+SERVICE, SOLDIER, ACTIVE MILITARY, SAILOR,
BANKRUPTCY	BANKRUPTCY, CRAM DOWN
TITLE ISSUE	TITLE ISSUE RESOLVED, CONTESTED FORECLOSURE, VENDOR'S LIEN, HOMESTEAD DEED, %INSTR%TAKING% , %CERT%SALE%, %TAX CERT%, %VENDOR LIEN%, %TAX%DEED%, %TAX%SALE%, %QUIET%TITLE%, %ORDER OF NOTICE%, INTENDED JUNIOR, CONTRACT FOR DEED, AGREEMENT FOR DEED, LAND SALE CONTRACT, REAL ESTATE INSTALLMENT AGREEMENT, LAND CONTRACT, CONTRACT SALE, BOND FOR TITLE, INSTALLMENT SALES, CONTRACT, NOT RELEASE, HOLD, SERVICE RELEASE, RELEASE FUNDS, INFORMATION RELEASE, RELEASE FORM, CHAIN, CLAIM, COMMERCIAL, ENCROACH, JUDGEMENT, SATISFACTION, QUIT CLAIM DEED, OWNERSHIP TRANSFER, PENDING, TAX LIEN, LIEN ISSUE, LIEN STRIPPING, REPURCHASE, WIPED OUT
OCCUPANCY	ABANDON, UNSECURED, VACANT LAND, VACANT LOT, MISSING, NOT SECURE,
LOSS MITIGATION	CHARGE-OFFS/COLL ON CBR, CHARGE OFF OR COLLECTIONS, COLLECTIONS/CHARGE-OFFS/JUDGMENTS/LIENS, CURTAILMENT OF INCOME, DEED IN LIEU, FORECLOSURE, REO
PROPERTY CONDITION	CONDEMNATION, CONTAMINATION, DESTROY, DAMAGE, TORN DOWN, TEAR DOWN, COLLAPSE, DOMAIN, EMINENT, ENVIRONMENTAL, EXPOSED WIRING, INFEST, INHABITABLE, SECURITY, VANDALISM, SMOKE DAMAGE, STORM DAMAGE, STRUCTURAL, TERMITE, WATER DAMAGE, WIND DAMAGE, WINTERIZE, GROUND, HAIL DAMAGE, HURRICANE DAMAGE, HAZARD LOSS, FREEZE, FOUNDATION, BURN/BURNT/BURNED, DEMOLISH, UNINHAB%, VACANT LAND, VACANT LOT, UNSAFE, ROOF, REPAIR, %GAS%TANK%, CASUALTY+LOSS, LEAK, HISTORICAL LANDMARK, MOLD, OIL, ORDINANCE, POISON,
PROPERTY TYPE	MOBILE HOME, MANUFACTURED HOME, MANUFACTURED HOUSING, MODULAR HOME, TRIBAL LAND, INDIAN LAND, TRIBAL LOAN, INDIAN LOAN, RURAL LAND, MULTIFAMILY, APARTMENT, CONDOMINIUM, PRE%FAB
LIFE EVENT	SICK, ILL, PASSED AWAY, MARITAL, DIVORCE, EMPLOYMENT TRANSFER, UNEMPLOYMENT, BUSINESS+FAILURE, DEATH, DECEASED, CANCER, EMERGENCY, HOSPITAL, MOVE
CRIME	INCARCERATED/INCARCERATE/INCARCERATION, FRAUD, IDENTITY THEFT, ILLEGAL, INVALID, JAIL, THEFT, VANDAL, PRISON.
LEGAL	COMPLIANCE, DEFENSE, DELAY, DISPUTE, DISTRESS, ERROR, INSURANCE CLAIM, SETTLEMENT, SEIZE, SKIP, VIOLATION, ZONING, SERVICING + PROBLEM, UNABLE+TO+CONTACT, PAYMENT ADJUSTMENT, PAYMENT DISPUTE, COLLECTIONS/CHARGE-OFFS/JUDGMENTS/LIENS, CURB APPEAL, FDCPA, COUNTER CLAIM, COUNTERSU%, SANCTIONS, LITIGIOUS, STATUTE OF LIMITATIONS, BARRED BY LIMITATIONS, RESTRICTIVE COVENANT, LAWSUIT, LIABILITY, PREDATORY, PROBATE, REAFFIRMED, AFFIRMATIVE DEFENSE
LOCATION	SALE DATE, SALES PRICE, INABILITY+TO+SELL, INABILITY+TO+RENT, LISTED,
COMPLIANCE	TEXAS 50(A)(6), TEXAS 50A6, TEXAS 50(A), TEXAS 50A, ARTICLE 50(A), CONST. ARTICLE 50(A), CONSTITUTION ARTICLE 50, HIGH COST LOAN,

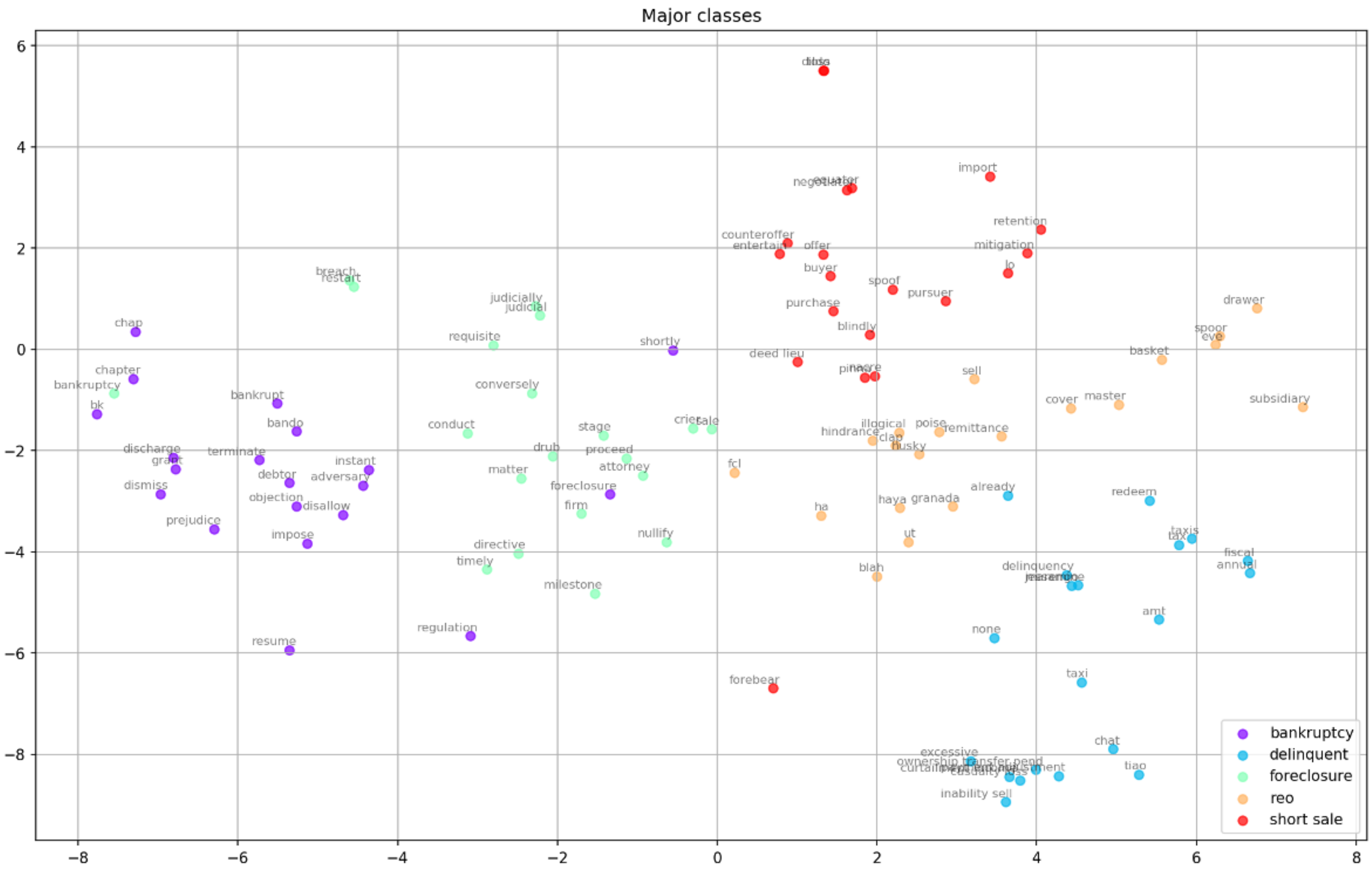


Figure 42: Key Delinquency States.

