Fore Casting Employee Turn Over

P Sruthi[1], Naeemali Ahamed[2], Pavin shaji[3], Ashid A P[4], Prof. S Kavitha[5],
[1, 2, 3, 4, 5] Dept of Computer Science,
[1, 2, 3, 4, 5] Rajadhani Institute of Science and Technology, Palakkad

Abstract- *Supervised machine learning methods are described, demonstrated and assessed for the prediction of employee turnover within an organization. In our project, numerical experiments for real and simulated human resources datasets representing organizations of small-, medium- and large-sized employee populations are performed using random forest method and logistic regression method .Through a robust and compre- hensive evaluation process, the performance of each of these supervised machine learning methods for predicting employee turnover is analyzed and established using statistical methods. Additionally, reliable guidelines are provided on the selection, use and interpretation of these methods for the analysis of human resources data sets of varying size and complexity.Employee turnover has been identified as a key issue for organizations because of its adverse impact on work place productivity and long term growth strategies. To solve this problem, organizations use machine learning techniques to predict employee turnover. Accurate predictions enable organizations to take action for retention or succession planning of employees. However, the data for this modeling problem comes from HR Information Systems (HRIS); these are typically under-funded compared to the Information Systems of other domains in the organization which are directly related to its priorities. This leads to the prevalence of noise in the data that renders predictive models prone to over-fitting and hence inaccurate The main contribution of our project is to explore the application of Random forest and Logistic Regression technique which is more robust because of its regularization formulation.*

Keywords- machine learning, employee turnover, random forest, logistic regression, attrition rate

I. INTRODUCTION

Employee turnover is one of the most significant problems an organization can encounter throughout its life cycle, as it is difficult to predict and often introduces noticeable voids in an organization's skilled workforce . Service firms recognize that the timely delivery of their services can become compro- mised, overall firm productivity can decrease significantly and, consequently, customer loyalty can decline when employees leave unexpectedly. As a result, it is imperative that organizations formulate proper recruitment, acquisition and retention strategies and implement effective mechanisms to prevent and diminish employee turnover, while understanding its underly- ing, root causes. Most recently, the prevalence of intelligent machine learning algorithms in the field of computer science has led to the development of robust quantitative methods to derive insights from industry data. Supervised machine learn- ing methods—wherein computers learn from analyses of large-scale, historical, labelled datasets—have been shown to garner insights in various fields, like biology and medical sciences, transportation , political science , as well as many other fields. Owing to the advancements in information technology, researchers have also studied numerous machine learning approaches to improve the outcomes of human resource (HR) management. The performance evaluation of machine learning algorithms has also been studied previously by various re-searchers Notably, Punnoose and Ajit compared the predictive capabilities of seven different machine learning algorithms, including recently developed algorithms, like Extreme Gra-dient Boosting , on employee turnover. Similarly, Sikaroudi and co-researchers conducted simulations to predict employee turnover using ten different data mining algorithms, including tests on various types of neural networks and induction rule methods. In addition to placing focus on classification and prediction ability, many researchers have also made substantial efforts to better understand which features (e.g. compensation, age, work experience, etc.) are most influential in predicting employee turnover . These features seldom carry equal value in data mining applications, so it is useful to gain a better understanding of their importance . For instance, many of the studies using tree-based quantified feature importance by cal- culating the impurity reduction by node split in decision trees . Moreover, modified genetic algorithms and sensitivity analysis have been used to understand relative feature importance as well. Numerous studies have also generated classification rules or visualized the classification procedure to provide further insight and confidence in using machine learning methods . Despite the breadth of research outcomes mentioned above, the findings for predicting employee turnover that stem from using machine learning methods are often problem-specific and diffi- cult to generalize. First and foremost, this is primarily because HR data is confidential , which inherently impedes conducting in-depth analyses on multiple datasets. In addition, HR data is often noisy, inconsistent and contains missing information , a problem that is exacerbated by the small proportion of employee turnover that typically exists within a given set of

