Employee Retention Analysis Using Machine Learning

[1]Prof. Kemparaju N, [2]Ms. Latha JA, [3]Ms. Likitha M B, [4]Ms. Prerana Hiremath,
[5]Ms. Vaishnavi Ram J
[1,2,3,4,5]Department of Information Science and Engineering, East Point College of Engineering and Technology, Bengaluru, India.

Abstract- In the 21st century of information science and enormous data analytics, people analytics help organizations and their Human resources (HR) managers to reduce attrition by changing the way of attracting and retaining talent. during this context, employee attrition presents a critical problem and an unlimited risk for organizations because it affects. The continuity of their planning is as significant as their productivity. during this context, the salient contributions of those researches are as follows. Firstly, we propose land analytics approach to predict employee attrition. It focuses on the quality of knowledge instead of the amount of data, shifting from a big-data to a deep-data context. In fact, this deep data-driven approach relies on a mixed method to construct a relevant employee attrition model so as to identify key employee characteristics that contributes to attrition. during this method, we started thinking 'big' by collecting most of the common features from the literature (an exploratory research) then we tried thinking 'deep' by filtering and selecting the foremost critical features using survey and have selection algorithms (a quantitative method). Secondly, this attrition prediction approach relies on machine, deep and ensemble learning models and is experimented on a large-sized and a medium-sized simulated human resources datasets then generated a small low dataset from an entire of 450 responses. Our approach achieves higher accuracy (0.96, 0.98 and 0.99 respectively) for the three datasets as compared to previous solutions. Finally, while rewards and payments are generally considered because the sole keys to retention, our Findings indicate that 'business travel', which may be a smaller amount common within the literature, is that the leading motivator for workers and must be considered within HR policies to stay up them.

Keywords – Data Science, Big Data Analytics, Human Resource Managers, IEEE 802.15.4, Employee Retention Analysis, Attrition Prediction.

I. INTRODUCTION

Employee attrition or voluntary turnover is a key issue. This is a critical issue not only for organizations but also for the world at large. The sustainability of their work, but also their long term growth strategies. On this path, employee retention is a major challenge for recruiters and employers alike, since employee attrition. This includes not only the loss of skills, experiences, and personnel but also the loss of business opportunities. In the era of Big Data, people analytics help organizations and their human Human resource managers can reduce attrition by changing the, A way to attract and retain talent. Considered together with the surrounding words or circumstances, HR analytics is considered as a must have capability for the HR profession and a tool for creating it. The ability to predict employee departures can help an organization Streamline HR management and save money on it.

It is imperative for HR managers to understand A better understanding of the kinds of employees who leave and what reasonably features will influence them to depart. Most commonly, organizations desire to form sure the proper The employee is within the right place at the correct time and identifying employees' intention to go away by means of analytics. Descriptive analytics are accustomed summarize or turn data analyze relevant information so investigate what has occurred. In other words, descriptive analytics have some meaningful However, explaining what has already happened can have a bearing. they're not much helpful in predicting what's going to be possible that this can happen within the future.

On the contrary, predictive to forecast what is going to happen, analysts have proposed and used analytics happen within the future. within the field of HR, predictive analytics lead to organizational benefits and help surely in better decision-making within the organization with none

 DOI : 10.5281/zenodo.7697798

biasness. Data era and data science basing on machine and deep learning techniques. In fact, data is considered as one of the five essential ingredients of a people analytics team to be effective. Otherwise, HR is set to fail in handling large projects. Data challenges since Big Data focuses on capturing every piece of available information and collecting every suitable and unsuitable data. But, in HR analytics context, the issue must move from the size of the data to its smartness and making better use of data to create and capture value.

Big data is a prerequisite for more advanced forms analysis. Additionally, highlighted the limits of the The use of Big Data within a contextual HR case study whilst also noting the need to shift the focus from a quantitative to a qualitative analysis of HR data. In this context, the concept of deep data was born to deal with collecting only relevant and specific information and excluding information that might be unusable or otherwise redundant. Thus, in this paper, we mainly focus on two dimensions. There is a data dimension and a functional dimension. From a functional dimension, we aim to test, compare and select the highest quality accurate predictive model that predicts employee attrition in advance. We also aim to interpret the positive attrition to find reasons As a result, to support HR managers to build retention plan. From a data dimension, the key property of the proposed approach, we aim to shift from big data to deep data to address data issues that organizations may face when implementing HR analytics.

