Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction:
An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach

Praveen Ranjan Srivastava, Indian Institute of Management, Rohtak, India

Prajwal Eachempati, Indian Institute of Management, Rohtak, India

ABSTRACT

The paper aims to examine the factors that influence employee attrition rate using an employee records dataset. It also aims to establish the predictive power of deep learning for employee churn prediction over ensemble machine learning techniques like random forest and gradient boosting on real-time employee data from a mid-sized fast-moving consumer goods (FMCG) company. The results are further validated through a regression model and a multi-criteria fuzzy analytical hierarchy process (AHP) model that considers the relative variable importance and computes weights. The machine learning models' empirical results indicate that deep neural networks (91.6% accuracy) is a better predictor of churn than random forest and gradient boosting algorithm (82.5% and 85.4%, respectively). These findings provide useful insights for human resource (HR) managers in an organizational workplace context. When recalibrated by organizations' human resource team, the model helps in better incentivization and more targeted employee retention.

INTRODUCTION

The world is a land of opportunities, and talent and skills are abundant among people from different walks of life. This talent is found in all sections and strata of the society irrespective of age and demographical variations. Recruiting such talent in organizations is important, but it is equally essential to satisfy and retain them since employees have their considerations for choosing and working in an organization, and if they are not satisfied, they can resign. This will lead to employee attrition and, thus, a phenomenon is known as employee churn. The reasons for this churn (Cheng et al.,2019) are varied, and there is a need to identify and target prospective employees who are more likely to resign. Therefore, there is a need to forecast churn (also called 'Turnover_status') to retain organizations' talent. All people today have their own needs and wants according to the Maslow Hierarchy of Needs

(Maslow,1943), and they make real-time decisions to stay in an organization or not based on many factors. According to Hertzberg Two Factor (Motivation and Hygiene) Theory (Herzberg,2005), they need incentives and motivation to work in the organization since they drive the employee's passion towards his work. Additionally, every employee feels a sense of ownership about his work, and only if his contribution to the organization is valued and rewarded will he be satisfied according to the Theory of Organizational Equilibrium (Simon and March,1976).

Thus, every employee needs to be motivated and perceive his job as meaningful according to the Job Embeddedness Theory (Mitchell *et al.,*2001); otherwise, he may resign from the organization, causing job churn.

To alleviate this problem, there is a need to predict accurately churn with the key factors that influence this decision using an employee's customized data available online. This is accomplished using state-of-the-art machine learning techniques that consider churn reasons as input in their model and are trained based on past employee data to recognize patterns and identify prospective employees who are likely to quit the organization.

There are existing studies that attempt to predict employee churn (Sisodia *et al.,* 2017; Ogbonnaya *et al.*, 2017; Ma *et al.*, 2019; Keegan and Hartog, 2019) using the above machine learning techniques but a multi-factor hybrid approach for validating the reasons for employee churn was not adopted. Further, the models are not tuned to provide recommendations for better employee retention and talent management.

This study's main objective is to build a robust hybrid employee churn prediction model for the organizations. For this purpose, the factors determining employee churn were inspired by an employee record dataset from the website Kaggle (In Class Prediction Competition,2017), a data analytics platform that contributes datasets and performs research in data science.

The proposed model aims to identify the critical factors that govern employee churn and provide customized recommendations to HR managers on how to retain valuable and disgruntled employees.

This study's data is collected by interviewing Information Technology sector employees in India and this data is used for training the prediction model. The model is then validated on a real-time employee dataset of a mid-sized Fast-Moving Consumer Goods (FMCG) company for comparing the predictive power of the classifiers.

The paper is structured as follows: The literature review of the existing studies in churn prediction and the rationale for adopting the above predictors is detailed. The data collection and research methodology are then elucidated. The results and discussion are illustrated. The study is then concluded and the scope for future research is discussed. The references follow this.

LITERATURE REVIEW

The existing studies conducted in the domain of organizational behavior (Trauth, Quesenberry & Huang, 2008**;** Edwards and Edwards, 2012; Fuchs and Edwards,2012) analyze the factors considered by employees to continue their job at the current organization the following factors governing employee churn emerge as significant predictors of employee churn outcome variable 'Turnover_status' as depicted in Figure 1.

The factors that emerge from the existing literature and the rationale for considering them as the study's critical factors as per defined variables in Figure 1 is provided below:

- **Employee_Satisfaction:** The extent to which the employee is content with his allocated responsibility in the project is the most critical factor for employee retention since only if the employee is satisfied can he deliver the best performance (Brown *et al.*,2008; Ma *et al.*,2019; Al-Ansari and Alshare, 2019). The variable is given out of 2.

Figure 1. The emergence of factors from literature
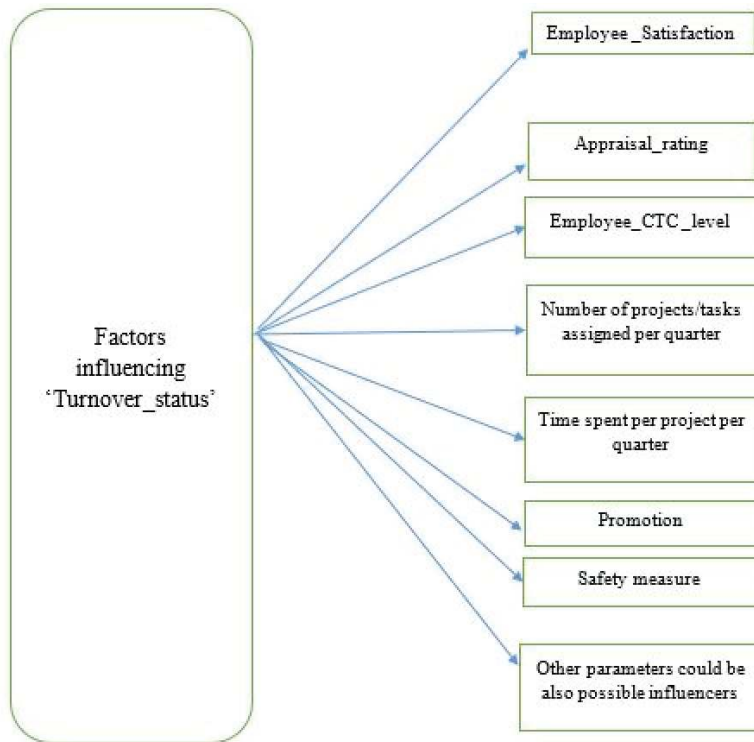
- **Appraisal_rating:** The metric that measures how well the employee performs a task within the time and budget constraints in the form of Appraisal_rating is also an essential factor for the organization provided out of 2 (Mohr *et al.,*2012; Keegan and Hartog, 2019).
- **Employee_CTC_level:** The level of annual remuneration of the employee (measured in Cost to Company i.e., CTC) and whether he is promoted or not commensurate to his performance evaluation score also in-directly impacts 'Employee_Satisfaction' and, in turn, 'Turnover_status' (Ajit,2016; Ogbonnaya *et al*.,2017; Ma *et al.,*2019; Raman *et al.,*2019; Mohr *et al.,*2012; El-Rayes, Fang, Smith, & Taylor,2020). The 'Employee_CTC_level' is classified into low, medium, and high, which is recoded in the dataset for further analysis as 0, 1, and 2.
- **Number of projects/tasks assigned per quarter:** The number of projects assigned to the employee under a specified timeline such as a quarter (3 months period) also affects the psychological thought process of employees (Keegan and Hartog, 2019).
- **Time spent per project per quarter:** The number of hours that an employee spends to complete a particular project successfully in a quarter is 'Time spent per project per quarter' (Ma *et al.,*2019).
- **Safety measure:** This is a flag status variable to indicate whether the work environment is safe and accident-free (indicated by 0 if accident-free else denoted by 1 in the dataset).
- **Promotion:** This is a flag status variable to indicate whether the employee has been promoted (indicated by 1 if promoted else denoted by 0 in the dataset).

The outcome variable for predicting employee churn is the 'Turnover_status' flag variable which is 1 if employee quits the organization and 0 if the employee stays in the organization.

Table 1. Existing Studies in Churn Prediction

S.No	Cited Study	Research Objective	Methodology /Technique	Limitations	Implications
1.	Al-Ansari & Alshare, 2019	This paper studies how employee churn is impacted by employee satisfaction in Qatar.	The Transaction-Based Model of Stress and Coping Theory was developed using a web-based survey questionnaire.	Structural Equation Modelling approach is adopted.	'Employee_ Satisfaction' has a positive effect on organizational commitment.
2.	El-Rayes, Fang, Smith, & Taylor (2020)	This paper examines the impact of 'Employee_ Satisfaction' on employee churn.	Regression techniques were adopted to investigate the impact of 'Employee_ Satisfaction'.	Machine learning techniques were not adopted for churn prediction and feature importance validation.	'Employee_ Satisfaction' plays a vital role in employee churn.
3.	Ogbonnaya et al., (2017)	The influence of 'Employee_CTC_ level' on a positive work ethic is investigated.	Structural Equation Modelling is used for investigating the impact of contingent pay.	The significance of the variable was not validated through hybrid machine learning and multi-factor approaches.	'Employee_CTC_ level' is a significant predictor.
4.	Mohr et al., (2012)	'Turnover_status' as a moderator variable for organizational culture is investigated.	A hierarchical regression analysis technique is adopted.	Though 'Turnover_status' is studied as a moderator, the causes that, in turn, influence 'Turnover_ status' were not investigated, and their importance is not validated.	'Employee_ Satisfaction' and 'Appraisal_rating' are critical factors.
5.	Keegan and Hartog(2019)	This paper examines the influence of the performance 'Appraisal_rating' for churn prediction.	The epistemological qualitative analysis approach is adopted through interviews to understand employee performance appraisal impact on 'Turnover_ status'.	This qualitative approach is not validated with data analysis for checking the consistency of interview responses with variable significance.	The 'Appraisal_rating' and the 'Number of projects/tasks assigned per quarter' is a critical factor.
6.	Ajit (2016)	To identify factors that influence employee churn in organizations.	Gradient Boosting models are applied for 'Turnover_status' prediction.	The model is deductive but not prescriptive and is not compared with a more robust machine learning approach like deep neural networks.	'Employee_CTC_ level' and 'Promotion' are significant factors to be considered.
7.	Sisodia et al.,(2017)	Employee churn is predicted using machine learning.	Machine learning baseline models like Support Vector, Naïve Bayes, and Decision Tree have been adopted for churn prediction.	Ensemble machine learning approaches were not explored, and the feature importance is not validated through multi-criteria models.	The results reveal that Random Forest is the best predictor.
8.	Ma et al., (2019)	The present study predicts the causes of employee 'Turnover_ status' occurrences in real-time.	Decision trees are applied for prediction.	The feature importance was not validated using a hybrid regression and multi-criteria approach.	'Time spent per project per quarter', 'Employee_ Satisfaction' and 'Promotion' are found to be critical.
9.	Raman et al., (2019)	The objective of this paper is to study the factors that drive employee retention and reduce attrition.	Predictive analytics has been adopted.	Complex machine learning algorithms are not adopted for providing a more accurate prediction, and variable importance is not validated.	The 'Appraisal_rating' and 'Employee_CTC_ level' are significant factors that drive retention.
10.	Hunter et al., (2008)	This investigation examines the factors influencing voluntary turnover decisions of software professionals in Singapore and New Zealand.	A narrative inquiry interview approach was conducted with the professionals to identify reasons for turnover.	Churn is forecasted, but the root cause for employee churn is not identified by examining the variable importance.	Factors influencing employee turnover can be both culturally independent and culturally sensitive.