HR data. Secondly, gaps tend to persist in model performance evaluation. Specifically, previous research on the assessment of machine learning algorithms has generally focused on a narrow evaluation of metrics across various models

II. PROBLEM DEFINITION

Employee turnover is a serious problem for organizations since it can negatively affect their competitive edge, prof- itability, and productivity. Increased expenditures for hiring and training new staff as well as a loss of key information, skills, and experience can all be consequences of high turnover rates. So, forecasting employee turnover can aid businesses in creating tactics that will keep workers on board and cut down on the expenditures that come with high turnover rates. The goal of this study is to assess how well the machine learning algorithms Random Forest and Logistic Regression perform at predicting employee turnover. The study will use a dataset made up of various employee demographic, job-related, and performance-related factors to train and evaluate Random Forest and Logistic Regression models. The performance of the models will be evaluated using accuracy. The study's findings can help organization to identify the factors that affect employee churn and develop effective retention plans, which will increase productivity, profitability, and competitive advantages.

III. LITERATURE SURVEY

A literature review is a crucial step in the software development process since it highlights the many studies and analyses done in the area of interest, as well as any theoretical and methodological advancements made in that area. It is the most crucial section of the report because it directs your research and aids in establishing the objectives for the study. The reader is intended to understand what information and concepts have been established about a subject as well as its advantages and disadvantages.

A. Research on Employee Turnover Prediction Based On Machine Learning Algorithms

Human resources (HR) have received more attention in recent years as a result of the fact that skilled employees are a key driver of growth and a genuine competitive advantage for organizations. Artificial intelligence is starting to help HR management make decisions relating to employees after being introduced to the sales and marketing departments. The goal is to assist in making judgements that are based on objective data analysis rather than on subjective considerations. Analyzing how objective factors affect employee attrition is the aim of this effort. In order to determine the primary factors that influence a worker's choice to quit a company and to be able to foretell whether a certain employee will stay or go. After training, the model for predicting employee attrition is tested using an actual data set that IBM Analytics provided. This set has 35 attributes and roughly 1500 samples. Outcomes are given in terms of conventional metrics, and the Gaussian Nave Bayes classifier algorithm delivered the best outcomes for the supplied data set. Since it gauges a classifier's capacity to identify all instances of positivity and reaches an overall false negative rate of 4.5, it reveals the highest recall rate (0.54).

B. ARROW: A Web-Based Employee Turnover Analysis Tool for Effective Human Resource Management in Large Scale Organizations

To gain the competitive advantage, organizations need to adapt to the dynamic market. Therefore, many researchers have tried to find different ways for adapting to competitive conditions. Most of these research have finally ended up focus- ing on the human resource, which is the major and important resource in any organization. Currently human beings are treated as assets rather than resources. The System, ARROW is a unique web application developed to satisfy the requirements of company management in employee understanding process. The main objective of the system, ARROW is to fulfil the gap between employees' past, present and future behavior and the management's ability to understand the behavior of the organization's employees at the HR practices. Natural Language Processing and Data Mining techniques were used to accomplish the main objective. ARROW is a survey based employee analysis web tool for worldwide HR researchers and organizations, and it will facilitate the user to develop direct and indirect survey questions according to the user's insights. It will provide facilities to conduct the survey online by distributing the survey questionnaire online or manually by downloading and distributing it among the participants. When doing an online survey it will track answers of the participants who fill in the survey questionnaire. However, after gathering the data of the survey, ARROW will help to analyze results set by hypothesis testing module. It will provide a hypothetical analysis results and reports. This process is to help managers to understand employees' current behavior in the company.