Big data may be a label commonly wont to identify large volumes of (structured or unstructured) data that may generally be defined with the assistance of the 3Vs volume, velocity, and variety. Volume refers to the amount of information that are produced by various sources like sensors, social media, business transactions, etc. Velocity represents the speed at which data are produced, and variety refers to the various formats of information. Over the last decade, the exploitation of huge data has become very hip among organizations and these ones tend to adopt new data-driven strategic decision-making recommendations to fight this possible attrition and to require necessary HR management policies.

II. RELATEDWORKS

Literature reports several employee attrition and voluntary yield predictive models. during this study, we particularly examine recent works that are supported machine and deep studying models applied to the simulated HR datasets of IBM and Kaggle. This choice is motivated through the existence of experiments and also the results of predictive models. Rightness for these open datasets so we will compare them with our proposed models.

IBM HR simulated dataset could be a medium sized-data block given by IBM and it contains 1470 samples with 34 input features (Age, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, legal status, Monthly Income, Monthly Rate, the amount Of Companies Worked, Over 18, Over Time, Percent Salary Hike, Performance Rating, Relationship, Satisfaction, Standard Hours, option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager) and it's target variable is turnover, represented as ''No''(employee didn't leave) or ''Yes'' (employee left). Kaggle HR dataset may be a large dataset provided by Kaggle that contains 15000 samples where its target variable is ''left'' and its 9 features are satisfaction level; last evaluation; number project; average monthly hours; time spend company; Work accident; promotion last 5 years; sales and Salary. In Table 1 authors present an overview of recent solutions to predict employee turnover. For each solution, used datasets and proposed models are presented such as Support vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) and K Nearest Neighbors (KNN). While these solutions put forward accurate predictive models to predict employee attrition, they suffer from two major critics:

The following problems are present: 1) There are no deep studies of employee features selected this is then used to predict attrition, thereby justifying the choice of features. 2) They generally focus only on the employee Attrition prediction is critical for a HR it is imperative to not only predict as soon as possible an employee's intention to leave but also to interpret and explain why the employee has this intention to leave.

References	Used simulated HR dataset	Proposed models
[14]	IBM	SVM, RF, KNN
[15]	IBM	LR, KNN, RF, XGB
[16]	IBM	DT, LR, KNN, RF, XGB, SVM, KNN
[17]	IBM	RF, LR
[18]	IBM	SVM
[19]	IBM	SVM, LR, RF
[20]	Kaggle	KNN, SVM, NB, DT, RF
[21]	Kaggle	DT, RF, SVM, MLP, KNN
[22]	IBM	LR, DT, SVM, Voting Classifier
[23]	IBM	LR, DT, KNN, RF
[24]	IBM	LR
[25]	IBM	DT, XGB
[26]	Kaggle	SVM, LR, RF. DT, XGB

Table 1. Recent Related Works

Data Set Overview

 DOI : 10.5281/zenodo.7697798

The databank includes all the employment history of 613 employees with a single employer. The details captured in the data are as follows:

- Employee ID
- Employee name
- Training count
- Date of birth
- Gender
- Marital status
- Engagement date
- Contract type
- Leave entitlement
- Grade
- Employment title
- Employee's department
- Last action type taken against an employee by the employer with respect to their
- employment
- Action date
- The employment status
- Degree level achieved by an employee
- Degree major for the employees

The above mentioned variables are important in helping an employer check the quantity of their human resources that is leaving, the factors contributing to it, forecasting future terminations, and estimating the validity of these estimates.

III. A MIXED METHOD FOR EMPLOYEE ATTRITION MODELING

As employee attrition or voluntary turnover could be a non-Modeling an unavoidable phenomenon may be a key issue for the procedure of attrition prediction. In addition, as we aim to embrace a deep data-driven approach, a search methodology that enables us to match theoretical models and experiments must be adopted. That's why we propose to conduct a mixed investigation method supported the mixture of an exploratory research and a quantitative method where the aim is to know and explain employee attrition phenomena. These two incorporate methods are used sequentially (e.g., findings from one method will tell the opposite. Thus, such a combined method can leverage the strengths and weaknesses of exploratory and quantitative methods and offer detailed insights on an event that every of those methods individually cannot suggests.

In fact, so as to realize a deeper understanding of the circumstance of high attrition and identifying the factors at the rear of it, an exploratory study supported reviewing available literature is firstly established thoroughly using studies, papers and open datasets provided by HR experts and researchers.

Secondly, these collected features are compared with causal factors for attrition identified through a questionnaire and have selection techniques (a quantitative research method).

The architecture of the conducted research methodology. we'll explain within the subsequent sections the various steps of the proposed mixed technique.

