AEA Papers and Proceedings 2019, 109: 1–5
https://doi.org/10.1257/pandp.20191000

BIG DATA IN THE MODERN ECONOMY

The Impact of Big Data on Firm Performance: An Empirical Investigation†

By Patrick Bajari, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki*

The use of data as an input into the production function of a firm dates back at least to the emergence of the modern industrial firm in the nineteenth century (Chandler 1977). Dramatic improvements in data collection, storage, and analysis technologies have made it easier for firms to use data in their decision processes. A question that has arisen in academic and policy discussions is the manner in which the performance of data-enabled decision systems scales with the size of the datasets; in particular, whether increasing returns or a "data feedback loop" exists (e.g., Newman 2014; Grunes and Stucke 2015; Lerner 2014; and Lambrecht and Tucker 2017).

In this paper, we conduct an empirical study of Amazon's retail forecasting system, an important input into purchase ordering decisions. In the context of forecasting, one might interpret the "data feedback loop" hypothesis as with more data, firms can produce better forecasts, which in turn allows them to better serve customers, which in turn leads to more data.

We consider four hypotheses regarding how dataset size may influence the accuracy of forecasts. First, statistical learning theory typically suggests diminishing returns to dataset size in terms of estimation error or predictive performance (e.g., Lerner 2014). In our particular setting, the underlying data is a panel, and the theory suggests that forecast accuracy should improve as we increase N, the number of products within a product category. Second, we would expect forecast errors to decrease as T, the number of observations per product, increases. Third, as pointed out by Bresnahan, Brynjolfsson, and Hitt (2002); Bloom, Sadun, and Van Reenen (2012); and Lambrecht and Tucker (2017), complementary investments and organizational practices often play a very important role in generating productivity gains from the use of data. In our particular setting, learning by doing applies to this case. Finally, there can be "diseconomies of scale" in forecasts. As the amount of data grows, the team may need to apply the same model to an increasing number of products with different demand patterns.

We test the empirical relevance of these factors by estimating forecast errors as a function of N, T and other factors using panel data for many product groups. Our dependent variable is a measure of forecast accuracy for a given product.

Our results suggest that (i) there is improvement in forecast performance in the T dimension, (ii) N has relatively little impact on forecasting performance, except when the number of products is relatively small, (iii) there is a trend rate of improvement in the forecasts, consistent with "learning by doing."

With the important caveat that our empirical results are specific to the context studied, we

* Bajari: University of Washington, 311 Savery Hall, UW Economics Box 353330, Seattle, WA 98195, and NBER (email: Bajari@uw.edu); Chernozhukov: Department of Economics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (email: vchern@mit.edu); Hortaçsu: Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, and NBER (email: hortacsu@uchicago.edu); Suzuki: Amazon Inc., 345 Boren Avenue North, Seattle, WA 98109 (email: sjunich@amazon.com). Bajari is VP and Chief Economist at Amazon, Inc. Chernozhukov and Hortaçsu carried out research work for this paper as independent contractors for Amazon, Inc. We thank Zhiying Gu and Liquan Huang with assistance in data collection and analysis, George Monokroussos for providing feedback, and seminar participants at INFORMS, Bruegel, The World Trade Organization, UCLA, NBER Summer Institute, and ASSA annual meeting for insightful comments.

† Go to https://doi.org/10.1257/pandp.20191000 to visit the article page for additional materials and author disclosure statement(s).

002_P20191000.indd 1 3/20/19 9:10 AM

note that our results are inconsistent with a naive "data feedback loop," where the addition of new products always results in better models. We interpret our results as instead being consistent with the statistical theory for the size of forecasting errors.

I. Data and Empirical Model

Our analysis builds on a proprietary panel dataset on weekly historical forecasts and actual unit sales at Amazon ranging from December 8, 2012 to June 3, 2017 for 36 major product lines including apparel, books, and consumer electronics.