C. Construction And Evaluation Of Employee Turnover Pre- diction Model

Employee turnover has been identified as a key issue for organizations because of its adverse impact on work place productivity and long term growth strategies. To solve this problem, organizations use machine learning techniques to predict employee turnover. Accurate predictions enable orga-

nizations to take action for retention or succession planning of employees. However, the data for this modeling problem comes from HR Information Systems (HRIS); these are typically under-funded compared to the Information Systems of other domains in the organization which are directly related to its priorities. This leads to the prevalence of noise in the data that renders predictive models prone to over-fitting and hence inaccurate. This is the key challenge that is the focus of this paper, and one that has not been addressed historically. The novel contribution of this paper is to explore the application of Extreme Gradient Boosting (XGBoost) technique which is more robust because of its regularization formulation. Data from the HRIS of a global retailer is used to compare XGBoost against six historically used supervised classifiers and demonstrate its significantly higher accuracy for predicting employee turnover. The problem of employee turnover has shot to prominence in organizations because of its negative impacts on issues ranging from work place morale and productivity, to disruptions in project continuity and to long term growth strategies. One way organizations deal with this problem is by predicting the risk of attrition of employees using machine learning techniques thus giving organizations leaders and Human Resources (HR) the foresight to take proactive action for retention or plan for succession. However, the machine learning techniques historically used to solve this problem fail to account for the noise in the data in most HR Information Systems (HRIS). Most organizations have not prioritized investments in efficient HRIS solutions that would capture an employee's data during his/her tenure. One of the major factors is the limited understanding of benefits and cost.

D. A Model Based Prediction of Desirable Applicants through Employee's Perception of Retention and Performance

One of the biggest issues in human resource management is choosing qualified candidates who have a lower inclination to leave their positions. The performance and productivity of an organisation are significantly impacted by proper pre- employment selection. Nonetheless, some companys are ex- periencing significant turnover rates as a result of the intense rivalry on the labour market. That happens when a candidate decides to leave and moves on to a company that provides alluring benefits and pay. This study aids decisions in reducing employee turnover by choosing suitable candidates who have a high likelihood of staying longer in an organisation. It assists human resource professionals in understanding psycho- logical climate. The researchers create a custom application that can assist organisations in making decisions regarding the employment process using the 12 retention dimensions utilised in the generation of association rules and naive bayes classifier. The method analysed seven retention characteristics with bad psychological climate responses, which led to a higher likelihood of people quitting their jobs willingly, based on the results of 12 retention dimensions. Also, the model that was created for the position of counter cashier revealed that applicants who are older than 20 and live outside of the workplace had a higher likelihood of staying long-term than other applicants. In order to generate association rules and conduct a classification analysis of the past and present employee profiles in an organisation, this study makes use of the PHP-ML Library. To guarantee the accuracy of the asso- ciation rule mining and naive bayes classifier implementation, the results were validated using the Rapid-Miner software. a company's previous and present employees' profiles can be examined using association rules and classification.

E. Survey on Employee Attrition Prediction

The most valuable component of human capital is the employee, whose performance reflects the success of the company. As a result of globalisation, workers have been mobilised from one organisation to another, from one region to another, and occasionally from one country to another. As a result, the biggest difficulty facing HR managers now is how to deal with attrition of skilled workers. Regarding the industry and its causes, the terms employee turnover and attrition are synonymous in HR practise. Attrition refers to the retirement of employees. either resign or die. Employee attrition is a severe problem that deals with the forced or voluntary termination of talented and skilled workers, which has an impact on the reputation and productivity of the company. Employee attrition is a sign that workers are quitting due to either personal or professional issues. Most studies have revealed that a greater employee attrition rate is mostly caused by factors relating to the employee's job. High attrition results in a loss of the company's investment in hiring and training new employees. Employee attrition has a long-lasting detrimental effect on an organization's reputation. In plain English, it may be noted that failure to meet an employee's view or expectation of an employer or a failure on the part of the employer to fulfil that commitment are the main causes of employee turnover. The most valuable assets in a company are its employees. They benefit the company in terms of both number and quality. Hence, maintaining a stable and promising staff is essential. Employers now have a challenging task, which has led to an increase in organisational attrition over time. In any form of company, a small amount of employee attrition is desirable for the influx of fresh concepts. In a rapidly changing world, it aids firms in maintaining their agility.