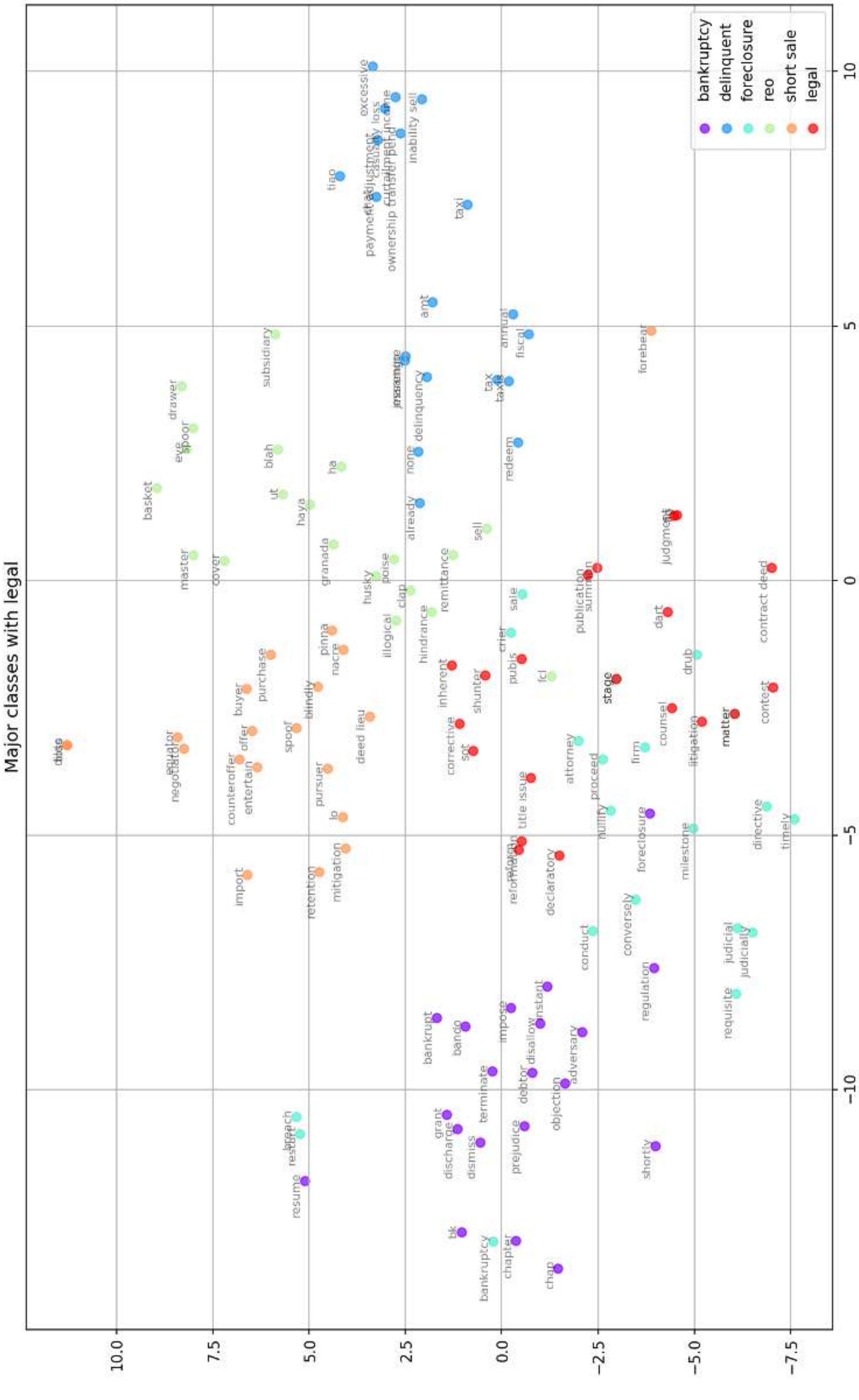


Figure 43: Key Delinquency States and Legal Keyword.

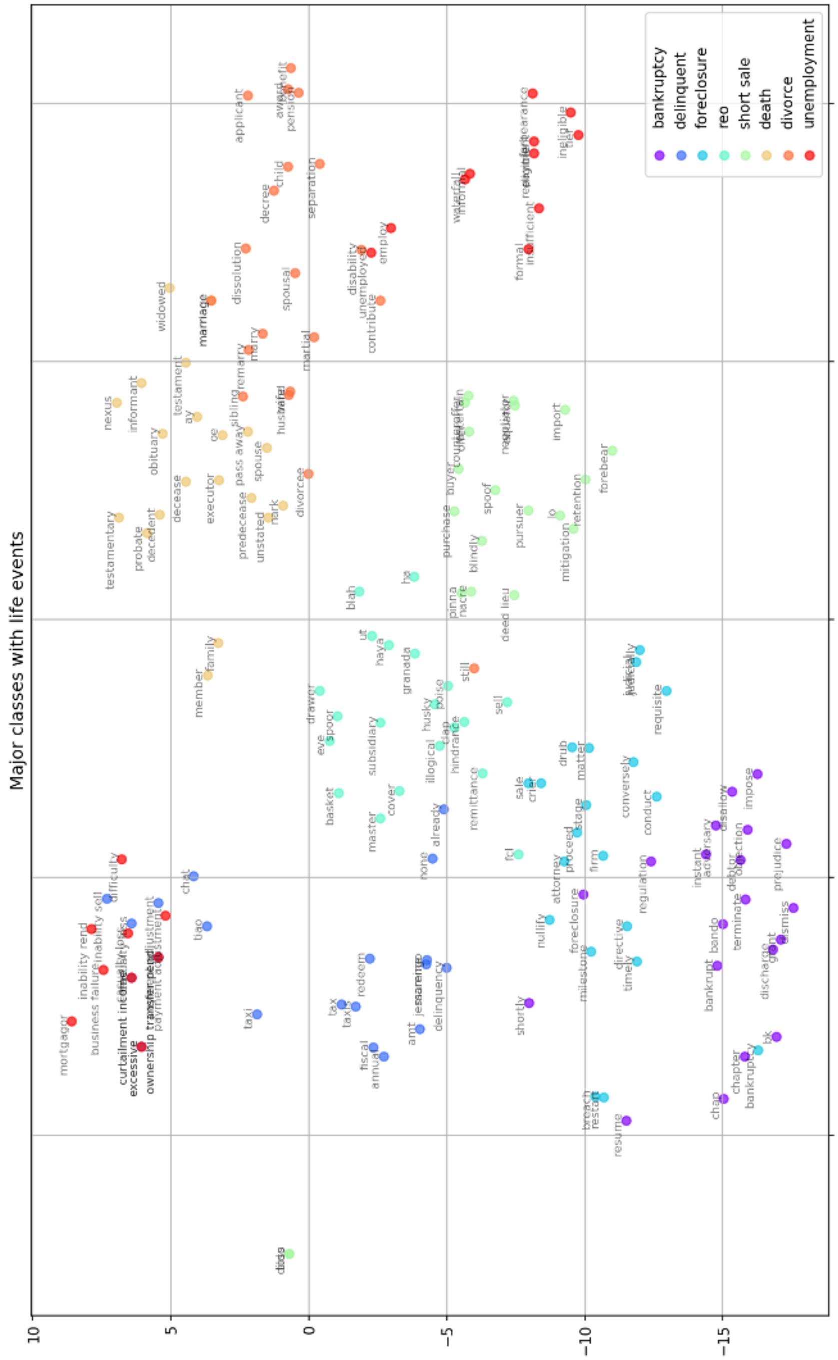


Figure 44: Key Delinquency States and Life Event-related Keywords.

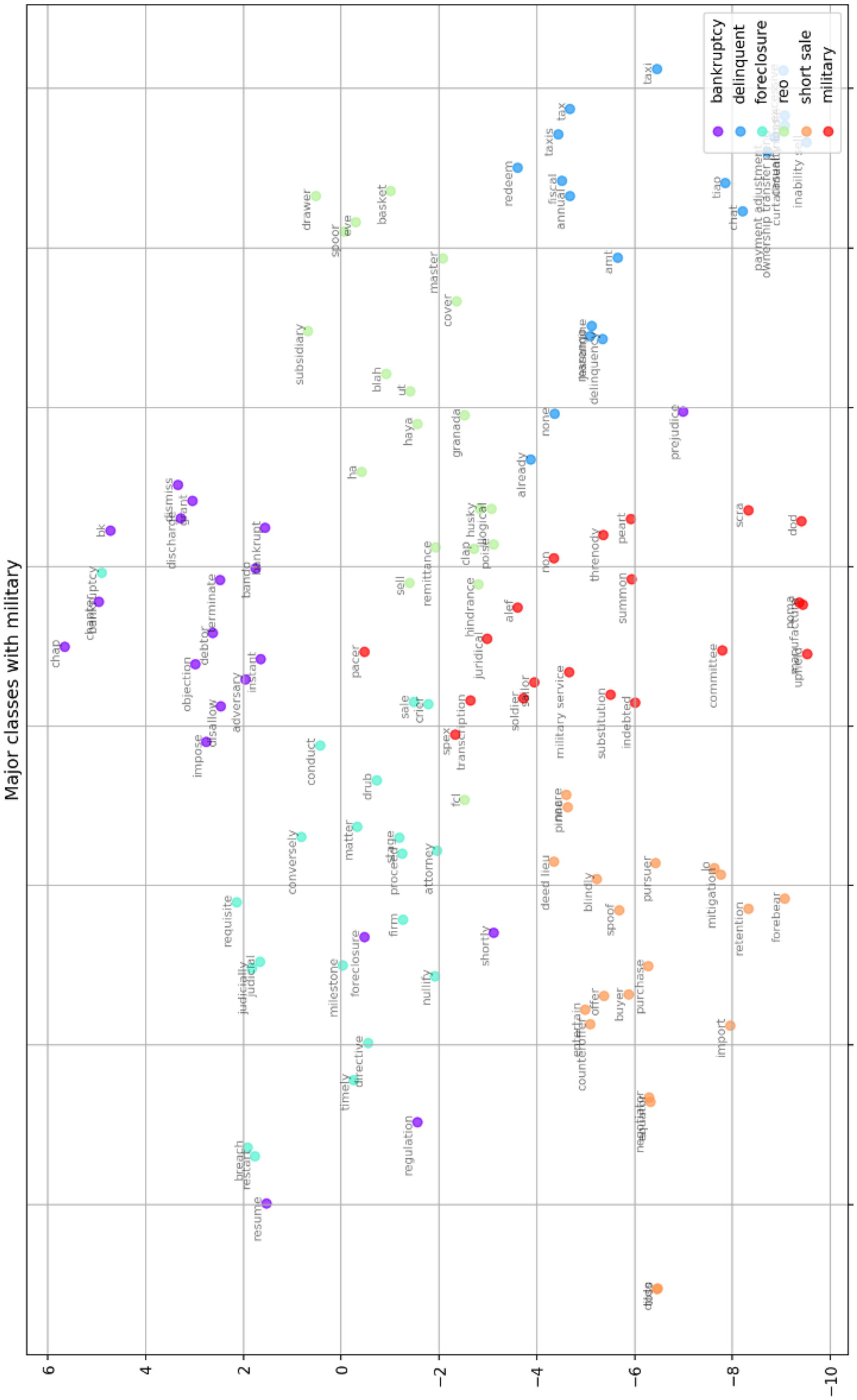


Figure 45: Key Delinquency States and Military Keyword.