We observe Q_{it}, the quantity of product i sold during week $t = 1, \ldots, H$ as well as the corresponding one-week ahead mean forecast $\hat{Q}_{it}$ that was produced using data available at one week ago, that is, at $t-1$. We also use the following variables:

- $T_{i,t}$ is the "Age" of product at the company, which describes the length of data available for product i at time t, as consumed by the forecasting model. The own history data, especially if available for substantive ranges, can be used to build a high quality forecast.
- $N_{i,t}$ is the number of products in the same product line as product i at time $t-1$, whose data were consumed by the forecasting model. It is reasonable to expect that data on products in the same product line can be used to capture seasonality and fashion patterns, improving the forecast based upon a product's own history alone.

We would like to assess how the size of the available data $(T_{i,t}, N_{i,t})$ affects the relative forecast error. Specifically we define the following binary dependent variable

$$Y_{i,t} = \mathbf{1}\left\{\left|Q_{i,t} - \hat{Q}_{it}\right|/(Q_{i,t}+1) > X\right\}.$$

This variable describes the event of making a *big forecast error*, namely $X\%$ error. The threshold of X is chosen for each product line separately so that this threshold is exceeded for any random product-week pair (t,i) with probability $P = 30\%$.

Our primary empirical model, motivated by the theoretical discussion in Bajari et al. (2018)[1], is the following linear prediction equation for $Y_{i,t}$:

$$\begin{aligned}(2)\quad Y_{i,t} = {} & \alpha_i + \beta_t + \delta_1 \mathbf{1}(T_{i,t} > 20) \\ & + \delta_2 \mathbf{1}(T_{i,t} > 20)/\sqrt{T_{i,t}} \\ & + \delta_3 \mathbf{1}(T_{i,t} > 20)/T_{i,t} \\ & + \gamma_1 \mathbf{1}(N_{i,t} > 200) \\ & + \gamma_2 \mathbf{1}(N_{i,t} > 200)/\sqrt{N_{i,t}} \\ & + \gamma_3 \mathbf{1}(N_{i,t} > 200)/N_{i,t} + \epsilon_{i,t},\end{aligned}$$

where the terms α_i are the product specific fixed effects, which capture the notion that some products are inherently harder to forecast than others, even using the same data size. We believe it is reasonable to always include product fixed effects as the leading empirical specification. Indeed, the product fixed effects control for changing product mix and for the fact that the products are added in a nonrandom manner. The terms β_t are the time effects, which capture that on some dates it may be harder or easier to forecast than on other dates; moreover, these time effects also partly capture the overall improvement in the forecasting model that produces $\hat{Q}_{i,t}$. We consider three specifications for time effects:

(i) No Trend: $\beta_t = 0$ for all $t = 1, \ldots, H$.

(ii) Smooth Trend: $\beta_t = b_1(t/H) + b_2(t/H)^2$.

(iii) Time Fixed Effects: β_ts are unrestricted parameters.

We expect that the coefficients will exhibit the following signs although we will not restrict the signs in the estimation:

$$\delta_1 \leq 0, \quad \delta_2 \geq 0, \quad \text{and} \quad \delta_3 \leq 0;$$

$$\gamma_1 \leq 0, \quad \gamma_2 \geq 0, \quad \text{and} \quad \gamma_3 \leq 0.$$

[1] A short summary of this theoretical discussion is available in the online Appendix.

002_P20191000.indd 2 3/20/19 9:10 AM

We call the three terms with δ_j "Product Age Effects," or the "T Effect." These effects are meant to capture the nonlinear effect of the age of product i on the relative forecast error. The first two terms are motivated by the theoretical bounds on the relative forecasting error derived in the theoretical stylized model, whereas the third term is meant to moderate the effect of the first-order term. Similarly, the next three terms with γ_j represent "Number of Products Effect," or the "N Effect." These effects are meant to capture the nonlinear effect of the number of products $N_{i,t}$ in a similar broad product line as i, on the relative forecast error. The threshold of $T_{i,t} > 20$ in the definition above is meant to signify a minimum sample size for time series dimension. Likewise, $N_{i,t} > 200$ was chosen similarly, based upon our practical experience that substantial cross-sectional size is needed to start to learn factors (e.g., seasonality) properly. Third, our model is a predictive model. It describes how changes in the data sizes $(N_{i,t}, T_{i,t})$ affect the predicted probability of the large forecast errors. We don't necessarily ascribe a causal interpretation to these results.