Figure 2. Limitations in the existing studies of churn prediction
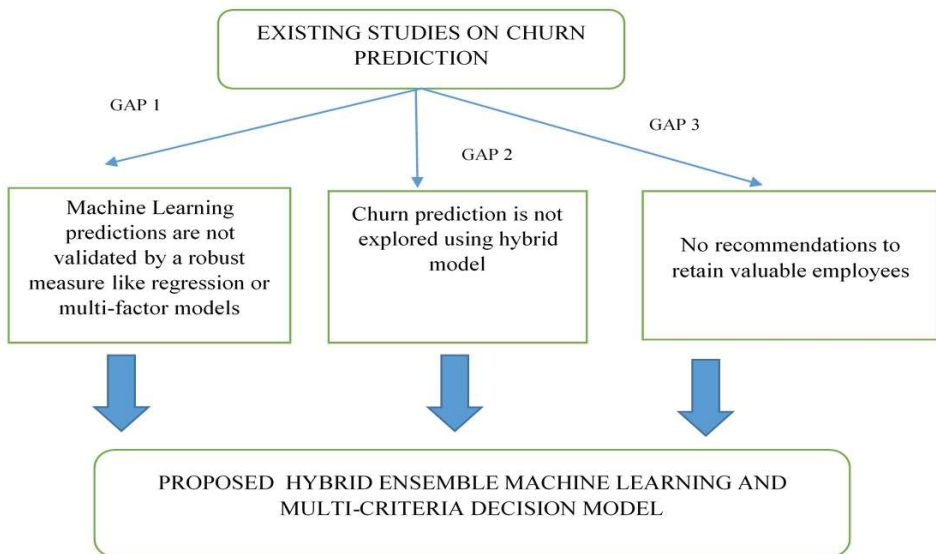
Existing Studies in Churn Prediction

The current work in the domain of employee churn prediction is tabulated in Table 1.

The existing studies have attempted to predict the employee churn in multiple contexts. majority of the studies have adopted a qualitative treatment to the employee churn problem. However, with the advent of state-of-the-art machine learning techniques for employee churn prediction, there is a need to investigate how these techniques can be leveraged to their full capability to not only predict employee churn but also to identify to what extent, each factor determines employee churn. The limitations of the studies and the development of the hypotheses in the paper is thus illustrated and elucidated in Figure 2.

## Limitations of Existing Studies and Hypothesis Development

As illustrated in Figure 2, the gaps in existing research are as follows.

Firstly, the machine learning approaches stated above are used for prediction. However, the feature importance that emerges is not validated by a robust statistical model like multiple linear regression or by a multi-stakeholder multi-criteria model that considers multiple factors that interplay in employees' minds before making a decision. The employees' personalized preferences and the importance they attach to different factors are not thus validated, and thus, the results are not dynamic and not applicable in a real scenario.

Secondly, the existing studies predict churn but do not compute or consider the relative importance of the factors influencing employee attrition rate, providing incentive strategies when prioritized. No hybrid model is adopted to predict churn and analyze the critical factors responsible for 'Turnover_status'.

Thirdly, none of the existing studies stated above have taken into consideration the essential factors. Understanding the key factors influencing employee chum, customized recommendations for employee retention, and boosting 'Employee_Satisfaction' can be provided. For instance, if the

Figure 3. Hypotheses development
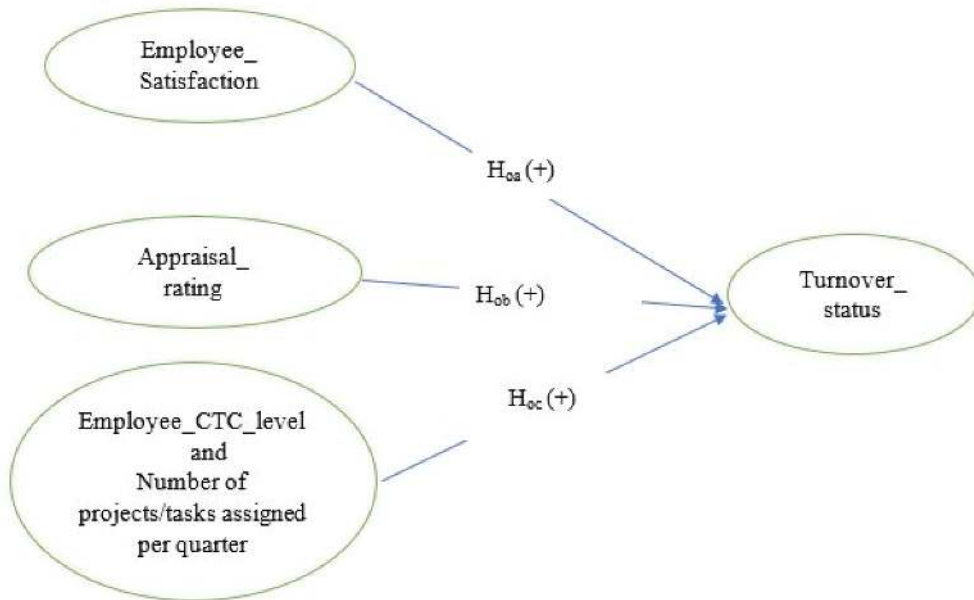
employee is predicted to leave the organization due to low 'Employee_CTC_level' and high workload (in terms of the 'Number of projects/tasks assigned per quarter'), incentives can be provided to the employee, and his workload can be distributed more amicably, thus increasing 'Employee_Satisfaction'.

Thus, a hybrid machine learning and multi-criteria model is proposed in this paper, predicting and analyzing the critical factors determining employee churn.

The implications from existing studies (Raman *et al.,*2019; El-Rayes, Fang, Smith, and Taylor,2020) indicate that the primary factors that influence employee churn are: 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level', and workload ('Number of projects/tasks assigned per quarter').

This forms a premise for testing the following hypotheses in the paper, as illustrated in Figure 3. The hypotheses are defined as follows:

H_{0a}: The factor 'Employee_Satisfaction' influences 'Turnover_status'.
H_{0b}: The factor 'Appraisal_rating' impacts 'Turnover_status'.
H_{0c}: 'Employee_CTC_level' and the 'Number of projects/tasks assigned per quarter' are significant predictors of 'Turnover_status'.

The data collection procedure and research methodology adopted to prove this hypothesis's validity are discussed below.

RESEARCH METHODOLOGY

Data Collection and Organizational Context

For collecting data for the study, the following procedure was adopted.

Table 2. Summary statistics of the input/predictor variables

Variable	Min	Max	Mean	Stdev
Employee_Satisfaction	0.38	1.95	0.99	0.34
Appraisal_rating	0.85	1.98	1.40	0.24
Employee_CTC_level	0	2	0.36	0.5
Number of projects/tasks assigned per quarter	1	3	2	0.6
Time spent per project per quarter	100	300	180	60
Safety measure	0	1	0.09	0
Promotion	0	1	0.0212	0.001

Initially, the factors considered for determining employee churn (variables and attributes) were inspired by an employee record dataset from the website Kaggle data analytics platform (In Class Prediction Competition,2017) that contributes datasets and performs research in data science.

The factors that propelled employee churn (as learned from the Kaggle dataset) were identified to be 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level' (level of Cost to Company i.e., CTC per annum) and the workload assigned to each employee quantified as the 'Number of projects/tasks assigned per quarter','Time spent per project per quarter' on completing the project deadlines. The factors were found relevant to the IT industry employees working in India. Data about the factors and the weightage assigned to the factors was collected by interviewing 365 respondents in the age group of 25-35.

The questionnaire designed for the interview is enclosed in the Appendix.

Since all the variables are in different ranges and are used to predict the outcome variable 'Turnover_status' which is a binary status variable with only two possible values (1 for employee leaving and 0 for employee staying in the company), the responses provided for the 'Employee_ Satisfaction' and 'Appraisal_rating' of the employee are normalized to a score between 0-2 in both the initial training set (IT) and validation set (FMCG). The reason for choosing the FMCG dataset as the validation set is due to the sector's vibrancy and scope for innovation and growth for the employees. There is also ample scope for employee churn in such a sector if the workplace is not challenging. If reasons could be identified for employees quitting for this sector and the significant factors can be identified, the problem of employee churn can be solved, and the insights from this can be applied to all other industries also. Secondly, a different sector dataset (FMCG) was chosen to minimize overfitting in test accuracy due to using the same sector (IT) for both train and test set. This overfitting is caused by the model "memorizing" the patterns in the original IT employee train dataset and not "learning" the patterns, which is the primary purpose of machine learning. Employee churn is a general workplace problem (not sector-specific), and the model should be generalizable in different contexts. This will help in recalibrating the model for different sector contexts. The rationale for allocating a range of 0-2 for relative measures like 'Employee_Satisfaction' and 'Appraisal_rating' is that the satisfaction and appraisal scores can be categorized based on merit of the employee into either low (0), medium(1) or high (2). However, the variables are discretized to contain values even in the range of 0-2. For instance, an employee with satisfaction score of 0.34 is found to be in the range of 0-1 and is categorized as 'low'. This normalization before feeding the inputs into the machine learning algorithm helps in more accurate prediction. 'Employee_CTC_level' is also recoded to 0,1 and 2 (indicating low, medium, and high) in the FMCG dataset as similarly performed for the IT employee training set.

This dataset contains the following attributes of employees in an organization: 'Employee_ Satisfaction', 'Appraisal_rating', 'Number of projects/tasks assigned per quarter', 'Time spent per

Table 3. A snippet of the employee dataset

Employee_ID	Employee_ Satisfaction	Appraisal_ rating	Employee_CTC_ level	Number of projects/ tasks assigned per quarter	Time spent per project per quarter	Safety measure	Promotion
1	0.87	1.12	0	2	103	0	0
2	1.5	1.76	1	2	115	0	0
3	0.72	1.76	1	3	258	0	0
4	1.44	1.74	0	2	239	0	0
5	0.74	1.04	0	2	105	0	1
6	0.82	1	0	2	228	1	1
7	0.6	1.54	0	2	183	1	1
8	1.84	1.7	0	2	235	1	1
9	1.78	1.84	0	3	182	1	1
10	0.87	1.12	0	2	103	0	0

project per quarter', 'Promotion', 'Employee_CTC_level' (whether low, medium or high), 'Safety measure' of the workplace and the 'Turnover_status' of the employee (whether he has left the organization or not). The employee (whether he is leaving the organization or not) is 'Turnover_status' to be predicted based on the above collected variable data.