FEATURES COLLECTION: EXPLORATORY STUDY

The first step during this study was to spot and collect hired men features that are suitable for our analysis. This step is administrated using an exploratory research of things chargeable for turnover rate that relies on sec-Reviewing secondary literature for secondary research. Thus, the exploratory research method, outlined in the first step, assisted us recognize and collecting adequate and impactful characteristic for our troublesome that are most commonly used in different related works and researches in the available literature. In fact, through this exploratory study, based on

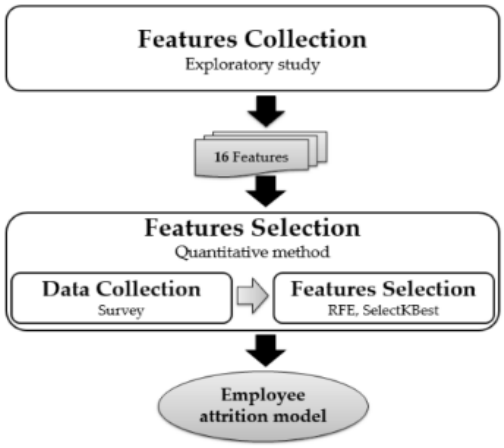

Figure 1. Our mixed method for employee retention modeling.

evaluate of the many researches, experiments in HR management and open simulated HR datasets (the fictional dataset created by IBM data scientists, and also the simulated HR databank supported by Kaggle) referenced in Table 1, we found that the strongest consensual predictors for employee. There are many factors that contribute to voluntary turnover: Age, Education, Gender, and responsibilities, etc. (inference of employee in decision making), Job Job satisfaction (career satisfaction), legal status, and performance at work (skills adequacy), Tenure, Promo ability (promotions In the workplace), Business Travel, Grade, Rewards (Pay, organization-based) based-rewards, Motivation factors, Salary), Relationship Satisfaction (Hostile organization culture), Environment Satisfaction (favorable or unfavorable working conditions), Training (Training time number, displeasing Work environment), Work life/balance. In table 2, we summarize these most cited 16 features that are commonly and regularly used in the available literature.

MACHINE LEARNING BASED PREDICTIVE MODELS

 DOI : 10.5281/zenodo.7697798

1**. Decision Tree** is built through a recursive partitioning procedure where paths from root to leaf represent classification regulation. Each internal node represents a ‘‘test’’ on an In terms of attributes, each branch represents the partitioned outcome of the test, and each leaf represents a class label in classification instance or a numerical value in a regression analysis.

2. **A support vector machine (SVM)** is a supervised learning algorithm. that is used for linear as well as nonlinear classification obstacles. To achieve class separation, it uses a hyper-plane or a set of hyper-planes in higher dimensional space. The intuition in this statistical learning based algorithm is that a suitable separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class.

3**. Logistic Regression** is a simple statistical technique and one of the basic linear models for classification that uses the logistic function to model categorical or binary depending on variables. It’s often used with regularization in the structure of penalties based on L1-norm or L2-norm to avoid over-fitting.

DEEP LEARNING BASED PREDICTIVE MODELS

1. **Deep Neural Networks (DNN),** are deep Artificial Neural Networks (ANN) with multiple (at least two) hidden layers where the ‘‘deep’’ refers to the number of hidden layers through which the data is transformed from the input to the output layers. In classical DNN, each layer is composed. The brain has a set of neurons and an activation function and is fully attached. A set of weights is applied to each neuron in the following way: each weight is multiplied by one input into the neuron. They are then summed to form the output from the neuron after it has been fed through the activation function.

2. **Long Short-Term Memory Networks (LSTM)** are an enhancement of recurrent neural networks (RNN) that are able to model sequential and temporal data and to predict times progression. More specifically, a cell state is added in LSTM to store long-term states and to build more stable RNN for time series prediction by detecting and memorizing the long- In the time series, there are term dependencies.

3. **Convolution Neural Networks (CNN)**, contain more often than not four types of layers in their structure: an input layer, complexity layers, pooling layers, and fully connected layer (output). In the convolution layer, which represents the most prominent feature of CNN is that the input will be convoluted with different filters where each filter is considered as a minimized matrix. Then, corresponding feature maps will be give rise to after the convolution operation. The pooling operation. The technique consists in reducing the size, while preserving the important features. The efficiency of the network is thus improved, and over-fitting is avoided. So, the main role of convolutional and pooling layers is generally to extract features, and the principal goal of fully connected layers is usually to output the knowledge from feature maps together, and then provide them to concluding layers.