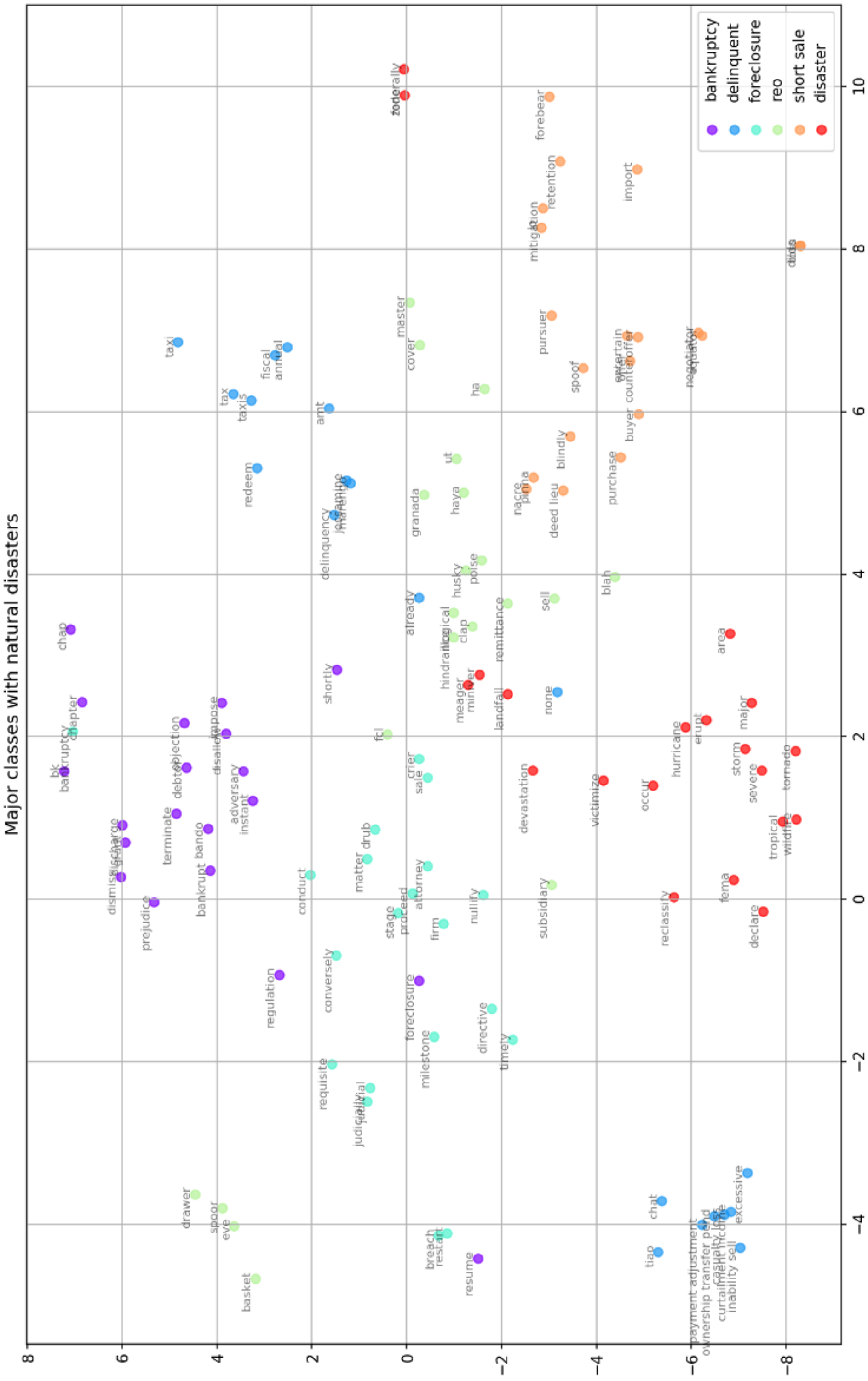


Figure 46: Key Delinquency States and Natural Disasters-related Keywords.

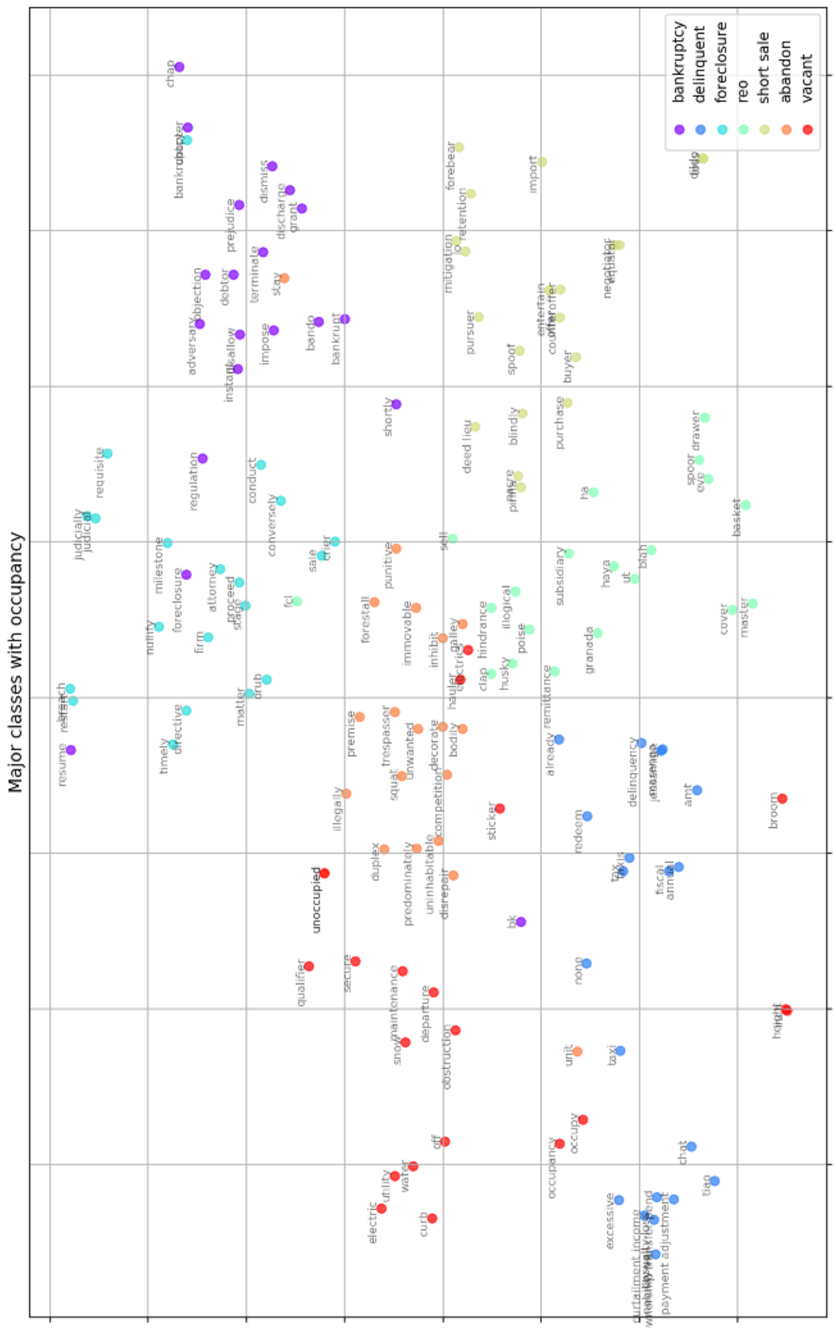


Figure 47: Key Delinquency States and Occupancy-related Keywords.

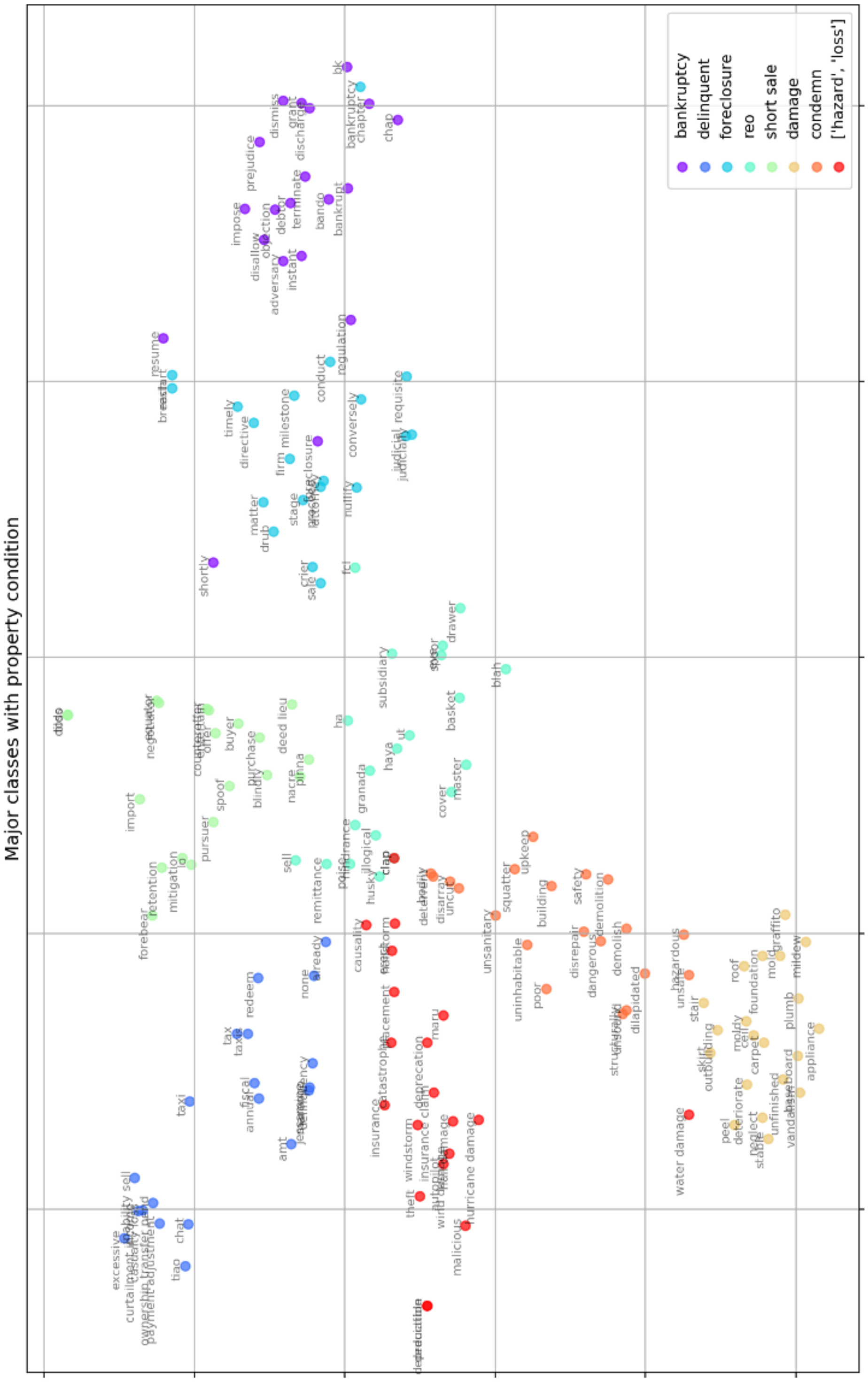


Figure 48: Key Delinquency States and Property Condition-related Keywords.