The data distribution of the dataset is analyzed through summary statistics of the predictor variables shown in Table 2.

The summary statistics of the input/predictor variables are provided in terms of minimum, maximum, mean and standard deviation. For instance, the 'Employee_Satisfaction' variable has the minimum score of 0.38, the maximum score of 1.95, mean of 0.99 and standard deviation of 0.34. Further, the distribution indicates that the variables 'Employee_Satisfaction', 'Appraisal_rating' and 'Employee_CTC_level' are normalized to a score out of 2 with other variables represented as number. The 'Time spent per project per quarter' is quantified in terms of number of hours per project spent for each quarter ranging from minimum value of 100 hours per project per quarter to maximum to 300 hours per project per quarter with mean of 180 hours per project per quarter and standard deviation of 60 hours per project per quarter. The 'Safety measure' variable is recoded in terms of 0/1 (indicating absence or presence of work safety). Similarly, 'Promotion' variable is also recoded as 0/1 to indicate whether the employee is promoted(indicated by 1) or not promoted (0).

A snippet of the adapted employee dataset is illustrated in Table 3.

Consider the scenario of Employee 2 (2nd row in Table 3). The 'Employee_Satisfaction' score is 1.5 out of 2 (high score), the current 'Appraisal_rating' is 1.76 out of 2(high score), his 'Employee_CTC_level' is medium level(recoded in the dataset as 1), the 'Number of projects/tasks assigned per quarter' is 2 (he is assigned 2 projects per quarter i.e., 3 months), the 'Time taken per project per quarter' is 115 hours per quarter for each project, his workplace is not safe and has no 'Safety measure' (hence recoded as 0 in the dataset), he was not given 'Promotion' in the last five years. Given the above parameters, a churn prediction model is to be implemented for predicting whether the employee would continue to work or would leave the organization. The methodology adopted is elucidated below.

Figure 4. Research Methodology adopted in the study
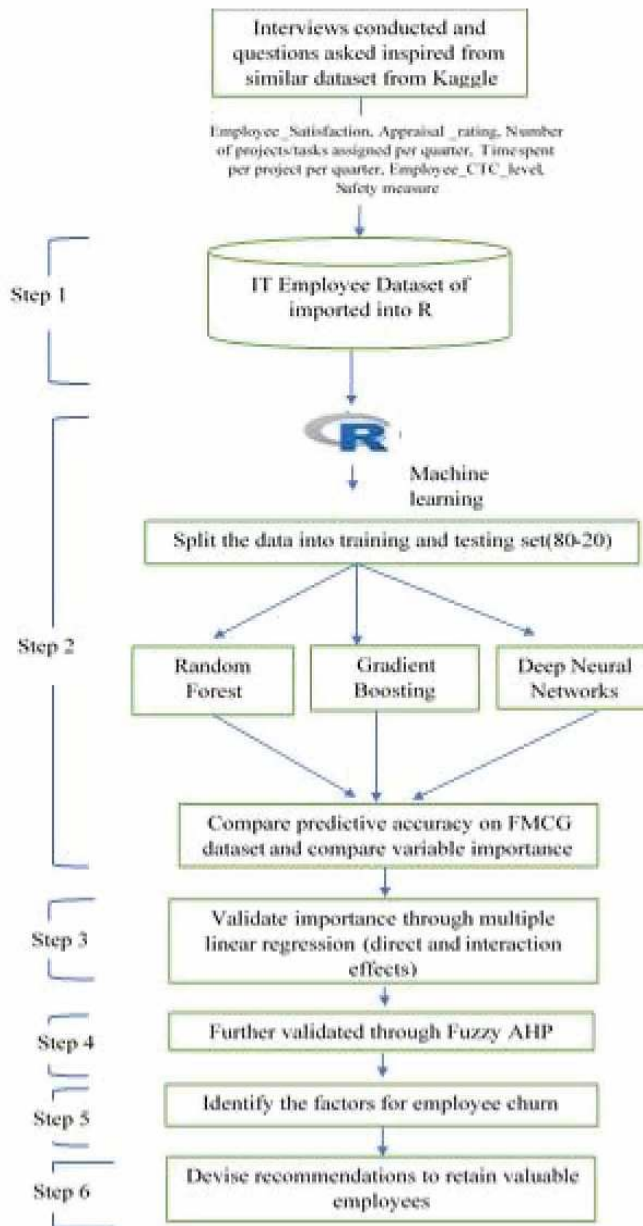
Methodology Adopted for the Study

This paper attempts to examine the role of various factors that influence employee churn. The factor variables are selected based on their significance in the organizational behavior domain and collected, as explained above.

This objective is accomplished using three steps as illustrated in Figure 4.

The first step is to perform predictive modeling using machine learning on the dataset, considering 'Turnover_status' as the outcome (to be forecasted) and the factors mentioned above, as illustrated in Figure 1. The models are then evaluated in terms of predictive accuracy and terms of the importance assigned to each variable, also known as relative importance.

The second step is to validate the relative importance of variables generated from the predictive models through a multiple regression model that computes how each predictor variable is significant and generated by these predictive models.

The third step is to corroborate the relative importance of variable generated from the predictive models through a multi-criteria Fuzzy Analytical Hierarchy Process (AHP) that takes the input pairwise comparison matrix criteria and computes their respective weights. The weights computed are validated from a stakeholder perspective and understand employees' and domain experts' real-time preferences in an uncertain environment; the Fuzzy Analytical Hierarchy Process (AHP) is adopted.

Thus, the overall research methodology is illustrated in Figure 4.

According to Figure 4, the following step-wise computation is elucidated:

Step 1: Import the dataset of IT employees into the R tool.

Step 2: Split the dataset into 80% training and 20% testing for implementing machine learning.

Step 2.1:   Build machine learning classifiers, namely, a Random Forest, Gradient Boost Model, and a deep neural network-based model with the same input variables for prediction through a training data set.

Step 2.2:   The prediction accuracies and performance on a real-life employee dataset of a mid-sized Fast-moving Consumer Goods company (FMCG) from a tier-2 city containing 5% of the number of data points in the training set 50 datapoints are then compared to determine which model outperforms the others in terms of churn prediction.

The data is collected by interviewing 50 respondents in the age group of 25-35 and contains the same attributes, i.e., 'Employee_Satisfaction', 'Appraisal_rating', 'Number of projects/tasks assigned per quarter', 'Time spent per project per quarter', the status of 'Promotion' from last five years, 'Employee_CTC_level' (whether low, medium or high), 'Safety measure' of the workplace and the 'Turnover_status' of the employee (whether he has left the organization or not) described above for the original dataset. While in the original dataset, data points based on employees from the IT sector are used to train the model, the validation set is chosen on an unseen FMCG employee dataset. The relative importance of the variables is, therefore, also compared. The attributes collected are generic to all sectors, and hence the responses gathered by interviewing the FMCG employees are synchronized to the response data in the training set of IT employees. Attributes like the 'Number of projects/tasks assigned per quarter', 'Appraisal_rating' and 'Employee_CTC_level' (salary) are equally crucial to both the IT and FMCG sectors. The criteria used to arrive at the 'Appraisal_rating' and allocate 'Employee_CTC_level' are generic and applicable to both the sectors. The questionnaire framed for collecting data from the FMCG employees for the validation interview is enclosed below in the Appendix.

Step 3: For validating whether the same variables are significant, a multiple linear regression model is built as a 'Turnover_status' prediction model using the above eight predictors. Further, since the employees consider multiple factors simultaneously before deciding whether to leave an organization, interaction variables are also factored in the regression model. This is done to investigate which factors simultaneously considered motivate employees to either leave an organization or stay.

Step 4: To validate whether the same variables are significant, a Fuzzy AHP methodology for computing variable weights is mentioned.

Step 5: The critical weights derived are analyzed to finalize the critical factors that employees consider before leaving an organization.

Figure 5. Working of the machine learning predictive models
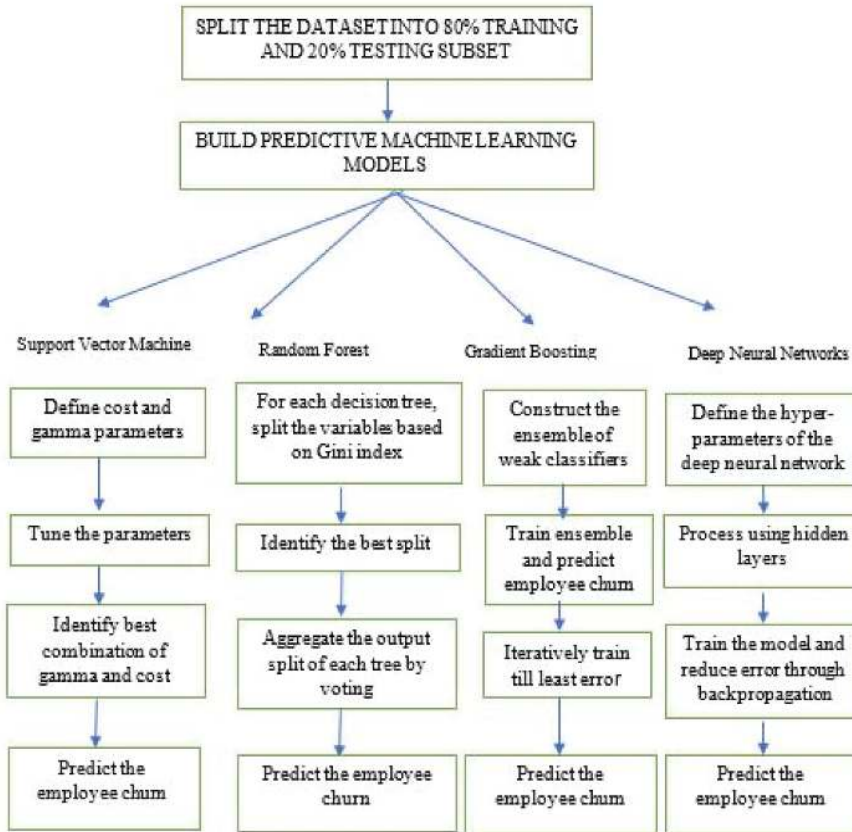
Step 6: The highly talented employees who are likely to leave the organization are then targeted for retention.

Building the Predictive Models

Predictive modeling techniques can be broadly categorized into supervised and unsupervised techniques (Sisodia *et.al.,*2017). In this paper, the predictive techniques, i.e., Random Forest model, Gradient Boosting, and more complex machine learning-based models like deep learning, have been adopted to derive the relative significance of factors and improve the extent to which prediction can be made in terms of accuracy.