II. Results

Figures 1 and 2 graphically present the estimated Product Age (T) Effect and the Number of Products (N) Effect on the predicted probability of a big forecast error event on electronic products[2] with the three different specifications for the time effects. Figure 3 presents the estimated date/time effects in the model with smooth trend and time fixed effects. The point estimates and standard errors of these regressions are available in the online appendices. The signs on the coefficients for the terms containing age T are consistent with the model, and the signs for the N follow less systematic patterns.

Result 1: The T effect (Figure 1) is the strongest in the model without trends, reaching -0.35 for high T. The T effect is more modest in models that allow for smooth trends or time fixed effects, flattening at -0.06 in the time fixed effects models. There are essentially no improvements due to large T in the two models

[2] See Bajari et al. (2018), the working paper version of this paper, for results on other product lines.

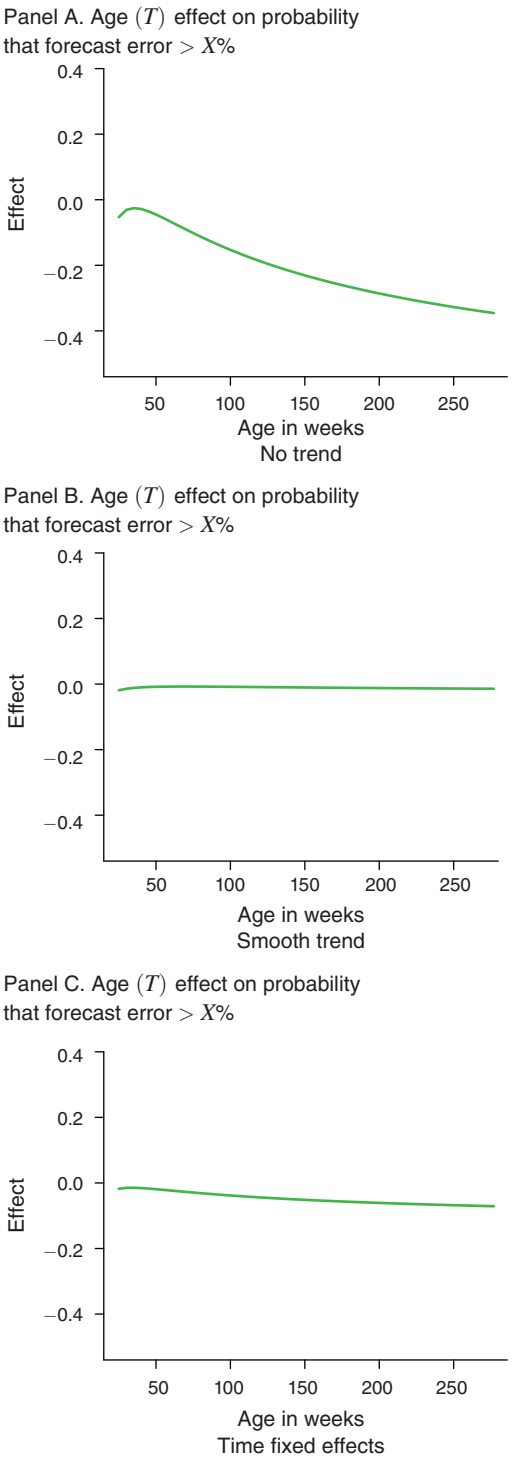

FIGURE 1. THE ESTIMATED IMPACT OF T ON THE QUALITY OF LEARNING IN THE MODEL WITH NO TREND, SMOOTH TREND, AND TIME FIXED EFFECTS