ENSEMBLE LEARNING BASED PREDICTIVE MODELS

The main goal of ensemble learning (EL) is to amalgamate several models in order to find a better infusion that gives better results. So, EL is used here to unite the classifiers and their forecasts in order to improve robustness over a single classifier. In this study, we will test three ensemble learning models:

1. **Random Forest** is a popular tree-based ensemble learning approach and a bagging algorithm where successive trees are constructed using a different bootstrap sample of the dataset. By the end, a simple majority vote is taken for prediction. Random forests are different from standard trees as each node is split using the most accurate among a subset of predictors randomly chosen at that node which makes it robust against over-fitting .

2. **XGBoost** is a gradient boosted tree algorithm that involves fitting a set of weak learners and creating a final foretell is produced by the combination of predictions from all of them through a weighted majority vote (or sum). This boosting algorithm is based on the use of a regularized-Control over-fitting by formalizing models. It provides better presentation and is highly robust.

3. **Voting Classifier** is an ensemble learning model that tutors on an ensemble of classifiers and then predicts the harvests class construct on a majority vote according to two dissimilar strategies. The first one is the Hard Voting where the predicted output class is the class which had the highest possibility of being predicted by each of the classifiers.

The second one is the Soft Voting where the output entails: is the prediction based on the average of probability given to that class. In our case, we use a Voting classifier that unites our chosen ML models and that is based on the most of the vote strategy (difficult vote) to predict the output class. Such a classifier can be useful for a set of equally well This model performs well in order to balance out their individual fragility.

4. **Stacked ANN-based model** where outputs of the three Deep learning models (DNN, LSTM and CNN) are collected to create a new dataset encompassing also for each row. This is the expected value that will be used to train a new DNN learning model, called meta-learner. It is helpful to recall here that we used GridSeach for 10% the databank provides a validation set to identify the most appropriate hyperparameters for each model (such as decision criterion and max-depth for DT, the hidden layers number and units or neurons number in each layer for DNN, LSTM and CNN).

 DOI : 10.5281/zenodo.7697798

IV. RESULTSANDDISCUSSION

The random forest model orders the importance of the variables in predicting, as shown in figure below. The variables are arranged in ascending order of their importance. The most crucial variable in explaining retention is Training count, while gender is the least important.

	Active	Inactive	MeanDecreaseAccuracy	MeanDecreaseGini
Gender	2.33	-1.35	0.56	0.35
Grade	10.59	0.17	10.01	3.44
Department	11.99	-1.85	10.71	2.44
Leave.Entitlement	12.89	3.29	11.85	1.67
ActionYear	14.35	7.14	15.39	23.19
Training.Count	83.36	41.34	64.65	53.34

Table showing importance of variables in the prediction.

The section explains the research findings, specifically to the objectives and how they were met with regard to the quantitative and qualitative approaches. The primary purposes of the research included identifying the talents which the company wants to develop and maintain to predict employee training; identify the first-hand productive analysis techniques and implementation in a real-world scenario; highlighting who is eligible to be retained in the company and have successful career path; and identifying the employee retention rate in the organization under study. The research achieved these objectives through discussing the employee retention rate, percentage of gender leaving and remaining, strategies for leaving departments, termination by training count, termination by action year and Contract Type, and termination by degree major and degree level.

Based on the research objectives, and the survey conducted among the HR professionals for Company X, the research found out the pattern that associates the employee characteristics in general and employee retention. The machine-learning algorithm employed in previous research was also used to enhance accuracy. The dataset collected from the Random Forest analysis helped in predicting employee retention and provide the most accurate output. The study built the data model by partitioning the data by employing various algorithms to test and train data. Concerning the R models, they helped in providing excellent procedures and functions. These were essential in predicting the employees who might leave the organization in the future and predicted whether an employee is active or inactive in Company X. With the random forest, the research was able to provide a prediction of the employees who were at the risk of retention in the data set. The Random forest consisted of 500 trees with two variables tested at each split. The model employed a default setting of m tries in the model to obtain the smallest OOB error. The OOB estimate rate was 0.36%, while the area under the curve was 0.09927. Similarly, the Random Forest model established that the confidence level was 95%.

Concerning this, the higher the area under the curve, the better. The model also helped in highlighting the key variables that were essential in predicting employee retention in the company through decreasing order on the action year. These were employee training count, action year, employee grade, department, leave entitlement, and gender. The findings agree with those of Mitchell (2018), who found out that training was one of the factors leading to attrition. Other factors identified by Mitchell (2018) are low incentives, below expectation salary, and relationship with the superior, lack of appreciation, and unsatisfied work culture. Therefore, it is possible to evaluate the models to check the possibilities of predicting employee retention. The study also employed error matrices as a way of tabulating the outcomes of the research with the predicted values.