Considering existing studies (Mohr *et al.,*2012; Keegan and Hartog, 2019) that build churn prediction models and in order to provide a consistent output that can handle multiple features with high predictive accuracy, machine learning techniques are adopted in the paper.

The model building phase starts by simulating highly sophisticated and layered machine learning models, such as random forest, gradient boost, and artificial neural network, and then validates the feature's importance through multiple linear regression models and further by Fuzzy AHP.

The reasons for this multi-stage approach (validation of the machine learning techniques through multiple linear regression and Fuzzy AHP) are as follows:

1.   The model building phase aims to predict the rating by multiple layered deep neural networks.

Table 4. Fuzzy linguistic terms with defined triangular scales (Wind and Saaty, 1980)

Saaty Scale	Definition	Fuzzy Triangular Scale
1	Equally important	(1,1,1)
3	Weakly Important	(2,3,4)
5	Fairly important	(4,5,6)
7	Strongly important	(6,7,8)
9	Absolutely important	(9,9,9)

2. Though machine learning techniques compute the variable importance, there is a need to confirm these variables' significance to provide more personalized and reliable recommendations to stakeholders. In this context, a valid mechanism to confirm the significance is the multiple regression techniques. Further, to understand the variables' relative importance from the stakeholder's perspective, Fuzzy AHP was also simulated to compute the criteria weights.

Neural networks are generally better at processing and providing a generic output (Zhou et al., 2002; Sisodia *et al.,*2017). Neural networks, particularly deep neural networks (Loureiro *et al*., 2018; Gleue *et al*.,2018), process the data through multiple hidden layers and boost predictive accuracy. The open-source data analytics tool R (Kaya *et al.,* 2019) was adopted to build a deep learning model, and a multiple linear regression model is built. Random forest results could be analyzed easily (Ogbonnaya et al.,2017). Gradient Boost is an ensemble model that combines the predictive power of simple models and is a robust predictor.

The open-source data analytics tool R (Angrave *et al*.,2016; Frisk and Bannister,2017; Huang *et al.,*2017; Kaya *et al*.,2019; Singh and Giudice,2019) was adopted to build the above three machine learning models and for simulating the Fuzzy AHP (Singh and Prasher,2019) process while Excel tool was adopted for the regression model. The machine learning models work on the algorithm as illustrated in Figure 5.

Figure 5 illustrates the working procedure of machine learning predictive models.

*   **Random Forest based predictive model (Bendazzoli *et.al.*,2019):** A random forest is a supervised machine learning classifier that combines the output of several Decision trees using a voting algorithm and predicts the outcome resulting from the aggregated outcomes. They are easy to implement with cleaner output and fit on a large set of data. For this model, all predictors are converted to numeric, and the output variable, i.e., 'Turnover_status', is predicted.
*   **Gradient Boosting predictive model (Ajit,2016):** The gradient boost ensemble model is also run to predict 'Turnover_status' to boost the predictive accuracy and interpretability. The boosting technique can be modeled as an optimization problem where the objective is to minimize the error rate of the ensemble model gradually and iteratively using a gradient descent like procedure. A weak algorithm like a decision tree is combined with a robust predictive model to form an ensemble that boosts the predictive accuracy. They help to deal with an unbalanced and large set of data to provide accurate results. For this model, all predictors are converted to numeric, and the output variable, i.e., 'Turnover_status', is predicted.
*   **Deep neural network (DNN) based model (Wang *et al.,*2018; Raman *et al.,*2019):** A Deep Neural Network (DNN) model is simulated based on the working of the human brain. A typical architecture is layer-wise: the input layer takes the normalized input data, and the last layer provides the output. In between, the processing of inputs is performed in hidden layers (one or more), which process the input values and compute an activation function (preferably sigmoid)

based on the importance of variables. A DNN is trained to learn the input weights and consequently generate the output incrementally. DNNs effectively handle a large number of multiple input variables, for example, in big data scenarios.

In this paper, the DNN has been constructed with five hidden layers (as illustrated below in Figure 8) due to the minimum root mean square error (RMSE) of 0.25, which also minimizes the probability of over-fit of the deep neural network model, i.e., the model fitting only on some data points and under-performing on other data points in the dataset(Loureiro et al.,2018). The number of input nodes is 7, considering seven individual predictors.

The variable importance of the predictors and the model's performance measured in terms of accuracy are compared and illustrated in the Results section.

- **Multiple linear regression-based predictive model(Kong *et al.,*2019):** Regression analysis is a statistical technique that predicts the outcomes from one or more factors and evaluates the explanatory power of these factors here, between 'Turnover_status' and the above-defined 7 predictors, to validate the importance of the variables given by the machine learning models.

Fuzzy-Analytical Hierarchy Process

Fuzzy Analytic Hierarchy Process [F-AHP] is applicable when multiple factors are uncertain to quantify. Fuzzy Analytic Hierarchy Process [F-AHP] (Singh and Prasher,2019) integrates the fuzzy concept to Analytic Hierarchy Process [AHP] (Wind and Saaty,1980). AHP makes the pair-wise comparison of the factors using crisp notation. In Fuzzy-AHP, the pair-wise comparisons of both criteria and the alternatives are carried out through the fuzzy triangular scale devised by Wind and Saaty(1980).

Table 4 depicts the Saaty scale scores from 1 to 9 and the corresponding fuzzy equivalent in the form of: (lo,m,u); lo being the lower bound of the fuzzy scale and u being the upper bound, and the median value is captured by the variable m.

The importance of a factor compared to another factor is assigned. For instance, a Saaty scale value of (1,1,1) indicates that factor i is highly un-important, while (6,7,8) indicates strong significance. The alternatives and criteria are compared based on this scale for drawing conclusions regarding which factors are more influential.

As illustrated above in Figure 4, the methodology undertaken for computing criteria weights is as follows:

Step 1: Data pertaining to the relative importance of factors that impact employee churn is collected through online surveys and questionnaires.
Step 2: The responses collected in the form of scores from 1-9 are converted to fuzzy notations of the form: (l,m,h) mentioned above for assigning the criteria.
Step 3: The fuzzy weights for the factors are allocated and converted to a pair-wise comparison matrix to ascertain the priority.
Step 4: The formula calculates the geometric mean of fuzzy values for each factor:

$$r_i = \left(\prod d_{ij}^{1/n} \right) \tag{1}$$

where; d represents the pair-wise relative importance of one criterion over the other for each of the alternatives, and i is from 1 to n.

Table 5. Predictive Accuracy of the Random Forest Model

mtry	ntree	RMSE	R-squared
2	100	290.9729	0.825
4	100	293.6134	0.824
6	100	295.4065	0.8232

Step 5: The fuzzy weights for each factor are computed by multiplying r_i with the inverse vector:

$$w_i = r_i * \left(r_1 * r_2 * r_3 * \ldots * r_n \right)^{-1} \tag{2}$$

$$= (lw_i, mw_i, hw_i)$$

where r_1, r_2 represent the geometric mean calculated above for each criterion.

From the fuzzy weights, first, the crisp average factor criterion M_i is computed by taking the average of the three fuzzy weights, i.e.:

$$M_i = (lw_i + mw_i + hw_i) / 3 \tag{3}$$

The average factor criterion M_i is normalized to normalized weight criterion N_i by the formula:

Figure 6. Variation of RMSE v/s number of predictors tuned by number of trees
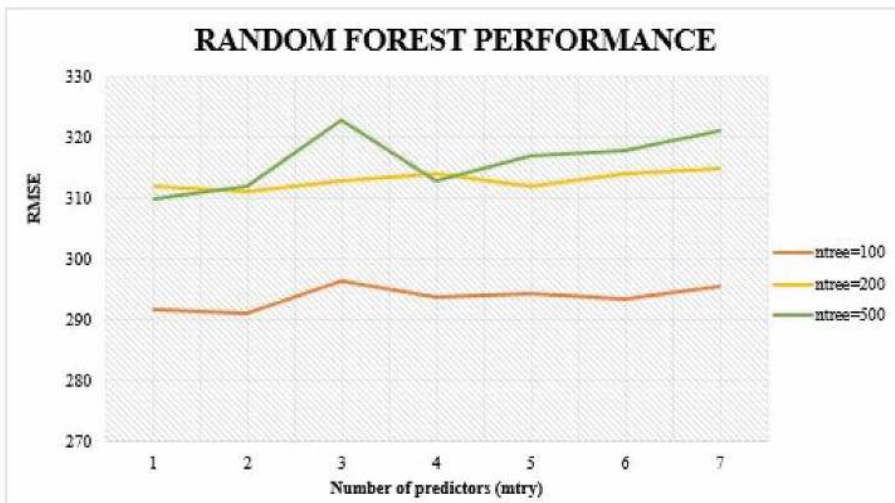
Table 6. Relative Importance of the variables in the Random Forest Model

Variable	Importance
Employee_Satisfaction	72.6
Appraisal_rating	63.4
Employee_CTC_level	54.2
Number of projects/tasks assigned per quarter	41.2
Time spent per project per quarter	30
Safety measure	13
Promotion	20

$$N_i = M_i / \sum M_i \tag{4}$$

where; $\sum M_i$ denotes the total of all the average crisp weights for all criteria.

These computations are pre-defined in the R tool library 'fuzzyAhp', and hence used to simulate the Fuzzy AHP method with a comparison matrix as input.

The weights computed and the relative importance of features are then examined to evaluate the critical factors and provide incentives to retain talented employees. The results of the analysis and subsequent discussion are detailed below.

RESULTS AND DISCUSSION

Analysis of Random Forest-Based Model

Outcomes of the random forest-based model generated using 'randomForest' package of the R tool are displayed in Tables 5 and 6.

The training set comprises 292 data points, while the test set had 73 data points (total 365 employees). The parameters i.e., the number of predictors 'mtry' and the number of trees 'ntree' are tuned to choose the best model, a sample of the tuning dataset is illustrated below in Table 5 for the number of trees i.e., ntree=100 and mtry=2,4 and 6.

The number of optimal predictors is supposed to be $n/3$; where n are the number of variables considered in the model (Adam *et al.*, 2014) while mtry values vary from 1 to $n-1$ i.e., 6 in this case for 7 predictors. This is to be tested for which the RMSE values are plotted against the number of predictors 'mtry' and tuned by changing the number of trees 'ntree'.