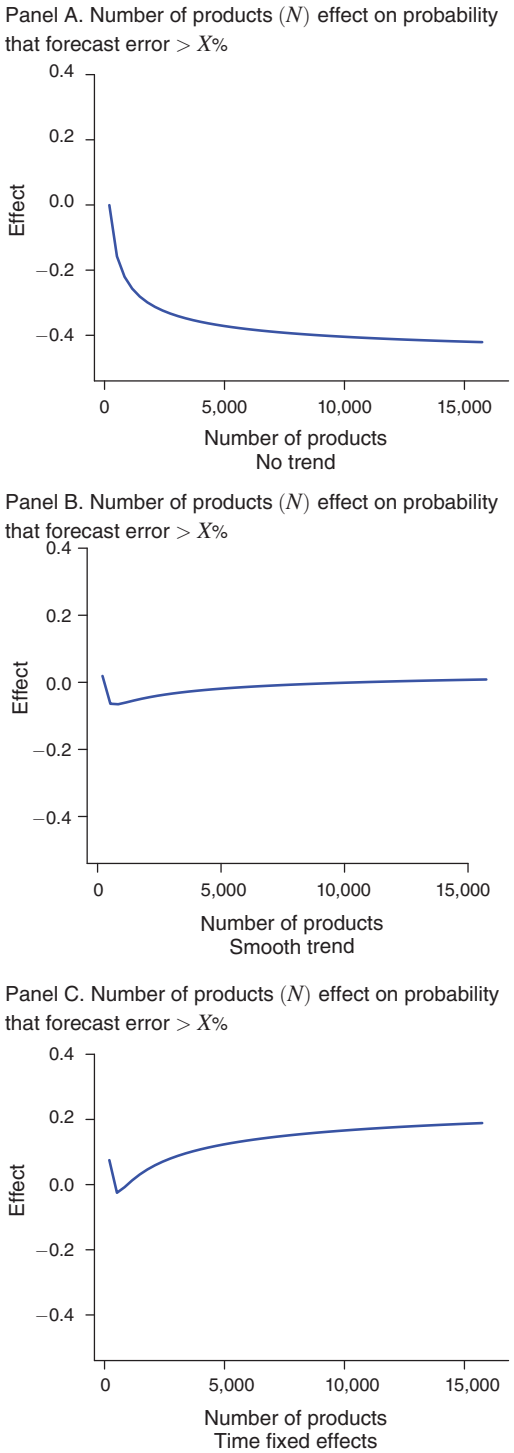

FIGURE 2. THE ESTIMATED IMPACT OF *N* ON THE QUALITY OF LEARNING IN THE MODEL WITH NO TREND, SMOOTH TREND, AND TIME FIXED EFFECTS

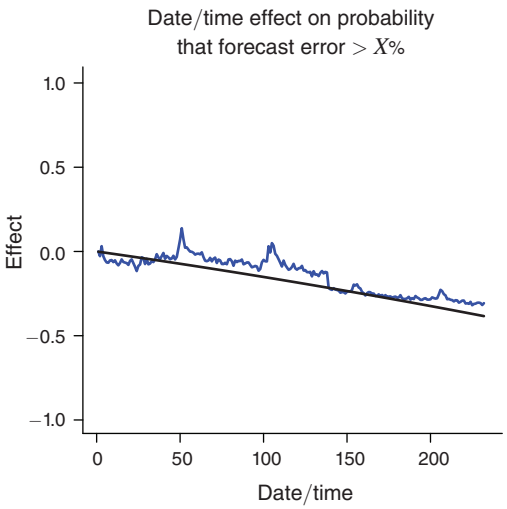

FIGURE 3. THE ESTIMATED DATE/TIME EFFECTS ON THE QUALITY OF LEARNING IN THE MODEL WITH SMOOTH TREND AND TIME FIXED EFFECTS

that adjust for time fixed effects. Overall, the *T* effect seems to agree with predictions of the theoretical model, and there are diminishing returns (improvement in forecast error) to large *T*.