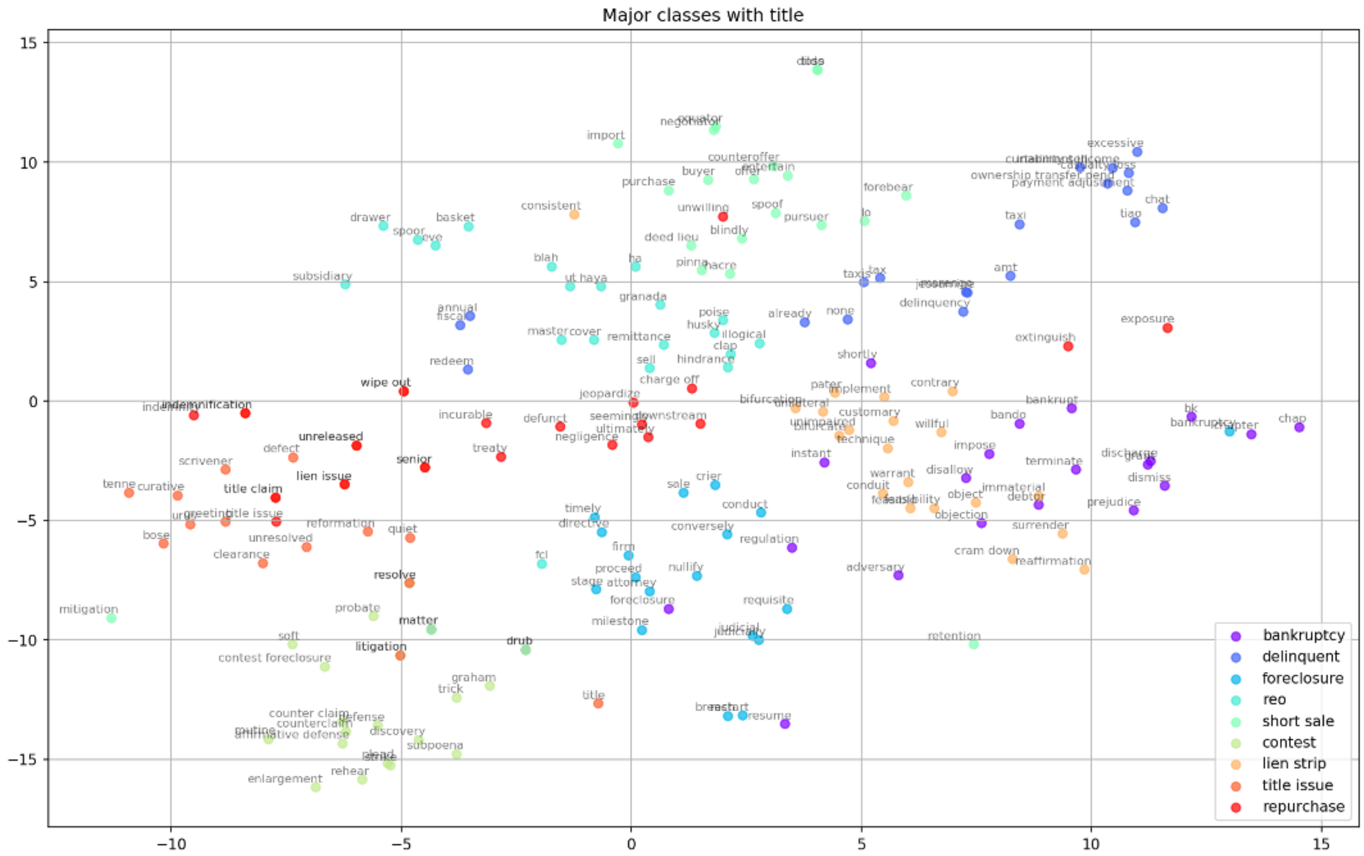


Figure 49: Key Delinquency States and Title-related Keyword.

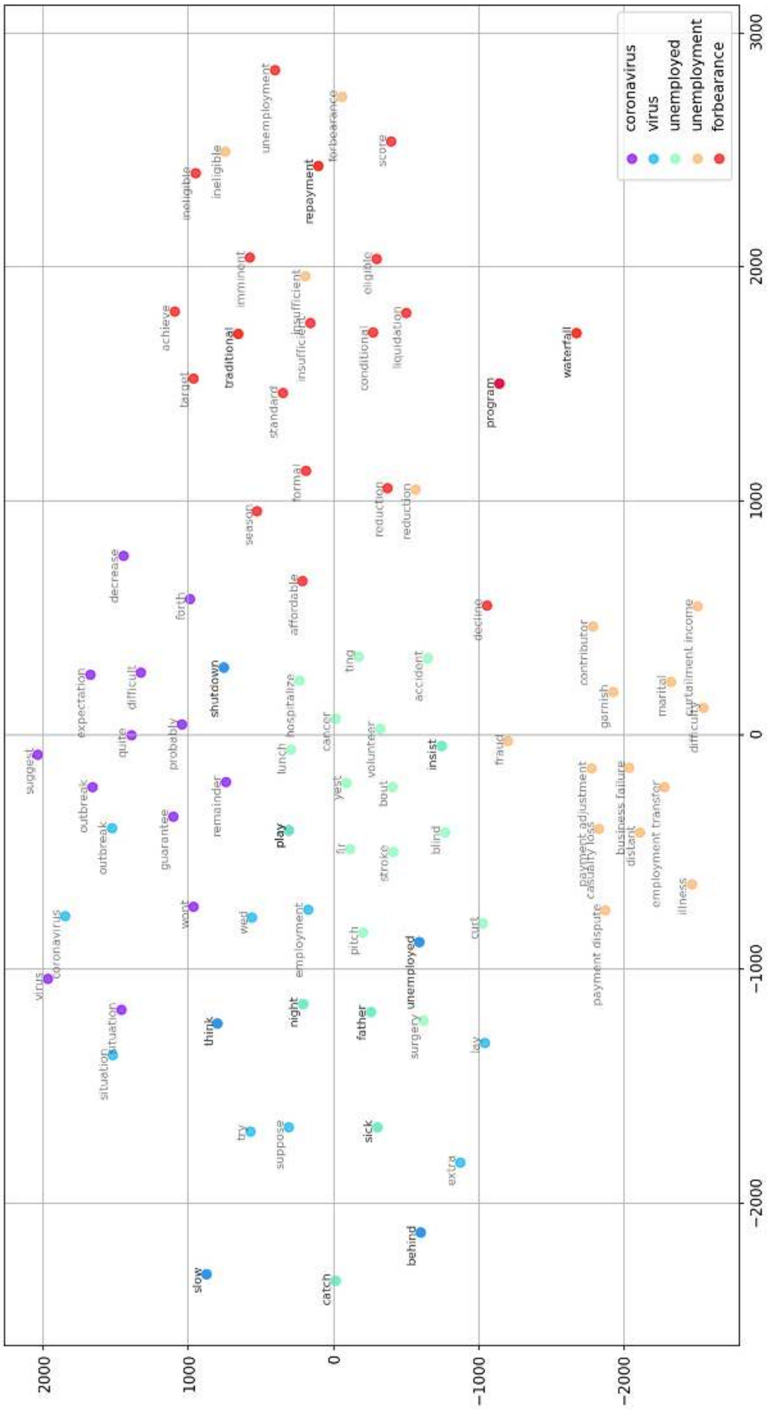


Figure 50: Coronavirus Unemployment Forbearance.

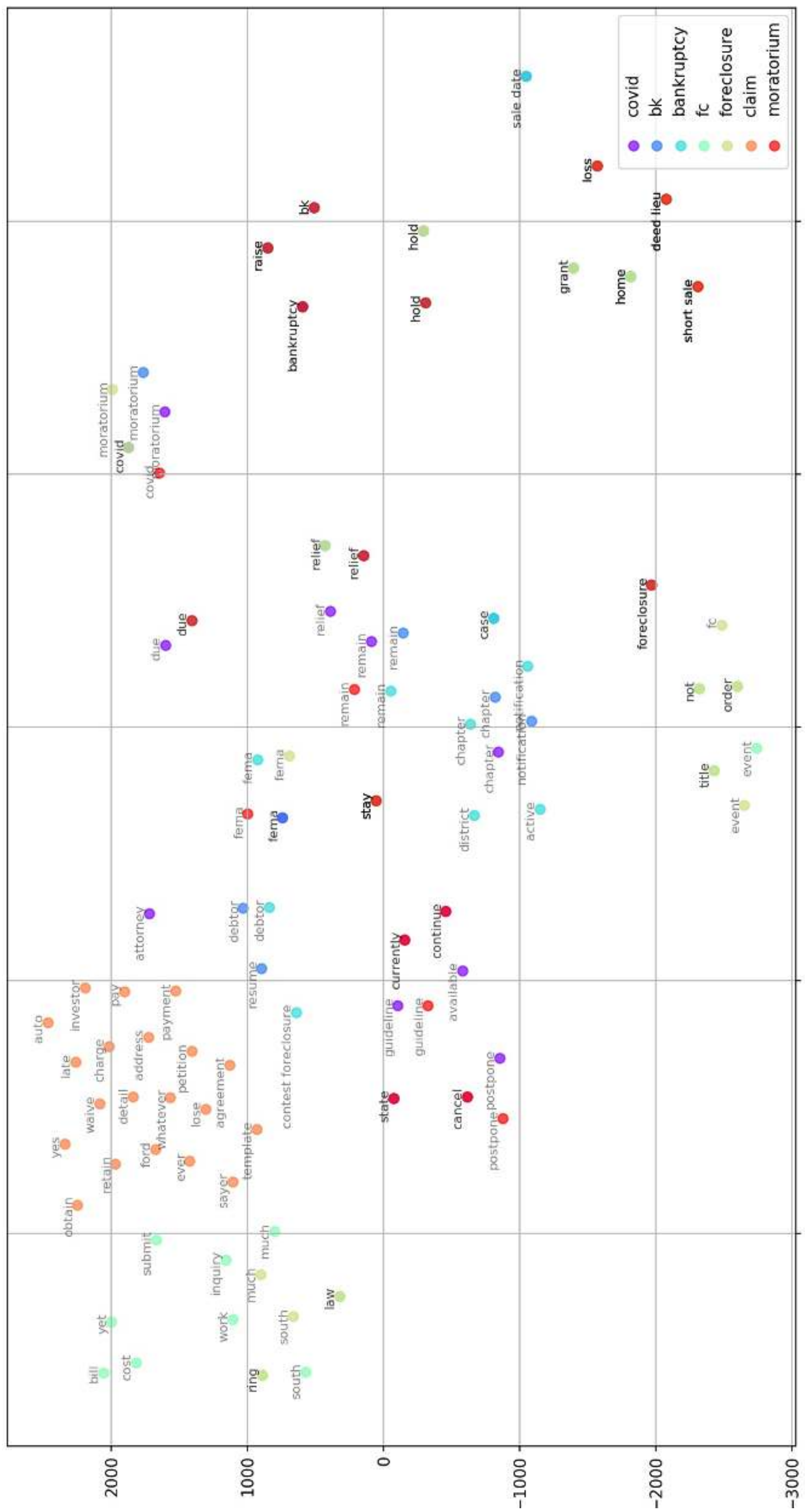


Figure 51: Corona Foreclosure Bankruptcy.