V. CONCLUSION

The company should strive to coach more employees since those that have attended more training are retained. Besides, other variables like ensuring employees are entitled to leaves and good salary grades, would make them feel motivated and remain within the company. The departments with reduced retention rates must be checked, and any dissatisfaction among employees cleared. The model highlighted that there are 79 active employees in Company X in 2019 and 1 inactive

employees. the info indicated that there's a high number of people who left Company X between 2006 and 2019. Company X could spend plenty of cash in training the staff, but doesn't gain from such investment thanks to the massive number of people who leave the corporate to work for other organizations. However, a keen look reveals that the majority employees who leave the corporate are people who have not attended any training. The HR department of Company X can utilize the obtained data to formulate a data-driven employee retention decision about the result of the model since the foremost critical variables for prediction are highlighted. The revelation that the majority trained employees remain with the corporate indicates that the retention strategies within the company are good. The RF model is applicable within the HR department to predict the general state within the HRM process, which include employee retention, hiring new employees, attrition, the number spent on training and developing new employees.

REFERENCES

[1] Masum, A.K., Beh, L.S., Azad, A.K., & Hoque, K., (2015). Intelligent human resource information system, 15(1), 121-130.

[2] Alao, D., Adeyemo, & A.B., (2013). Analyzing employee attrition using decision tree algorithms.

 DOI : 10.5281/zenodo.7697798

computing, information systems & development informatics, 4(1), 17-28.

[3] Alaskar, L., Crane, M., & Alduailij, M. (2019). Employee turnover prediction using machine learning. International Conference on Computing, 301-316.

[4] Crossman, A., & Zaki, B.A. (2003). Job satisfaction and employee performance of Lebanese banking staff. Journal of Managerial Psychology, 18(4), 368-376.

[5] Banerjee, A., Gosh. R. K., & Gosh, M. (2017). A study on the factors influencing the rate of attrition in it sector: based on indian scenario. Pacific Business Review International, 9(7), 1-13.

[6] Attridge, M. (2009). Measuring and managing employee work engagement: a review of research and business literature. Journal of Workplace Behavioral Health, 24(4), 383-398.

[7] Gent B, Coussement. K., & Poel D.V.D. (2006). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques.

[8] Bhartiya, N., Jannu, S., Shukla, P., & Chapaneri, R. (2019). Employee attrition prediction using classification models. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 1-6.

[9] Ribes E, Karim, T., & Benoit P. (2017). Employee turnover prediction and retention policies design: A case study, US: Cornell University.

[10] Luthans, F., Steven M.N. Bruce J.A., & James B.A. (2008). The mediating role of psychological capital in the supportive organizational climate: Employee performance relationship. Journal of Organizational Behavior, 29(2), 219-238.

[11] Kotsiantis, S.B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(1), 249-268.

[12] Kumar, A. A., & Mathimaran, K. B. (2017). Employee retention strategies: An empirical research. Global Journal of Management and Business Research, 17(1).

[13] Hoffman, M., & Tadelis, S. (2018). People management skills, employee attrition, and manager rewards: An Empirical Analysis. National Bureau of Economic Research.

[14] Nafeesa Begum, A., & Brindha, G. (2019). Emerging trends of it industry policies for ensuring women employee retention. Indian Journal of Public Health Research & Development, 10(3).

[15] Onsardi, A.M., & Abdullah, T. (2017). The effect of compensation, empowerment, and job satisfaction on employee loyalty. International Journal of Scientific Research and Management, 5(2), 7590-7599.

[16] Pradhan, R. K., Jena, L. K., & Pattnaik, R. (2017). Employee Retention Strategies in Service Industries: Opportunities and Challenges. Employees and Employers in Service Organizations, 53-70.

[17] Ramos, P.C. (2019). Employee retention strategies for executive operation leaders in an academic nursing environment, Walden University.

[18] Raschka, S, & Mirjalili V. (2017). Python machine learning. S.l. Packt Publishing Ltd.

[19] Ribes, E., & Touahri, K., & Perthame, B. (2017). Employee turnover prediction and retention policies design: a case study, Cornell University.

[20] Punnoose, P., & Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, 5(9), 22-26.

[21] Salunkhe, T.P. (2018). Improving employee retention by predicting employee attrition using machine learning techniques, Dublin Business School.

[22] Staniak, M., & Biecek, P. (2018). Explanations of model predictions with live and break down packages. arXiv preprint arXiv:1804.01955.

 DOI : 10.5281/zenodo.7697798