Table 7. Predictive Accuracy of the Gradient Boosting Model

n.trees	RMSE	R-squared
50	203.4	0.854
50	204.9077	0.843
50	204.4687	0.852
100	204.7394	0.853
100	205	0.86

Figure 7. Relative Importance of the variables in the Gradient Boosting Model
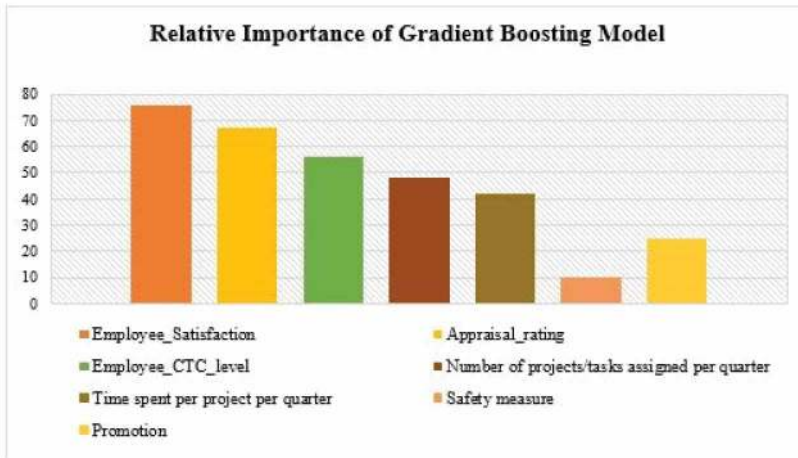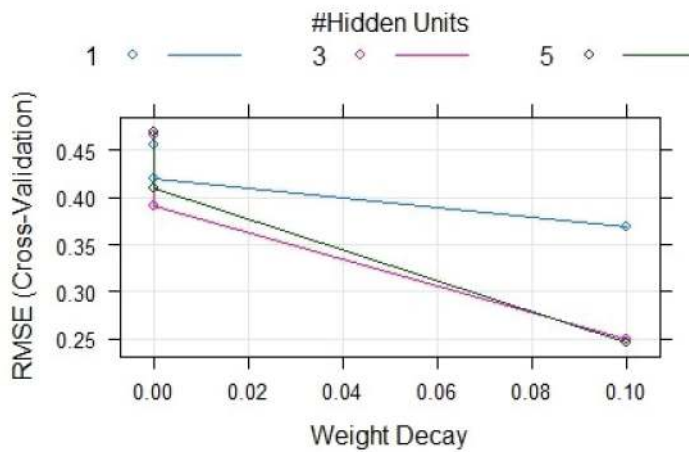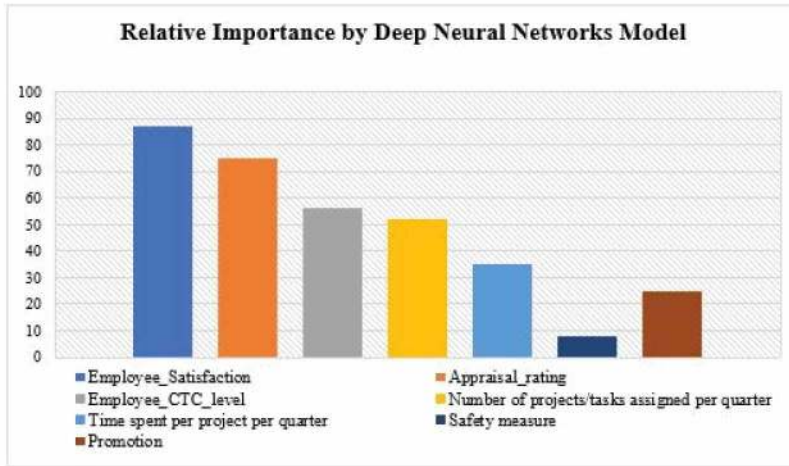

Figure 8. Neural Network model output

Table 8. Confusion Matrix of the Deep Neural Network Model

Classified	Data-points	Positive	Negative
		0	1
TRUE	0	6	1
FALSE	1	0	5

Figure 9. Relative importance plot of Deep Neural Networks model

From Table 5, the optimal accuracy of prediction of this model i.e., accuracy at the minimum root mean square error value (RMSE) is 82.5%, which implies that 82.5% of data points of the test set were accurately classified.

Figure 6 illustrates the variation of RMSE with the number of predictors i.e., mtry for values: mtry=2, 3, 4, 5 and 6, which is tuned by the number of trees =100,200 and 500. It is found that the optimal values of mtry=2 and the number of trees=100 achieves the best model indicated by minimum RMSE of 290.97. This is also in-line with the theoretical proposition that optimal mtry= number of predictors/3, which in this case =7/3~= 2. Further, the weightage assigned to the predictors for the Random Forest model is tabulated in Table 6.

It is observed from Table 6 that 'Employee_Satisfaction' and 'Appraisal_rating' are the most significant predictors, followed by 'Employee_CTC_level', 'Number of projects/tasks assigned per quarter' and 'Time spent per project per quarter'.

Analysis of Gradient Boosting Model

Table 7 shows the optimal accuracy of prediction of this model, i.e., accuracy at the minimum root mean square error value (RMSE) is 85.4%, which implies that 85.4% of the test set's data points were accurately classified. The ideal size of the classification tree, i.e., is n.tree=50. This is found to be an improvement over the Random Forest model due to the ensemble of classifiers and aggregation of output from multiple decision trees.

Further, weightage assigned to the predictors for the Gradient Boost model is illustrated in Figure 7.

It is observed from Figure 7 that 'Employee_Satisfaction' (bright orange color) is the most significant predictor followed by the 'Appraisal_rating' (yellow color), 'Employee_CTC_level' (green color) and 'Number of projects/tasks assigned per quarter' (brown color).

To boost accuracy through hidden layer processing, the deep neural networks model is built.

Analysis of Deep Neural Network-Based Model

The DNN based model performance is illustrated in Figure 8.

Figure 10. Performance Comparison of the models on the FMCG company data
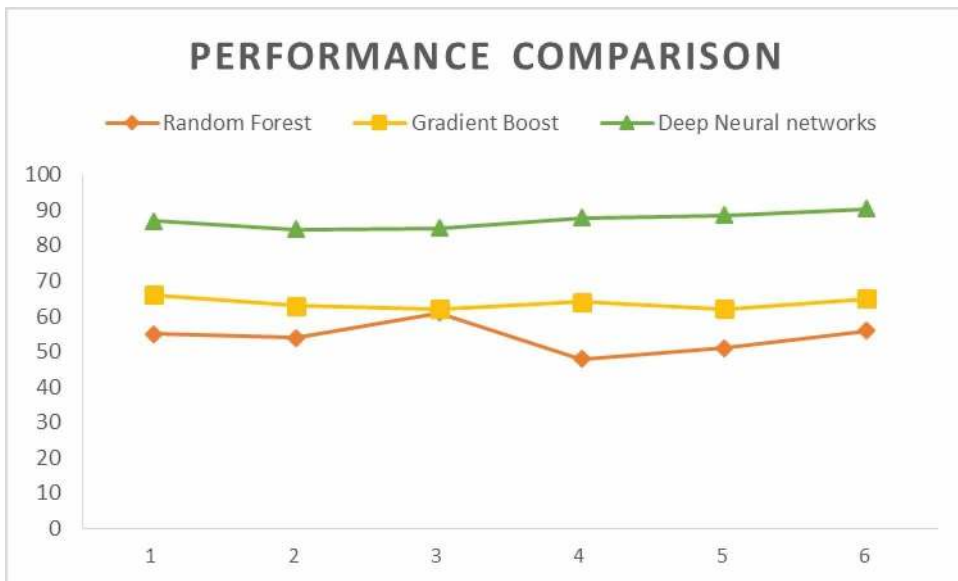
A simple model showed 69.8% accuracy, but an ensemble of 100 such simple DNN models (Sharma et al.,2018) boosted the accuracy to 91.6% with the least root mean square error (RMSE) of 0.25 for five hidden layer units and hence five hidden layers are used in the model.

Thus, in Figure 8, since the least root mean square error (RMSE) value is least (0.25) for the number of hidden layers=5 with minimum weight decay of 0.1, 5 hidden layers are considered for the deep neural network implementation.

From Table 8, the optimal accuracy of this model's prediction, i.e., accuracy at the minimum root mean square error value (RMSE), is computed from the above confusion matrix by the formula:

$$Accuracy = \left(Number\ of\ Correctly\ classified\ Instances\right) / Total\ Instances$$
$$= (6+5) / (6+5+0+1) = 91.6\%$$

Further, weightage assigned to the predictors is tabulated in Figure 9.

It is observed from Figure 9 that 'Employee_Satisfaction' (in blue) is the most significant predictor followed by the current 'Appraisal_rating' (orange), 'Employee_CTC_level' (grey), and the 'Number of projects/tasks assigned per quarter' (yellow).

In order to compare the models' performance on new employee records for validation, 50 datapoints from a mid-sized FMCG company are collected and the predictive performance is compared as illustrated in Figure 10.

From Figure 10, it can be inferred that the Deep Neural Network model predicts the 'Turnover_ status' of employees closest to the actual status followed by Gradient Boost and then the Random Forest model.

In order to determine which are the most important factors influencing employee 'Turnover_ status', a regression model is formulated below.

Table 9. Robustness test results across models

Variables	Model 1	Model 2	Model3	VIF
Employee_Satisfaction	-0.32***(0)	-0.32***(0)	-0.32***(0)	1.13
Appraisal_rating	0.04***(0.002)	0.04***(0.002)	0.04***(0.002)	1.02
Number of projects/tasks assigned per quarter	-0.04***(0)	-0.04***(0)	-0.04***(0)	1.52
Time spent per project per quarter	0.03***(0)	0.03***(0)	0.03***(0)	1.07
Promotion	-0.15***(0.006)	-0.15***(0.006)	-0.15***(0.006)	1.73
Employee_CTC_level	0.06***(0.004)	0.06***(0.004)	0.06***(0.004)	1.96
Safety measure	-0.13***(0.003)	-0.13***(0.003)	-0.13***(0.003)	1.01
Employee_Satisfaction* Appraisal_rating		-0.012***(0)	-0.009***(0)	1.03
Employee_Satisfaction* Number of projects/tasks assigned per quarter		-0.032 (0.75)	-0.029 (0.75)	2.56
Employee_Satisfaction* Time spent per project per quarter		0.034 (0.57)	0.042 (0.57)	2.08
Employee_Satisfaction* Promotion		0.062 (0.134)	0.049 (0.134)	5.69
Employee_Satisfaction* Employee_CTC_level		0.046 ***(0)	0.042 ***(0)	4.29
Employee_Satisfaction* Safety measure		0.065 (0.52)	0.068 (0.52)	1.2
Appraisal_rating * Number of projects/tasks assigned per quarter		0.004***(0.003)	0.008***(0.003)	1.65
Appraisal_rating * Time spent per project per quarter		0.04***(0)	0.037***(0)	1.44
Appraisal_rating * Promotion		-0.146 (0.524)	-0.153 (0.524)	5.41
Appraisal_rating * Employee_CTC_level		-0.092*** (0.005)	-0.097*** (0.005)	5.8
Appraisal_rating * Safety measure		-0.153 (0.229)	-0.156 (0.229)	1.81
Number of projects/tasks assigned per quarter * Time spent per project per quarter		-0.003 (0.932)	-0.005 (0.932)	2.02
Number of projects/tasks assigned per quarter * Promotion		0.001 (0.930)	0.0007 (0.930)	5.63
Number of projects/tasks assigned per quarter * Employee_CTC_level		0.0006 (0.468)	0.001 (0.468)	4.23
Number of projects/tasks assigned per quarter * Safety measure		-0.001 (0.485)	-0.002 (0.485)	1.14
Time spent per project per quarter * Promotion		-0.012 (0.348)	-0.011 (0.348)	5.74
Time spent per project per quarter * Employee_ CTC_level		-0.012 (0.168)	-0.017 (0.168)	3.73
Time spent per project per quarter * Safety measure		-0.011 (0.163)	-0.016 (0.163)	0.64
Safety measure *Promotion		0.156 (0.145)	0.15 (0.145)	1.09
Safety measure * Employee_CTC_level		0.092 (0.161)	0.095 (0.161)	2.52
Promotion * Employee_CTC_level		0.050 (0.167)	0.056 (0.167)	0.88
Employee_Satisfaction* Appraisal_rating * Employee_CTC_level			0.074***(0.001)	1.73
Appraisal_rating * Number of projects/tasks assigned per quarter * Time spent per project per quarter			0.082***(0.002)	1.23
Adj. R-Square	0.79	0.80	0.83	
DW Statistic	2.05(0.23)	2.29(0.03)	2.12(0.4)	
LM Statistic	2.19(0.4)	2.66(0.52)	2.65(0.61)	