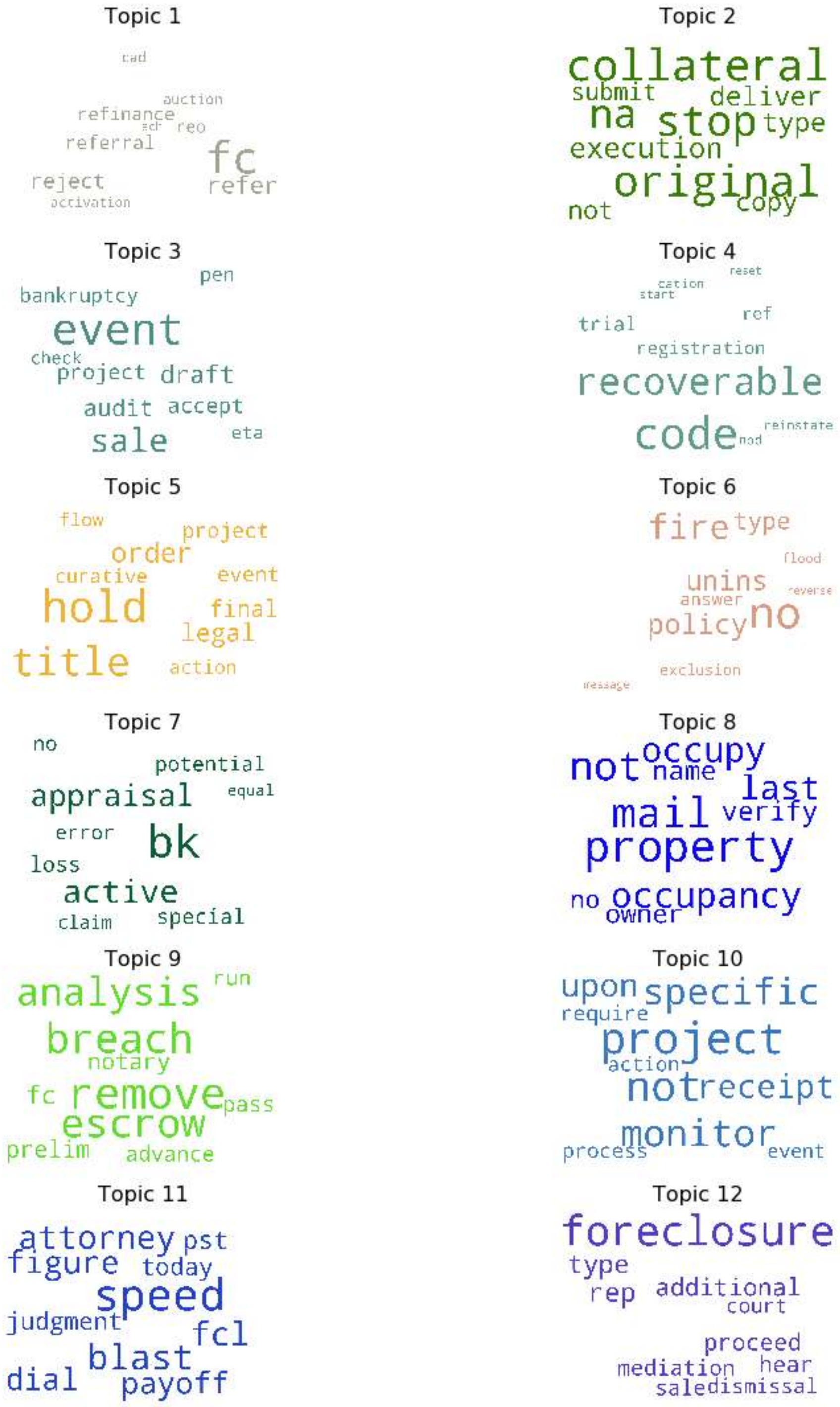


Figure 52: Key 12 Topics and related wordcloud.

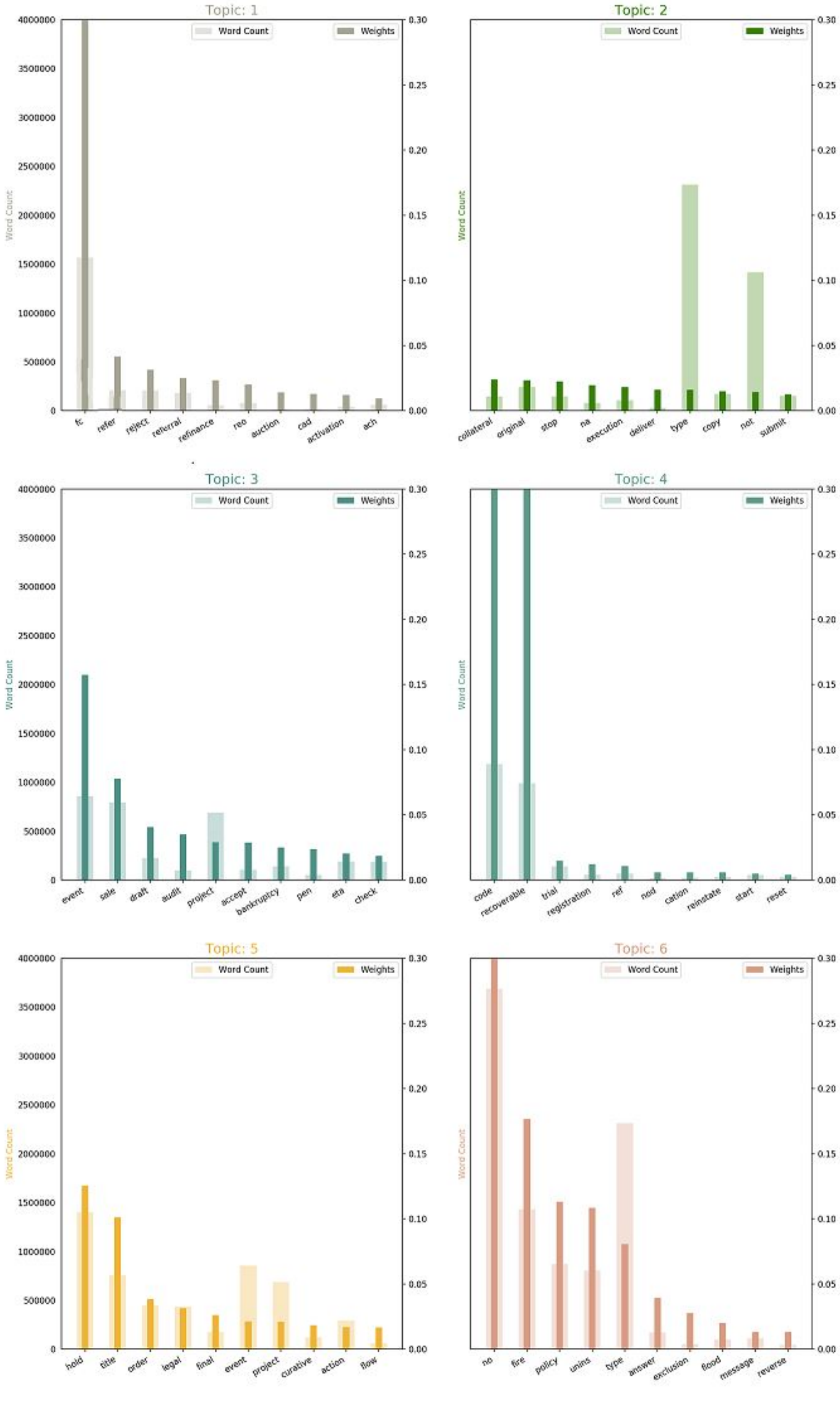


Figure 53: Relative weights and word counts of topics

In Figure 53 I plot the relative weights of words in each topic and also their frequencies for topics 1-6 and a similar analysis for topics 7-12 in Figure 54.