Result 2: Without time adjustments, the *N* effect (Figure 2) seems to be strong early on, but it does exhibit diminishing returns to scale and saturates at -0.4 once $N > 5{,}000$. Note, however, that this effect is marginally significant (insignificant at the five percent level but significant at the ten percent level, without any multiplicity adjustments). If we inspect the model with the smooth trends, we see that the *N* effect is almost zero at high *N*. This effect is also not statistically significant. In the third model with unrestricted time effects, large *N* seems to moderately increase the probability of the big error event. This contradicts the predictions of the data feedback loop hypothesis. This modestly negative effect of increasing *N* does appear to be significant, at least given the conventional time-product clustered standard errors we use. In summary, the *N* effect seems to exhibit either positive, quickly diminishing returns to scale or modestly negative returns, depending on how we control for time effects.

Result 3: The estimated time effects (Figure 3) indicate that there is a general improvement in the relative forecast error over time; this could reflect improvements in the forecasting engine itself, for example using more features (traffic data etc.), as well as the time evolution of the mix of products.

To interpret empirical Result 1 in light of Result 3, note that the ages $T_{i,t}$ and trends t are correlated in this data. Hence the model without trends attributes the aggregate changes in the occurrence of the Big Error Event to the longer T effect, whereas the model with the trends or time fixed effects projects out the aggregate changes or trends, before it attributes the effect to longer T. The aggregate changes suggest gradual improvements in forecasting performance, which occur due to

(i) learning-by-doing and other methodological improvements (e.g., using more features), and

(ii) availability of longer histories/ages (T) for many products.

Hence the model without trends attributes the aggregate changes to source (ii). The Time Fixed Effects model attributes the aggregate changes captured by time fixed effects to source (i) and the remainder of the effects to source (ii). The Smooth Trend model attributes aggregate changes captured by a smooth trend to source (i) and the rest of the changes to source (ii).

III. Discussion

There are important limitations of our study. As we noted above, the T effect that we identify may not be the true causal effect of having access to longer histories. This could be driven also by the continual improvements in forecasting technology over time. What we can recover, however, is bounds on the true effect of T. Another important caveat we should emphasize is regarding the scope of this study. Clearly, data is used in many aspects of company decision-making, and our focus is on the application to demand forecasting. We believe that this is a particularly good area to study, as the success of forecasting models is relatively straightforward to assess. However, our conclusions regarding the presence of scale benefits to data are limited to this particular application.

REFERENCES

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki. 2018. "The Impact of Big Data on Firm Performance: An Empirical Investigation." NBER Working Paper 24334.

Bloom, Nicholas, Rafaella Sadun, and John Van Reenen. 2012. "Americans Do IT Better: US Multinationals and the Productivity Miracle." *American Economic Review* 102 (1): 167–201.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. 2002. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence." Quarterly Journal of Economics 117 (1): 339–76.

Chandler, Alfred D., Jr. 1977. *The Visible Hand: The Managerial Revolution in American Business.* Cambridge, MA: Harvard University Press.

Grunes, Allen P., and Maurice E. Stucke. 2015. "No Mistake about It: The Important Role of Antitrust in the Era of Big Data." University of Tennessee Legal Studies Research Paper 269.

Lambrecht, Anja, and Catherine E. Tucker. 2017. "Can Big Data Protect a Firm from Competition." *Competition Policy International Antitrust Chronicle*. January 17, 1–8.

Lerner, Andres V. 2014. "The Role of 'Big Data' in Online Platform Competition." Unpublished.

Newman, Nathan. 2014. "Search, Antitrust, and the Economics of the Control of User Data." *Yale Journal on Regulation* 31 (2): 401–54.