Analysis of Multiple Linear Regression-Based Model

The variation in variable importance from the above importance results of Random Forest, Gradient Boost and Deep Neural Network techniques in Table 6, Figure 7, and Figure 9 is to be examined by implementing multiple linear regression-based models (summary) illustrated in Table 9.

The regression models formulated in Table 9 ensure that all the robustness tests for the assumption of linear regression, namely multi-collinearity, linearity, auto-correlation, and homoskedasticity are validated (Abadie et al.,2020). For the employees, the robustness tests, namely Durbin-Watson, Langrange Multiplier (LM Coefficient), and Variance Inflation Factor (VIF) are run to ensure the reliability of the model variable significance.

According to the thumb rule of the Durbin Watson test (Lumbantobing et al.,2020), the value of the Durbin Watson statistic (DW) must lie between 2 and 4 with a value tending closer to 2 implying that auto-correlation is not present in the dataset. Moreover, the significance value rho must be closer to 0. Similarly, for Langrange Multiplier (LM) test (Chauhan, Pande and Sharma,2020), if the p-value statistic is greater than level of significance 'alpha', the null hypothesis of homoscedasticity is validated.

The VIF (Variance Inflation Factor) (Vörösmarty and Dobos,2020) for all the predictors is expected to be <10 to indicate that there is no multi-collinearity in the data.

Table 9 reports the coefficients and p-value statistics (in parenthesis) and above robustness statistics (DW, LM and VIF). *** indicates a 1% statistical significance level.

In Table 9, three regression models are implemented: Model 1 is implemented by regressing the outcome variable 'Turnover_Status' on the individual predictors (direct effects) considered in the machine learning model. At 95% significance level, 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level' and 'Number of projects/tasks assigned per quarter' were found to be highly significant based on coefficient and p-value. The value of adjusted R-squared is 0.79 i.e., 79% is explained. Further, in Model 1, the Durbin Watson Statistic (DW) is reported to be 2.05 with a p-value of 0.23. The significance value rho is 0.002. Both these statistics imply that there is no presence of autocorrelation in the dataset. The Langrange Multiplier (LM) is reported to be 2.19 with a p-value of 0.4, which is greater than the level of significance alpha=0.05 (5% significance). This implies that the dataset is homoscedastic.

The above model relies on the assumption that employees only consider one factor at a time hierarchically before deciding to leave the organization. However, in real-time, employees tend to consider multiple factors. For instance, an employee will simultaneously consider factors like 'Employee_Satisfaction', 'Appraisal_rating' and 'Number of projects/tasks assigned per quarter' to decide whether to leave an organization. If he/she finds that though his 'Appraisal_rating' is high, his workload in terms of the 'Number of projects/tasks assigned per quarter' is also increased, he may still quit the organization though other factors are favorable. In light of this scenario, the regression model can be augmented with variable interactions. This implies that individual variables like 'Employee_Satisfaction', 'Appraisal_rating' can be multiplied pair-wise to form a new interaction term 'Employee_Satisfaction*Appraisal_rating'. Similarly, all the seven predictors are taken two at a time and assuming that no two predictor variables are multiplied twice, $^7C_2$, i.e., 21 pair-wise interaction terms are initially factored in the regression models. Thus, a new regression Model 2 is implemented augmenting Model 1 with pair-wise interaction effects.

From Model 2, it is found that interaction variables 'Employee_Satisfaction*Appraisal_rating', 'Appraisal_rating*Number of projects/tasks assigned per quarter', 'Employee_Satisfaction*Employee_CTC_level', 'Appraisal_rating * Time spent per project per quarter' and 'Appraisal_rating *Employee_CTC_level' are significant (highlighted in red), implying and corroborating the result in Model 1 that employees simultaneously consider 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level' followed by 'Number of projects/tasks assigned per quarter' and 'Time spent per project per quarter'. The value of adjusted R-squared is 0.80 i.e., 80% is explained. For Model 2, the Durbin Watson Statistic (DW) is reported to be 2.29 with a p-value of 0.03, showing

Figure 11. Feature contribution chart from the Multiple Linear Regression Model
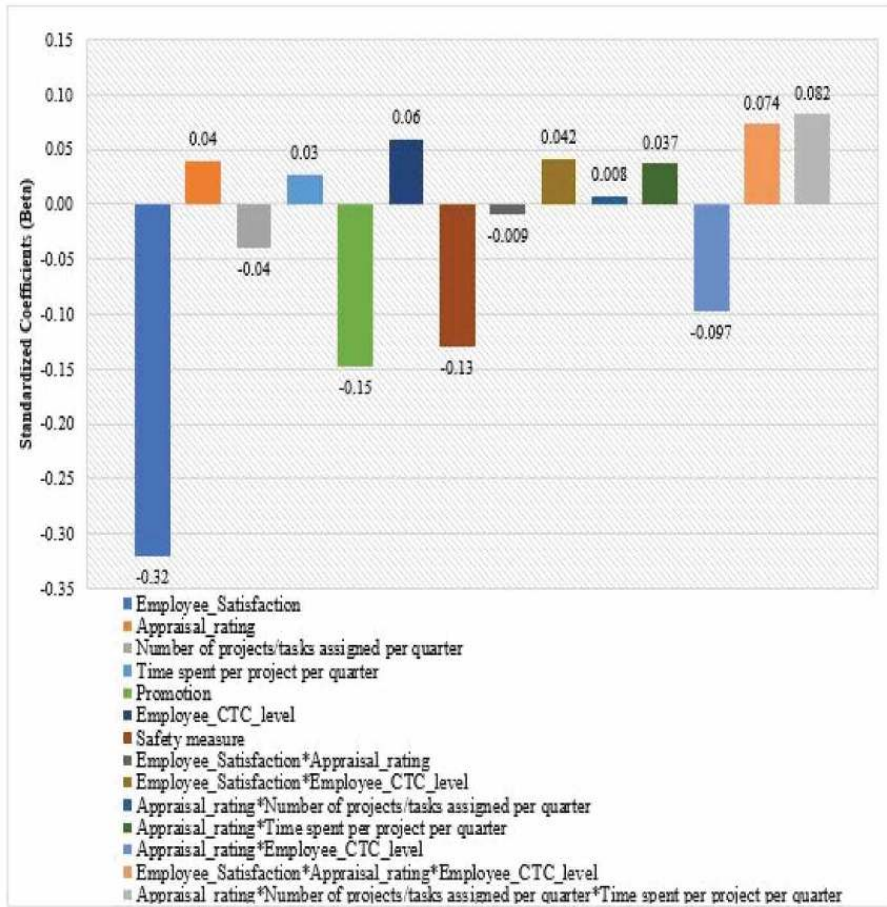
no autocorrelation presence in the dataset. Similarly, the Langrange Multiplier (LM) is reported to be 2.66 with a p-value of 0.52 (> alpha=0.05), which implies that the dataset is homoscedastic.

Model 3 is the extension of Model 2, with triplet interaction variables corresponding to Table 11. This is done to factor in the real-time consideration of employees who simultaneously consider three or more factors for leaving the organization. Other triplet interactions are not formulated since they are not significant variables and cannot form significant interactions. The triplet interaction terms are from the top three significant variables i.e., one from the interaction of 'Employee_Satisfaction', 'Appraisal_rating' and 'Employee_CTC_level'as 'Employee_Satisfaction*Appraisal_rating* Employee_CTC_level' and the other from the interaction of 'Appraisal_rating', 'Number of projects/ tasks assigned per quarter' and 'Time spent per project per quarter' as 'Appraisal_rating * Number of projects/tasks assigned per quarter * Time spent per project per quarter'. The variables included in Models 1 and 2 are still significant, along with the additional variables. Overall model fit, as evident Adj.R-Squared shows an increase from Model 2 with a value of .83. This shows that the included triplet interaction variable as anticipated. For Model 3, the Durbin Watson Statistic (DW) is reported to be 2.12 with a p-value of 0.4, showing no autocorrelation presence in the dataset. Similarly, the Langrange Multiplier (LM) is reported to be 2.65, with a p-value of 0.61(> alpha=0.05), which implies that the dataset is homoscedastic. Figure 11 depicts the features and most significant (in

Table 10. Comparison Matrix

Criteria	Employee_ Satisfaction	Appraisal_ rating	Employee_ CTC_level	Number of projects/ tasks assigned per quarter	Time spent per project per quarter	Promotion	Safety measure
Employee_ Satisfaction	(1,1,1)	(2,3,4)	(2,3,4)	(4,5,6)	(2,3,4)	(4,5,6)	(6,7,8)
Appraisal_ rating	(1/4,1/3,1/2)	(1,1,1)	(2,3,4)	(2,3,4)	(4,5,6)	(6,7,8)	(9,9,9)
Employee_ CTC_level	(1/4,1/3,1/2)	(1/4,1/3,1/2)	(1,1,1)	(2,3,4)	(4,5,6)	(6,7,8)	(2,3,4)
Number of projects/tasks assigned per quarter	(1/6,1/5,1/4)	(1/4,1/3,1/2)	(1/4,1/3,1/2)	(1,1,1)	(9,9,9)	(6,7,8)	(2,3,4)
Time spent per project per quarter	(1/4,1/3,1/2)	(1/6,1/5,1/4)	(1/6,1/5,1/4)	(1/9,1/9,1/9)	(1,1,1)	(4,5,6)	(9,9,9)
Promotion	(1/6,1/5,1/4)	(1/8,1/7,1/6)	(1/8,1/7,1/6)	(1/8,1/7,1/6)	(1/6,1/5,1/4)	(1,1,1)	(2,3,4)
Safety measure	(1/8,1/7,1/6)	(1/4,1/3,1/2)	(1/9,1/9,1/9)	(1/4,1/3,1/2)	(1/9,1/9,1/9)	(1/4,1/3,1/2)	(1,1,1)

Table 11. The weights for criteria

Criteria	N_i (Weights)
Employee_Satisfaction	0.351
Appraisal_rating	0.207
Employee_CTC_level	0.179
Number of projects/tasks assigned per quarter	0.131
Time spent per project per quarter	0.071
Promotion	0.034
Safety measure	0.027

terms of p-value and coefficients) [highlighted in red in Table 9] interactions (pair-wise and triplet) to predict the 'Turnover_status' variable.