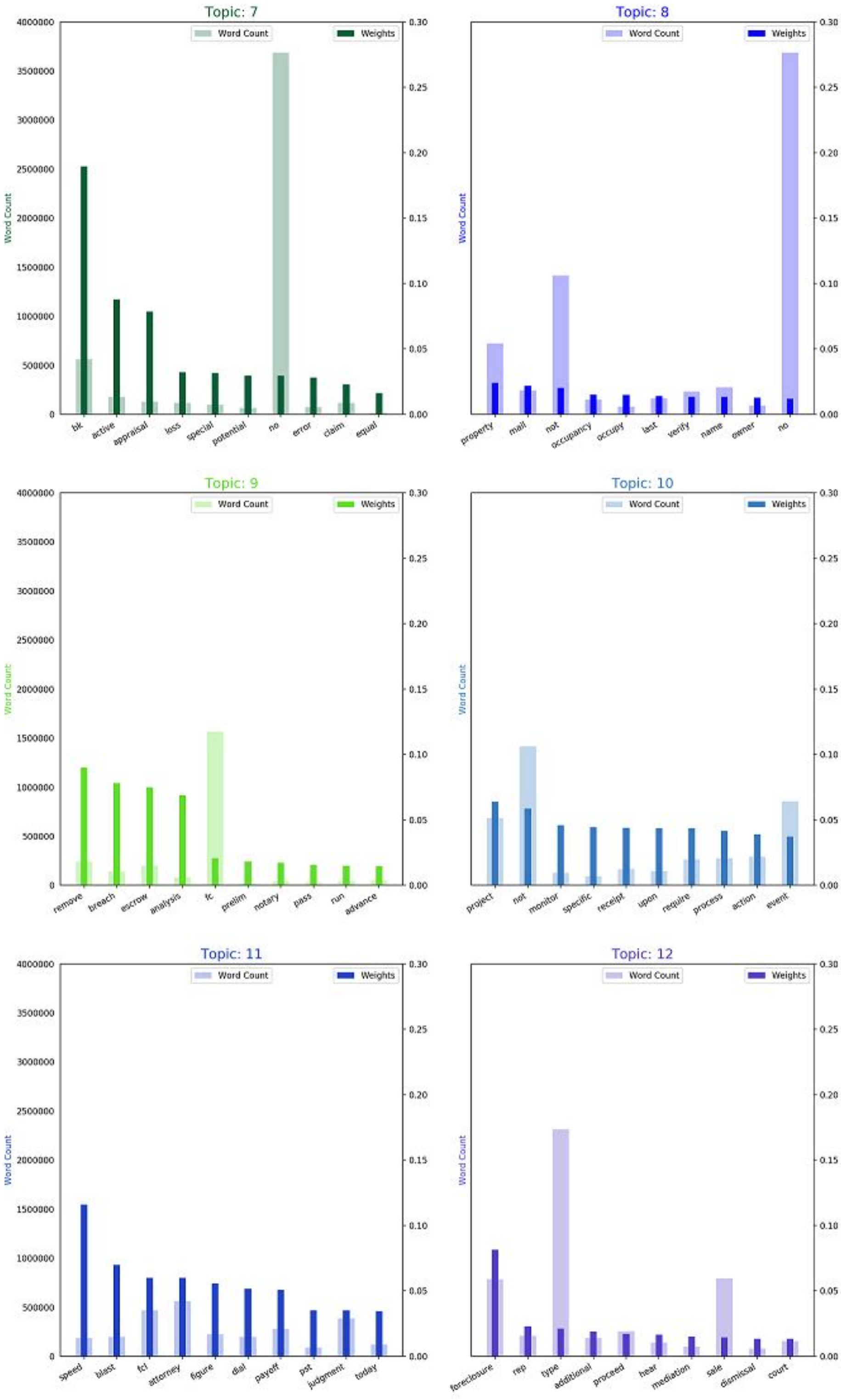
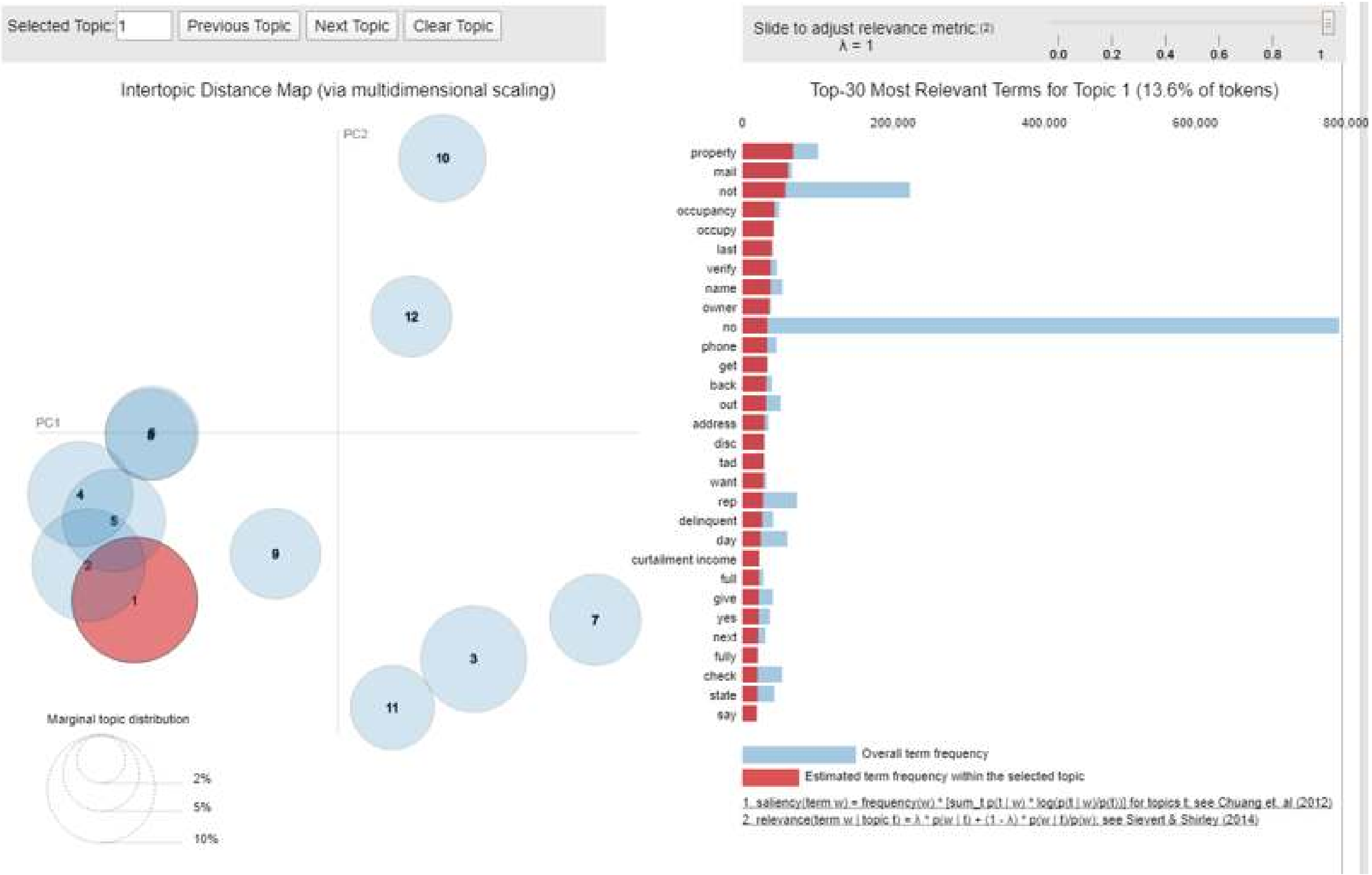


Figure 54: Relative weights and word counts of topics

In Figure 53 I plot the relative weights of words in each topic and also their frequencies for topics 1-6 and a similar analysis for topics 7-12 in Figure 54.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 55: Topic 1 and related word frequencies and weights.

Keynes, J. M. (1937). “The General Theory of Employment”. In: *The Quarterly Journal of Economics* 51.2, pp. 209–223.

Furstenberg, von and George (1969). “Default Risk on FHA-Insured Home Mortgages as a Function of the Terms of Financing: A Quantitative Analysis”. In: *Journal of Finance* 24.3, pp. 459–77.

Curley, Anthony J. and Jack M. Guttentag (1974). “The Yield on Insured Residential Mortgages”. In: *Explorations in Economic Research, Volume 1, Number 1*. National Bureau of Economic Research, Inc, pp. 114–161.

Jackson, Kenneth T. (1980). “Race, Ethnicity, and Real Estate Appraisal: The Home Owners Loan Corporation and the Federal Housing Administration”. In: *Journal of Urban History* 6.4, pp. 419–452. eprint: `https://doi.org/10.1177/009614428000600404`.

Campbell, Tim S and J Kimball Dietrich (1983). “The Determinants of Default on Insured Conventional Residential Mortgage Loans”. In: *Journal of Finance* 38.5, pp. 1569–81.

Foster, Chester and Robert Van Order (1984). “An Option-Based Model of Mortgage Default”. In: *Housing Finance Review*, pp. 351–372.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). “Learning Internal Representations by Error Propagation”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 318362. ISBN: 026268053X.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). “Multilayer feedforward networks are universal approximators”. In: *Neural Networks* 2.5, pp. 359 –366.

Cunningham, Donald and Charles Capone (1990). “The Relative Termination Experience of Adjustable to Fixed-Rate Mortgages”. In: *Journal of Finance* 45.5, pp. 1687–1703.

Hornik, Kurt (1991). “Approximation capabilities of multilayer feedforward networks”. In: *Neural Networks* 4.2, pp. 251 –257.

Riddiough, Timothy (1991). “Equilibrium mortgage default pricing with non-optimal borrower behavior”. In: *PhD Dissertation, Univseristy of Wisconsin*.

Sussmann, Hector J. (1992). “Uniqueness of the weights for minimal feedforward nets with a given input-output map”. In: *Neural Networks* 5.4, pp. 589 –593.

Albertini, Francesca and Eduardo D. Sontag (1993). “For neural networks, function determines form”. In: *Neural Networks* 6.7, pp. 975 –990.

Bansal and Viswanathan (1993). “No Arbitrage and Arbitrage Pricing: A New Approach”. In: *Journal of Finance* 48.4, pp. 1231–1262.

Lee, Tae Hwy, Halbert White, and Clive Granger (1993). “Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests”. In: *Journal of Econometrics* 56.3, pp. 269–290.

Hutchinson, James M, Andrew Lo, and Tomaso Poggio (1994). “A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks”. In: *Journal of Finance* 49.3, pp. 851–89.

Kuan, Chung-Ming and Halbert White (1994). “Adaptive Learning with Nonlinear Dynamics Driven by Dependent Processes”. In: *Econometrica* 62.5, pp. 1087–1114.

Goh, A. T. C. (1995). “Back-propagation neural networks for modeling complex systems”. In: *AI in Engineering* 9, pp. 143–151.

Granger, Clive (1995). “Modelling Nonlinear Relationships between Extended-Memory Variables”. In: *Econometrica* 63.2, pp. 265–79.

Childs, Paul, Steven H. Ott, and Timothy J. Riddiough (1996). “The Pricing of Multiclass Commercial Mortgage-Backed Securities”. In: *The Journal of Financial and Quantitative Analysis* 31.4, pp. 581–603.

Gedeon, Tamás D. (1997). “Data Mining of Inputs: Analysing Magnitude and Functional Measures”. In: *Int. J. Neural Syst.* 8.2, pp. 209–218.

Swanson, Norman and Halbert White (1997). “A Model Selection Approach To Real-Time Macroeconomic Forecasting Using Linear Models And Artificial Neural Networks”. In: *The Review of Economics and Statistics* 79.4, pp. 540–550.

Brown, Stephen, William Goetzmann, and Alok Kumar (1998). “The Dow Theory: William Peter Hamilton’s Track Record Reconsidered”. In: *Journal of Finance* 53.4, pp. 1311–1333.

Pace, Kelley, Ronald Barry, and C. F. Sirmans (1998). “Spatial Statistics and Real Estate”. In: *Journal of Real Estate Finance and Economics* 17.1, pp. 1573–045. eprint: `https://doi.org/10.1023/A:1007783811760`.

Chan, Sewin, Henry Schneider, and Joseph Tracy (1999). “Are stocks overtaking real estate in household portfolios?” In: *Current Issues in Economics and Finance* 5.Apr, p. 5.

Deng, Yong (1999). “Network Power: Japan and Asia. Edited by Katzenstein Peter and Shiraishi Takashi. Ithaca, NY: Cornell University Press, 1997. 399p. $55.00 cloth, $22.50 paper”. In: *American Political Science Review* 93.01, pp. 226–227.