The significance of these interaction variables 'Employee_Satisfaction* Appraisal_rating * Employee_CTC_level' and 'Appraisal_rating*Number of projects/tasks assigned per quarter* Time spent per project per quarter'also illustrated in the feature interaction chart in Figure 11. This significance indicates that employees can simultaneously consider 'Employee_Satisfaction', 'Appraisal_rating', and 'Employee_CTC_level' and, at the same time, 'Appraisal_rating', 'Number of projects/tasks assigned per quarter', and 'Time spent per project per quarter' for leaving the organization.

In Figure 11, the presence of the factors of the 'Appraisal_rating' (orange), 'Employee_CTC_ level'(navy blue) and 'Time spent per project per quarter' (dark blue colour) are found to be significant

Table 12. Summary of Variable Importance across all Models

Key factors	Random Forest	Gradient Boost	Neural Network	Regression Model	Fuzzy AHP
Top four	Employee_ Satisfaction, Appraisal_ _rating, Employee_CTC _level, Number of projects/tasks assigned per quarter	Employee_ Satisfaction, Appraisal _rating, Employee_CTC _level, Number of projects/tasks assigned per quarter	Employee_ Satisfaction, Appraisal _rating, Employee_CTC_ level, Number of projects/tasks assigned per quarter	Employee_ Satisfaction, Appraisal _rating, Employee_CTC_ level, Number of projects/tasks assigned per quarter	Employee_ Satisfaction, Appraisal _rating, Employee_CTC_ level, Number of projects/tasks assigned per quarter
Remaining factors	Time spent per project per quarter, Promotion and Safety measure	Time spent per project per quarter, Promotion and Safety measure	Time spent per project per quarter, Promotion and Safety measure	Time spent per project per quarter, Promotion and Safety measure	Time spent per project per quarter, Promotion and Safety measure

Table 13. Results and Inferences

SNo.	Hypothesis	Random Forest	Gradient Boost	Neural Network	Regression Model	Fuzzy AHP
H_{0a}	The factor 'Employee_ Satisfaction' influences 'Turnover_status'.	Validated	Validated	Validated	Validated with p-value of 0	Validated
H_{0b}	The factor 'Appraisal_ rating' impacts 'Turnover_status'.	Validated	Validated	Validated	Validated with p-value of 0.002	Validated
H_{0c}	'Employee_CTC_level' and the 'Number of projects/tasks assigned per quarter' are significant predictors of 'Turnover_status'.	Validated	Validated	Validated	Validated with p-values of 0.004 for 'Employee_CTC_level' and p-value of 0 for 'Number of projects/ tasks assigned per quarter' respectively	Validated

positive drivers of 'Turnover_status' with respective standardized coefficients 0.04, 0.03 and 0.06. This implies that keeping other factors constant, one unit of increase in 'Appraisal_rating' of the employee, one unit of increase in 'Employee_CTC_level' and one unit of increase in 'Time spent per project per quarter' can increase 'Turnover_status' rate by 0.04, 0.03 and 0.06. On the other hand, 'Employee_Satisfaction'(light blue), 'Number of projects/tasks assigned per quarter' (grey colour), 'Promotion' (light green) and 'Safety measure' (brown) are found to be significant negative drivers of 'Turnover_status', indicating that keeping other factors constant, one unit of increase in each of the above factors i.e., 'Employee_Satisfaction' and 'Promotion' can lead to -0.32, -0.04, -0.15 and -0.13 units in 'Turnover_status'.

Similarly, the pair-wise interactions 'Employee_Satisfaction*Employee_CTC_level' (mustard colour), 'Appraisal_rating*Number of projects/tasks assigned per quarter'(dark blue) and 'Appraisal_rating *Time spent per project per quarter' (dark green colour) are the most significant positive interactions with standardized coefficients of 0.042, 0.008 and 0.037, indicating that these

interaction variables are vital to 'Turnover_status' variable. One unit of increase in interaction term 'Employee_Satisfaction*Employee_CTC_level' implies 0.042 units increment in 'Turnover_status'.

Similarly, the triplet interactions 'Employee_Satisfaction*Appraisal_rating *Employee_CTC_level'(light orange colour) and 'Appraisal_rating*Number of projects/tasks assigned per quarter *Time spent per project per quarter' (light grey colour) are found to be significant positive interactions with standardized beta coefficients of 0.074 and 0.082 respectively. One unit of increase in triplet interaction term 'Employee_Satisfaction*Appraisal_rating*Employee_CTC_level' implies 0.074 units increase in 'Turnover_status'.

The pair-wise interaction 'Appraisal_rating *Employee_CTC_level' (sky blue colour) are found to be significant negative interactions with standardized beta coefficients of -0.097 respectively.

Thus, employees consider 'Employee_Satisfaction', 'Appraisal_rating', and 'Employee_CTC_level' as the three most critical factors before leaving an organization. Thus, from the robustness test results, it is confirmed that assumptions for multiple linear regression are satisfied, thus validating the variable significance computed by the machine learning models.

Further to corroborate the validation results from a stakeholder perspective to understand the weightage assigned to each factor by stakeholders like customers in an uncertain demand scenario as explicated in methodology, Fuzzy AHP is adopted below.

Weightages of Factors Using Fuzzy –AHP

Fuzzy-AHP procedure in R first inputs the pair-wise comparison matrix of the criteria as shown in Table 10.

The comparison matrix illustrates the relative importance values adopting the Fuzzy AHP scale defined in Wind and Saaty(1980). For instance, 'Employee_Satisfaction' is considered more important than the 'Appraisal_rating' represented by fuzzy number (2,3,4) while the factor 'Number of projects/tasks assigned per quarter' is less important as 'Appraisal_rating' represented by fuzzy number (1/4,1/3,1/2).

The function 'calculate_Weights()' pre-defined under the library in R tool 'fuzzyAhp' computes the normalized weights N_i [from the above pair-wise comparison matrix] illustrated in Table 11.

The 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level', 'Number of projects/tasks assigned per quarter', and 'Time spent per project per quarter' are the most critical factors, as illustrated in Table 11.

The overall results of variable importance across all the models (machine learning, regression, and multi-criteria) are illustrated in Table 12.

In all the models in Table 12, the top two predictors are found to be: 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level' and 'Number of projects/tasks assigned per quarter', and these are the critical factors that have implications for the management to take steps for employee retention.

The other significant factors include 'Time spent per project per quarter', 'Promotion' and 'Safety measure'. The overall inferences from the machine learning models validated through the regression model and Fuzzy AHP are summarized in Table 13.

Thus, the hypothesis is validated for all the above models that for an employee, his 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level' and workload in terms of 'Number of projects/tasks assigned per quarter' determine his 'Turnover_status'.

Theoretical Contributions

This study makes the following four contributions to literature.

Firstly, prior studies in the domain of employee churn prediction (Al-Ansari & Alshare, 2019) have attempted to model employee churn without considering the interaction or interplay between factors influencing churn. In this paper, the importance of the factors is also analyzed for making customized recommendations to alleviate employee attrition rate. Pair-wise interactions and triplet

interactions variables between individual factors are also included in the regression model to test for interaction effects between variables. This significance indicates that employees can simultaneously consider the factors 'Employee_Satisfaction', 'Appraisal_rating' and 'Employee_CTC_level'. At the same time, the factors 'Appraisal_rating', 'Number of projects/tasks assigned per quarter' and 'Time spent per project per quarter' have high interaction significance indicating these three-factor combinations are also considered by employees to decide whether to leave or stay in the organization.

Secondly, while some of the existing studies have adopted machine learning techniques for employee churn prediction (Ogbonnaya et al.,2017), the variable importance computed by machine learning is not validated by a robust methodology. In this paper, the results are validated using multiple regression techniques and multi-criteria approach Fuzzy AHP to confirm the validity. Thus, this model can be re-calibrated for different datasets to predict 'Turnover_status' of an employee real-time.

Thirdly, existing employee churn prediction models are not targeted towards the employees' specific needs leaving the company. This paper devises an employee churn prediction model that provides insights on which employees are valuable and devises strategies for employee retention when recalibrated. These strategies thus contribute to the realm of global IT management and workforce retention.

IMPLICATIONS FOR PRACTICE AND FUTURE SCOPE

The output from the study accurately predicts employee churn as discussed in the paper. The implications for practice are three-fold: first on the HR manager, second on the organization and third on the employees.

Implications for HR Manager

The constructs 'Employee_Satisfaction', 'Appraisal_rating', 'Employee_CTC_level' and 'Number of projects/tasks assigned per quarter' are found to be factors simultaneously considered by employees to decide whether to stay or leave an organization. This insight can be taken into account to devise effective strategies to retain loyal employees.

The HR manager can recalibrate the above model and increase 'Employee_Satisfaction' by reducing the 'Number of projects/tasks assigned per quarter' for over-worked employees and incentivizing them with 'Promotion', 'Employee_CTC_level' hikes, and high 'Appraisal_rating' appease them and reduce attrition levels in the organization.

By moderating the 'Number of projects/tasks assigned per quarter' and by providing regular incentives, the disgruntled employees can be appeased to continue in the organization and thus alleviate the employee 'Turnover_status'.

Implications for Organization and Global IT Management

By identifying the reasons behind valuable employees leaving the organization, the employee work and project duration allocation policies can be made more flexible. Further, culture-agnostic work policies like "work from home" can be provided to employees who need to put in long hours to complete projects. Employees can be entertained with recreational programmes to relieve the monotony of hard work, and valuable employees can be periodically awarded to contribute to the organization. The implications from the study thus are applicable and useful for global IT management across different cultures. The need to incentivize the employee and make work hours more flexible will help in increasing productivity in the workplace and strengthen employee commitment to the organization.

Implications for Employees

Employees can use the model's insights to voice their concerns more specifically to the manager in terms of work hours, viz-a-viz incentives. The model provides insights into the causes for employee resignation, which helps employees understand their organizational culture and enable them to plan

their future career path by eliminating career choices that are not in line with their work ethic. Thus, this enables better career planning and, thereby a meteoric progress up the corporate ladder which is critical in this competitive world.

This study can be extended to building a simulation model that considers the above factors and recommends how to retain employees real-time by tuning the factors and examining the sensitivity of these factors on the employee's decision-making.

Further, in the light of the current pandemic i.e., the Corona Virus (Song et al.,2020) is found to have transformed the work environment to the notion of 'remote workplace' or 'work-from-home' with limited commuting to work. However, there may be a need for people to visit the office occasionally. There may be a chance to contract the virus even during these rare visits, if appropriate precautions are not followed. This has led to the phenomenon of 'Social distancing'. This phenomenon involves setting up norms to ensure that people maintain physical distance from each other even during the occasional workplace visits. This is ensured to avoid the spread of the virus. The adherence to these norms can also influence the employee's decision-making whether or not to leave an organization. Employees may leave the organization if social distancing norms are not strictly followed. Thus, this factor can also be incorporated in future studies to investigate its impact on employee churn.

CONCLUSION

The paper has attempted to compare the predictive performance of random forest, gradient boost, and deep neural networks-based models and present the significance of the factors and the interplay between them in predicting employee 'Turnover_status'.

The deep neural network model is observed to out-perform the other two models in terms of predictive accuracy and performance on the FMCG company data used for validation.

The regression models are found to validate the variable importance given to predictors in the above models in terms of 'Employee_Satisfaction', 'Employee_CTC_level', 'Appraisal_rating' and 'Number of projects/tasks assigned per quarter' as significant predictors in terms of both direct effect (individual predictors) and interaction effect (pair-wise interaction variables and triplet interaction variables). This finding is useful to indicate that these factors are significant reasons for employee 'Turnover_status'. The Fuzzy AHP model is found to validate the variable importance given to predictors. This finding is useful to indicate that these factors play a significant role in a frustrated employee's decision-making.

The talented employees who are likely to leave the company can be incentivized by reducing their 'Number of projects/tasks assigned per quarter' (if high) and increasing their 'Employee_CTC_level' to satisfy the employee according to Maslow's Hierarchy of Needs (Maslow,1943), which emphasizes the role of self-actualization and physical needs in a person's life.

REFERENCES

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis. *Econometrica*, *88*(1), 265–296. doi:10.3982/ECTA12675

Adam, E., Mutanga, O., Abdel-Rahman, E. M., & Ismail, R. (2014). Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: Exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing*, *35*(2), 693–714. doi:10.1080/01431161.2013.870676

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *Algorithms*, *5*(9), 22–26.

Al-Ansari, M. A., & Alshare, K. (2019). The Impact of Technostress Components on the Employees Satisfaction and Perceived Performance: The Case of Qatar. *Journal of Global Information Management*, *27*(3), 65–86. doi:10.4018/JGIM.2019070104

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal*, *26*(1), 1–11. doi:10.1111/1748-8583.12090

Bendazzoli, S., Brusini, I., Damberg, P., Smedby, Ö., Andersson, L., & Wang, C. (2019). Automatic rat brain segmentation from MRI using statistical shape models and random forest. In Medical Imaging 2019: Image Processing (109492O). International Society for Optics and Photonics. doi:10.1117/12.2512409

Brown, A., Forde, C., Spencer, D., & Charlwood, A. (2008). Changes in HRM and job satisfaction, 1998–2004: Evidence from the Workplace Employment Relations Survey. *Human Resource Management Journal*, *18*(3), 237–256. doi:10.1111/j.1748-8583.2008.00069.x

Chauhan, S., Pande, R., & Sharma, S. (2020). The causal relationship between Indian energy consumption and the GDP: A shift from conservation to feedback hypothesis post economic liberalisation. *Theoretical & Applied Economics*, *27*(3), 203–212.

Cheng, L. C., Wu, C. C., & Chen, C. Y. (2019). Behavior analysis of customer churn for a customer relationship system: An empirical case study. *Journal of Global Information Management*, *27*(1), 111–127. doi:10.4018/JGIM.2019010106

Edwards, M. R., & Edwards, T. (2012). Procedural justice and identification with the acquirer: The moderating effects of job continuity, organisational identity strength and organisational similarity. *Human Resource Management Journal*, *22*(2), 109–128. doi:10.1111/j.1748-8583.2011.00176.x

El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M. (2020). Predicting employee attrition using tree-based models. *The International Journal of Organizational Analysis*, *28*(6), 1273–1291. doi:10.1108/IJOA-10-2019-1903

Frisk, J., & Bannister, F. (2017). Improving the use of analytics and big data by changing the decision-making culture: A design approach. *Management Decision*, *55*(10), 2074–2088. doi:10.1108/MD-07-2016-0460

Fuchs, S., & Edwards, M. R. (2012). Predicting pro-change behaviour: The role of perceived organisational justice and organisational identification. *Human Resource Management Journal*, *22*(1), 39–59. doi:10.1111/j.1748-8583.2011.00167.x

Herzberg, F. (2005). The motivation-hygiene theory. *Organizational Behavior: Essential Theories of Motivation and Leadership, 1*(1),61.

Huang, J., Wang, H., & Kochenberger, G. (2017). Distressed Chinese firm prediction with discretized data. *Management Decision*, *55*(5), 786–807. doi:10.1108/MD-08-2016-0546

Hunter, M. G., Tan, F. B., & Tan, B. C. (2008). Voluntary turnover of information systems professionals: A cross-cultural investigation. *Journal of Global Information Management*, *16*(4), 46–66. doi:10.4018/jgim.2008100103

In Class Prediction Competition. (2017). *Employee Churn Prediction*. Retrieved 2019 from: https://www.kaggle.com/c/employee-churn-prediction/data

Kaya, E., Agca, M., Adiguzel, F., & Cetin, M. (2019). Spatial data analysis with R programming for environment. *Human and Ecological Risk Assessment*, *25*(6), 1521–1530. doi:10.1080/10807039.2018.1470896

Keegan, A., & Den Hartog, D. (2019). Doing it for themselves? Performance appraisal in project-based organisations, the role of employees, and challenges to theory. *Human Resource Management Journal*, *29*(2), 217–237. doi:10.1111/1748-8583.12216

Kong, Y. S., Abdullah, S., Schramm, D., Omar, M. Z., & Haris, S. M. (2019). Development of multiple linear regression-based models for fatigue life evaluation of automotive coil springs. *Mechanical Systems and Signal Processing*, *118*, 675–695. doi:10.1016/j.ymssp.2018.09.007

Lumbantobing, I. P., Sulivyo, L., Sukmayuda, D. N., & Riski, A. D. (2020). The Effect of Debt to Asset Ratio and Debt to Equity Ratio on Return on Assets in Hotel, Restaurant, and Tourism Sub Sectors Listed on Indonesia Stock Exchange for the 2014-2018 Period. *International Journal of Multicultural and Multireligious Understanding*, *7*(9), 176–186. doi:10.18415/ijmmu.v7i9.1982

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396. doi:10.1037/h0054346

Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal*, *44*(6), 1102–1121.

Mohr, D. C., Young, G. J., & Burgess, J. F. Jr. (2012). Employee turnover and operational performance: The moderating effect of group-oriented organisational culture. *Human Resource Management Journal*, *22*(2), 216–233. doi:10.1111/j.1748-8583.2010.00159.x

Ogbonnaya, C., Daniels, K., & Nielsen, K. (2017). Does contingent pay encourage positive employee attitudes and intensify work? *Human Resource Management Journal*, *27*(1), 94–112. doi:10.1111/1748-8583.12130

Raman, R., Bhattacharya, S., & Pramod, D. (2019). Predict employee attrition by using predictive analytics. *Benchmarking*, *26*(1), 2–18. doi:10.1108/BIJ-03-2018-0083

Sharma, A., Bhatnagar, V., & Bansal, A. (2018). SENSEX price fluctuation forecasting comparison between global indices and companies making it. *Journal of Global Information Management*, *26*(3), 90–104. doi:10.4018/JGIM.2018070107

Simon, H. A., & March, J. (1976). Administrative Behavior Organization. New York: Free Press.

Singh, A., & Prasher, A. (2019). Measuring healthcare service quality from patients' perspective: Using Fuzzy AHP application. *Total Quality Management & Business Excellence*, *30*(3-4), 284–300. doi:10.1080/14783363.2017.1302794

Singh, S. K., & Del Giudice, M. (2019). Big data analytics, dynamic capabilities and firm performance. *Management Decision*, *57*(8), 1729–1733. doi:10.1108/MD-08-2019-020

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017). Evaluation of machine learning models for employee churn prediction. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 1016-1020). IEEE. doi:10.1109/ICICI.2017.8365293

Song, F., Shi, N., Shan, F., Zhang, Z., Shen, J., Lu, H., Ling, Y., Jiang, Y., & Shi, Y. (2020). Emerging 2019 novel coronavirus (2019-nCoV) pneumonia. *Radiology*, *295*(1), 210–217. doi:10.1148/radiol.2020200274 PMID:32027573

Trauth, E. M., Quesenberry, J. L., & Huang, H. (2008). A multicultural analysis of factors influencing career choice for women in the information technology workforce. *Journal of Global Information Management*, *16*(4), 1–23. doi:10.4018/jgim.2008100101

Vörösmarty, G., & Dobos, I. (2020, June). Green purchasing frameworks considering firm size: a multicollinearity analysis using variance inflation factor. In *Supply Chain Forum: An International Journal* (pp. 1-12). Taylor & Francis. doi:10.1080/16258312.2020.1776090

Wang, P., Wang, J., & Zhang, J. (2018). Methodological Research for Modular Neural Networks Based on "an Expert With Other Capabilities". *Journal of Global Information Management*, *26*(2), 104–126. doi:10.4018/JGIM.2018040105

Wind, Y., & Saaty, T. L. (1980). Marketing applications of the analytic hierarchy process. *Management Science*, *26*(7), 641–658. doi:10.1287/mnsc.26.7.641

APPENDIX: QUESTIONNAIRE FOR IT TRAINING DATASET AND FMCG VALIDATION DATASET OF EMPLOYEES

Q1. What is your age group?

Q2. What income category do you fall under? (Low, Medium or High)

Q3. What are the factors that you consider for continuing to work in your organization?

Q4. What weightage do you assign to the following factors: Employee_Satisfaction, Appraisal_rating, Number of projects/tasks assigned per quarter, Employee_CTC_level, Time spent per project per quarter (on a scale of 1-9, 1 being least important and 9 being most important)?

Q5. What is your Employee_Satisfaction in your current organization (on a scale of 0-2, 0 being low and 2 being the highest)?

Q6. How many projects are you assigned to currently for this quarter?

Q7. How would you rate your Employee_CTC_level on a scale of 0-2(0 is low, 1 is medium and 2 is high)?

Q8. How much time have you been allocated (in hours) for each project for this quarter?

Praveen Ranjan Srivastava has more than 15 years of teaching and research experience in the area of analytics and machine learning.

Prajwal Eachempati just finished his PhD degree.