Hofmann, Thomas (1999). “Probabilistic latent semantic indexing”. In: *Proceedings of SIGIR*, pp. 50–57.

Ambrose, BW and A Pennington-Cross (2000). “Local economic risk factors and the primary and secondary mortgage markets”. In: *Regional Science and Urban Economics* 30.6, 683–701.

Friedman, Jerome H. (2000). “Greedy Function Approximation: A Gradient Boosting Machine”. In: *Annals of Statistics* 29, pp. 1189–1232.

Ambrose, Brent W., Jr. Charles A. Capone, and Yongheng Deng (2001). “Optimal Put Exercise: An Empirical Examination of Conditions for Mortgage Foreclosure”. In: *The Journal of Real Estate Finance and Economics* 23, pp. 213–234.

King, Gary and Langche Zeng (2001). “Logistic Regression in Rare Events Data”. In: *Political Analysis* 9.02, pp. 137–163.

Ambrose, BW, A Pennington-Cross, and AM Yezer (2002). “Credit rationing in the US mortgage market: Evidence from variation in FHA market shares”. In: *Journal of Urban Economics* 51.2, 272–294.

Chawla, Nitesh V. et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *J. Artif. Intell. Res.* 16, pp. 321–357.

Ambrose, Brent W. and Anthony B. Sanders (2003). "Commercial Mortgage-Backed Securities: Prepayment and Default". In: *The Journal of Real Estate Finance and Economics* 26.2, pp. 179–196.

Bengio, Yoshua et al. (2003). "A Neural Probabilistic Language Model". In: *J. Mach. Learn. Res.* 3.null.

Faure-Grimaud, Antoine, Jean-Jacques Laffont, and David Martimort (2003). "Collusion, Delegation and Supervision with Soft Information". In: *The Review of Economic Studies* 70.2, pp. 253–279. eprint: `https://academic.oup.com/restud/article-pdf/70/2/253/4504589/70-2-253.pdf`.

Kuncheva Ludmila I.and Whitaker, Christopher J. (2003). "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy". In: *Machine Learning* 51.2, pp. 181–207.

Antweiler, Werner and Murray Frank (2004). "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards". In: *Journal of Finance* 59, pp. 1259–1294.

Manski, Charles F. (2004). "Measuring Expectations". In: *Econometrica* 72.5, pp. 1329–1376.

Olden, Julian D., Michael Kevin Joy, and Russell George Death (2004). "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data". In:

Titman, Sheridan, Stathis Tompaidis, and Sergey Tsyplakov (2005). "Determinants of Credit Spreads in Commercial Mortgages". In: *Real Estate Economics* 33.4, pp. 711–738.

Baker, Malcolm and Jeffrey Wurgler (2006). "Investor Sentiment and the CrossSection of Stock Returns". In: *Journal of Finance* 61.4, pp. 1645–1680.

Boukus, Ellyn and Joshua Rosenberg (2006). "The information content of FOMC minutes". In: *SSRN Electronic Journal*.

Mayer, C. and Yingjin Hila Gan (2006). “Agency Conflicts, Asset Substitution, and Securitization”. In: *Banking Financial Institutions*.

Das, Sanjiv and Mike Y. Chen (2007). “Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web”. In: *Management Science* 53.9, pp. 1375–1388.

Klabjan, Diego (2007). “Subadditive approaches in integer programming”. In: *European Journal of Operational Research* 183.2, pp. 525–545.

Ambrose, Brent W., Anthony B. Sanders, and Abdullah Yavas (2008). “CMBS Special Servicers and Adverse Selection in Commercial Mortgage Markets: Theory and Evidence”. In: *Real Estate Research Institute*.

Bajari, Patrick, Chenghuan Sean Chu, and Minjung Park (2008). *An Empirical Model of Subprime Mortgage Default From 2000 to 2007*. NBER Working Papers 14625. National Bureau of Economic Research, Inc.

Cashin, Sheryll (2008). “Race, Class, and Real Estate”. In: *Race, Poverty the Environment* 15.2, pp. 56–58.

Christopoulos, Andreas D., Robert A. Jarrow, and Yildiray Yildirim (2008). “Commercial Mortgage-Backed Securities (CMBS) and Market Efficiency with Respect to Costly Information”. In: *Real Estate Economics* 36.3, pp. 441–498. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6229.2008.00219.x`.

Daelemans, Walter, Bart Goethals, and Katharina Morik, eds. (2008). *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I*. Vol. 5211. Lecture Notes in Computer Science. Springer. ISBN: 978-3-540-87478-2.

Demers, Elizabeth and Clara Vega (2008). *Soft information in earnings announcements: news or noise?* Tech. rep.

Spalding, Ashley E. (2008). “Race, Class, and Real Estate: Neoliberal Policies in a Mixed Income Neighborhood”. In: *Graduate Theses and Dissertations*.

van der Maaten, L.J.P. and G.E. Hinton (2008). "Visualizing High-Dimensional Data Using t-SNE". In: *Journal of Machine Learning Research* 9, pp. 2579–2605.

Yildirim, Yildiray (2008). "Estimating Default Probabilities of CMBS Loans with Clustering and Heavy Censoring". In: *The Journal of Real Estate Finance and Economics* 37.2, pp. 93–111.

An, Xudong, Yongheng Deng, and Stuart A. Gabriel (2009). "Value Creation through Securitization: Evidence from the CMBS Market". In: *Journal of Real Estate Finance and Economics* 38.3, 302–326.

Duffie, Darell (2009). "Frailty Correlated Default". In: *Journal of Finance* LXIV.5, pp. 34–52.

Foote, Christopher et al. (2009). *Reducing Foreclosures: No Easy Answers*. NBER Working Papers 15063. National Bureau of Economic Research, Inc.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2009). *Moral and Social Constraints to Strategic Default on Mortgages*. Working Paper 15145. National Bureau of Economic Research.

Harding, John P., Eric Rosenblatt, and Vincent W. Yao (2009). "The contagion effect of foreclosed properties". In: *Journal of Urban Economics* 66.3, pp. 164 –178.

Lin, Zhenguo, Eric Rosenblatt, and Vincent W. Yao (2009). "Spillover Effects of Foreclosures on Neighborhood Property Values". In: *The Journal of Real Estate Finance and Economics* 38.4, pp. 387–407.

Loughran, Tim, Bill McDonald, and Hayong Yun (2009). "A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports". In: *Journal of Business Ethics* 89.1, pp. 39–49.

Whitelaw, Casey et al. (2009). "Using the Web for Language Independent Spellchecking and Autocorrection". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 890–899.

Elul, Ronel et al. (2010). "What "Triggers" Mortgage Default?" In: *American Economic Review* 100.2, pp. 490–94.

Geanakoplos, John (2010). "The Leverage Cycle". In: *NBER Macroeconomics Annual* 24, pp. 1–66. eprint: `https://doi.org/10.1086/648285`.

Khandani, Amir, Adlar Kim, and Andrew Lo (2010a). “Consumer Credit Risk Models Via Machine-Learning Algorithms”. In: *SSRN Electronic Journal*.

Khandani, Amir E., Adlar J. Kim, and Andrew Lo (2010b). “Consumer credit-risk models via machine-learning algorithms”. In: *Journal of Banking and Finance* 34.11, pp. 2767–2787.

Mikolov, Tomas et al. (2010). “Recurrent neural network based language model”. In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 1045–1048.

Piskorski, Tomasz, Amit Seru, and Vikrant Vig (2010). “Securitization and distressed loan renegotiation: Evidence from the subprime mortgage crisis”. In: *Journal of Financial Economics* 97.3, pp. 369–397.

Agarwal, Sumit et al. (2011a). “The role of securitization in mortgage renegotiation”. In: *Journal of Financial Economics* 102.3, pp. 559–578.

Agarwal, Sumit et al. (2011b). “The Role of Soft Information in a Dynamic Contract Setting: Evidence from the Home Equity Credit Market”. In: *Journal of Money, Credit and Banking* 43.4, pp. 633–655.

Demyanyk, Yuliya and Otto Van Hemert (2011). “Understanding the Subprime Mortgage Crisis”. In: *Review of Financial Studies* 24.6, pp. 1848–1880. (Visited on 07/27/2020).

Elul, Ronel (2011). *Securitization and mortgage default*. Tech. rep.

Agarwal, Sumit, Yan Chang, and Abdullah Yavas (2012). “Adverse selection in mortgage securitization”. In: *Journal of Financial Economics* 105.3, pp. 640 –660.

Ambrose, Brent W. and Richard J. Buttimer Jr. (2012). “The Adjustable Balance Mortgage: Reducing the Value of the Put”. In: *Real Estate Economics* 40.3, pp. 536–565.

Black, Bernard S., Woochan Kim, and Julia Nasev (2012). “The Effect of Board Structure on Firm Disclosure and Behavior: A Case Study of Korea and a Comparison of Research Designs”. In: *ECGI - Finance Working Paper*, p. 52. eprint: https://ssrn.com/abstract=2133283.

Favara, Giovanni, Enrique Schroth, and Philip Valta (2012). “Strategic Default and Equity Risk Across Countries”. In: *Journal of Finance* 67.6, pp. 2051–2095.

Levitin, Adam and Susan Wachter (2012). “Explaining the Housing Bubble”. In: p. 36.

Schafran, Alex and Jake Wegmann (2012). “Restructuring, Race, and Real Estate: Changing Home Values and the New California Metropolis, 1989-2010”. In: *Urban Geography* 33.5, pp. 630–654. eprint: `https://doi.org/10.2747/0272-3638.33.5.630`.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2013). “The Determinants of Attitudes toward Strategic Default on Mortgages”. In: *Journal of Finance* LXVIII.4.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). “Linguistic Regularities in Continuous Space Word Representations”. In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pp. 746–751.

Towe, Charles and Chad Lawley (2013). “The Contagion Effect of Neighboring Foreclosures”. In: *American Economic Journal: Economic Policy* 5.2, pp. 313–335.

Anenberg, Elliot and Edward Kung (2014). “Estimates of the Size and Source of Price Declines Due to Nearby Foreclosures”. In: *American Economic Review* 104.8, pp. 2527–51.

Ergungor, O. Emre and Stephanie Moulton (2014). “Beyond the Transaction: Banks and Mortgage Default of LowIncome Homebuyers”. In: *Journal of Money, Credit and Banking* 46.8, pp. 1721–1752.

Mayer, Christopher et al. (2014). “Mortgage Modification and Strategic Behavior: Evidence from a Legal Settlement with Countrywide”. In: *The American Economic Review* 104.9, pp. 2830–2857.

Montufar, Guido F et al. (2014). “On the Number of Linear Regions of Deep Neural Networks”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., pp. 2924–2932.

Murtagh, Fionn and Pierre Legendre (2014). “Wards Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Wards Criterion?” In: *Journal of Classification* 31.3, 274295.

Srivastava, Nitish et al. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15, pp. 1929–1958.

Agarwal, Sumit, Brent W. Ambrose, and Yildiray Yildirim (2015). "The Subprime Virus". In: *Real Estate Economics* 43.4, pp. 891–915. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.12108`.

An, Xudong et al. (2015). "Asymmetric Information and Subprime Mortgage Default". In: *SSRN Electronic Journal*.

Cordell, Larry et al. (2015). "The Cost of Foreclosure Delay". In: *Real Estate Economics* 43.4, 916–956.

D'Aurizio, Leandro, Tommaso Oliviero, and Livio Romano (2015). "Family firms, soft information and bank lending in a financial crisis". In: *Journal of Corporate Finance* 33, pp. 279 – 292.

Bayer, Patrick, Fernando Ferreira, and Stephen L Ross (2016). *What Drives Racial and Ethnic Differences in High Cost Mortgages? The Role of High Risk Lenders*. Working Paper 22004. National Bureau of Economic Research.

Butaru, Florentin et al. (2016). "Risk and risk management in the credit card industry". In: *Journal of Banking & Finance* 72.C, pp. 218–239.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *CoRR* abs/1603.02754. arXiv: `1603.02754`.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. New York: The MIT Press.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke (2016). "Deep Learning for Mortgage Risk". In: *SSRN Electronic Journal*.

Denton, Nancy (2017). "Race and Real Estate". In: *Contemporary Sociology* 46.5, pp. 542–544. eprint: `https://doi.org/10.1177/0094306117725085h`.

Fuster, Andreas et al. (2017). *Predictably Unequal? The Effects of Machine Learning on Credit Markets*. CEPR Discussion Papers 12448. C.E.P.R. Discussion Papers.

Heaton, J. B., N. G. Polson, and J. H. Witte (2017). “Deep learning for finance: deep portfolios”. In: *Applied Stochastic Models in Business and Industry* 33.1, pp. 3–12.

Agarwal, Sumit and Itzhak Ben-David (2018). “Loan prospecting and the loss of soft information”. In: *Journal of Financial Economics* 129.3, pp. 608–628.

Agarwal, Sumit et al. (2018). “The Politics of Foreclosures”. In: *Journal of Finance* 73.6, 2677–2717.

Buchak, Greg et al. (2018). “Fintech, regulatory arbitrage, and the rise of shadow banks”. In: *Journal of Real Estate Finance and Economics* 130.3, 453–483.

Chen, Hugh, Scott Lundberg, and Su-In Lee (2018). “Hybrid Gradient Boosting Trees and Neural Networks for Forecasting Operating Room Data”. In: *CoRR* abs/1801.07384. arXiv: `1801.07384`.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2018). *Bartik Instruments: What, When, Why, and How*. NBER Working Papers 24408. National Bureau of Economic Research, Inc.

Greenwell, Brandon M. et al. (2018). “Residuals and Diagnostics for Binary and Ordinal Regression Models: An Introduction to the sure Package”. In: *The R Journal* 10.1, pp. 381–394.

Heimer, Rawley and Alex Imas (2018). “Biased By Choice: How Financial Constraints Can Reduce Financial Mistakes”. In: *9th Miami Behavioral Finance Conference 2018*, pp. 1–83. eprint: `https://ssrn.com/abstract=3300456`.

Jiang, Cuiqing et al. (2018). “Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending”. In: *Annals of Operations Research* 266.1, pp. 511–529.

Kruger, Samuel (2018). “The effect of mortgage securitization on foreclosure and modification”. In: *Journal of Financial Economics* 129.3, 586–628.

Kvamme, Håvard et al. (2018). “Predicting mortgage default using convolutional neural networks”. In: *Expert Syst. Appl.* 102, pp. 207–217.

Soo, Cindy K (2018). “Quantifying Sentiment with News Media across Local Housing Markets”. In: *The Review of Financial Studies* 31.10, pp. 3689–3719. eprint: `https://academic.oup.com/rfs/article-pdf/31/10/3689/25693090/hhy036.pdf`.

Albanesi, Stefania and Domonkos Vamossy (2019). “Predicting Consumer Default: A Deep Learning Approach”. In: p. 70.

Ambrose, Brent et al. (2019). “Contractual Completeness in the CMBS Market: Insights from Machine Learning”. In: *Working Paper*.

Campbell, Dennis, Maria Loumioti, and Regina Wittenberg-Moerman (2019). “Making sense of soft information: interpretation bias and loan quality”. In: *Journal of Accounting and Economics* 68.2.

Conklin, James N. et al. (2019). “The Importance of Originator-Servicer Affiliation in Loan Renegotiation”. In: *Journal of Real Estate Finance and Economics* 59.1, 56–89.

Giglio, Stefano et al. (2019). *Five Facts about Beliefs and Portfolios*. NBER Working Papers 25744. National Bureau of Economic Research, Inc.

Heimer, Rawley and Alp Simsek (2019). “Should retail investors leverage be limited?” In: *Journal of Financial Economics* 132.3, pp. 1–21.

Akee, Randall K.Q. et al. (2020). *Dissecting the US Treasury Departments Round 1 Allocations of CARES Act COVID19 Relief Funding for Tribal Governments.* Working Paper. Harvard University, Cambridge, MA, 2020.

An, Xudong et al. (2020). “Inequality in the Time of COVID-19: Evidence from Mortgage Delinquency and Forbearance”. In:

Baker, Todd and Kathryn Judge (2020). *How to Help Small Businesses Survive COVID-19*. Working Paper 620. Columbia University School of Law.

Boar, Corina and Simon Mongey (2020). *Dynamic Trade-offs and Labor Supply Under the CARES Act*. Working Paper 27727. National Bureau of Economic Research.

Bybee, Leland et al. (2020). *The Structure of Economic News*. NBER Working Papers 26648. National Bureau of Economic Research, Inc.

Capponi, Agostino, RUIZHE JIA, and David Aaron Rios (2020). *Foreclosures and House Price Growth in the COVID-19 Period.*

Carroll, Christopher D. et al. (2020). *Modeling the consumption response to the CARES Act.* Working Paper Series 2441. European Central Bank.

Chetty, Raj et al. (2020). *How Did COVID-19 and Stabilization Policies Affect Spending and Employment? A New Real-Time Economic Tracker Based on Private Sector Data.* Working Paper 27431. National Bureau of Economic Research.

DLima, Walter, Luis A. Lopez, and Archana Pradhan (2020). "COVID-19 and Housing Market Effects: Evidence from U.S. Shutdown Orders". In: *Working paper.*

Ganong, Peter and Pascal Noel (2020). "Why Do Borrowers Default on Mortgages?" In: p. 43.

Humphries, John Eric, Christopher Neilson, and Gabriel Ulyssea (2020). *The evolving impacts of COVID-19 on small businesses since the CARES Act.* Cowles Foundation Discussion Papers 2230. Cowles Foundation for Research in Economics, Yale University.

Indarte, Sasha (2020). "Moral Hazard versus Liquidity in Household Bankruptcy". In: *Working Paper*, p. 77. eprint: `https://ssrn.com/abstract=3378669`.

Neilson, Christopher, John Eric Humphries, and Gabriel Ulyssea (2020). *Information Frictions and Access to the Paycheck Protection Program.* Working Paper 27624. National Bureau of Economic Research.

Petrosky-Nadeau, Nicolas (2020). *Reservation Benefits: Assessing job acceptance impacts of increased UI payments.* Working Paper Series 2020-28. Federal Reserve Bank of San Francisco.

Wilson, Fernando and Jim P Stimpson (2020). *US Policies Increase Vulnerability of Immigrant Communities to the COVID-19 Pandemic.*

McManus, Douglas and Elias Yannopoulos (2021). "COVID-19 Mortgage Forbearances: Drivers and Payment Behavior". In: *The Journal of Structured Finance.* eprint: `https://jsf.pm-research.com/content/early/2021/03/11/jsf.2021.1.120.full.pdf`.

VITA

Arka Prava Bandyopadhyay
PhD candidate, William Newman Real Estate Department, Baruch College, CUNY
Principal Data Scientist/Executive Director, Rocktop Partners, TX

Education
PhD candidate, William Newman Real Estate Department, Baruch College, CUNY
PhD (All-But-Dissertation) in Applied Mathematics, MS in Mathematical Finance
Courant Institute of Mathematical Sciences, NYU
MS, Computer Science, Louisiana State University
B.Math, Indian Statistical Institute

Professional
Principal Data Scientist, Rocktop Partners
Senior Director, Investments, Nuveen
Quantitative Risk, Director, UBS
Mortgage Prepayment, Default Risk, VP, Santander Bank
Senior Consultant, Deloitte & Touche
Quant Strat, Amber Capital

Selected working papers (since 2019)

- Deep Learning for disentangling Liquidity-constrained and Strategic Default (With Yildiray Yildirim), Conference Presentation: **ASSA 2020, FMA 2020, BlackRock FMG, Bloomberg, Baruch Annual Symposium, ARES 2020 & 2021, Baruch Real Estate Seminar**.

- Cost of Misaligned CARES Act: Overcrowding, Selective Verification and Unintended Racial Consequences **(under review)**, Conference Presentation: **IRMC 2020, AREUEA Virtual Seminar, Bloomberg, ARES 2021 Doctoral Session, CUNY Economics Seminar**.

- Communications Between Borrowers and Servicers: Evidence from the Covid-19 Mortgage Forbearance Program **Quarterly Journal of Finance, 2021**, Conference Presentation: **AREUEA PhD Poster, IRMC 2020, ARES 2021**.

- Quantifying Soft Information, Mortgage Market Efficiency and Asset Pricing Implications, Conference Presentation: **(under review) ASSA 2021, CUNY Economics Seminar, Baruch Brown Bag**.

- Market Microstructure of Luxury Properties: a data-driven approach, **(under review)** Conference Presentation: **Baruch Real Estate Seminar, CUNY Economics Seminar**.

- Regional Distributional Implications of COVID-19 on Housing Price Index and Investment Opportunities, Presentation: **Baruch Finance Seminar, CUNY Economics Seminar**.

Examples of industry-sponsored research projects

- Rocktop Partners, Nuveen, UBS, Santander Bank, Deloitte: Available on request.