IV. EXISTING SYSTEM

In various situations, the current research have made an effort to forecast employee turnover. The bulk of studies have chosen a qualitative approach to the issue of employee turnover. Yet, with the development of cutting-edge machine learning approaches for predicting employee turnover, it is now necessary to look into how these techniques may be used to their full potential to not only predict employee turnover but also to discover how much each element influences it.

1) First, predictions are made using the machine learning techniques mentioned above. A strong statistical model, such as multiple linear regression, or a multi-stakeholder multi-criteria model, which takes into account several elements that interact in employees' brains before mak- ing a decision, do not, however, validate the feature importance that emerges. The results are not dynamic and are not applicable in a real situation because the employees' unique tastes and the value they place on various aspects are not thus validated.
2) However, current research does not assess or priori- tise the relative importance of the elements affecting employee attrition rate, which would provide incentive methods. To forecast turnover and examine the crucial aspects influencing turnover status, no hybrid model is used.
3) Lastly, none of the aforementioned research have taken into account the crucial elements. It is possible to improve employee satisfaction and retention by un- derstanding the primary elements affecting employee morale and making tailored recommendations. If, for example, an employee is expected to leave the company owing to a low "Employee CTC level" and a heavy workload (measured by the "Number of projects/tasks allocated every quarter"), rewards can be given to the worker and his burden can be divided more fairly, raising "Employee Satisfaction."

V. PROPOSED SYSTEM

Methods for supervised machine learning that can fore- cast employee turnover inside an organisation are discussed, illustrated, and evaluated. In this study, the Random Forest Method and Logistic Regression Method are used to conduct numerical experiments for real and simulated human resources data sets representing firms with employee populations. The effectiveness of each of these supervised machine learning algorithms for forecasting employee turnover is studied and established utilising statistical methods through a thorough and rigorous evaluation procedure. Furthermore, trustworthy advice on the selection, application, and interpretation of these methodologies are offered for the analysis of human resources data sets of various sizes and complexity. A general approach for a project to forecast staff turnover using Random Forest and Logistic Regression is gather information on employee turnover, Pre-process and clean the data. Taking care of missing data, scaling and normalising the features, encoding categorical variables, and dividing the data into training and testing sets are all included in this process. Determine the key elements that most significantly affect employee churn. Depending on the features of the data, select the suitable algorithm. In this situation, the algorithms Random Forest and Logistic Regression are appropriate for classification issues. Use the training data set to train the chosen models. Analyze the models' performance using the test data. To obtain un- derstanding of the elements that lead to employee turnover, interpret the model results.

A. RANDOM FOREST

By integrating a number of weak learners into a stronger learner, random forests use an ensemble strategy that outper- forms the basic decision tree structure. Divide and conquer tactics are used by ensemble methods to boost algorithm performance. On the basis of bootstrapped training sets, a number of decision trees, or weak learners, are constructed in random forests. For each decision tree, a random sample of m predictors are selected as split candidates from the whole set of P predictors. In this study, the employee HR data collection is initially loaded into a Pandas DataFrame. The train test split function from Scikit-Learn was then used to divide the data set into training and testing sets. Then, with 100 trees and a random state of 42, we generate a RandomForestClassifier object. The model was then fitted using the fit method on the training set. Following that, we use the predict method to generate predictions on the test set and the accuracy score function to determine the model's accuracy. Lastly, we print out the model's accuracy.

B. LOGISTIC REGRESSION

The standard classification procedure known as logistic regression, which Cox first described in 1958, uses linear discriminants. A probability that the specified input point belongs to a particular class is the main output. The model establishes a linear border dividing the input space into two sections based on the probability value. One of the most popular classifiers is logistic regression since it is simple to apply and performs well on classes that can be divided into linearly distinct subsets. The employee HR dataset is initially loaded into a Pandas DataFrame. The train test split function from Scikit-Learn was then used to divide the dataset into training and testing sets. A Logistic Regression object with a

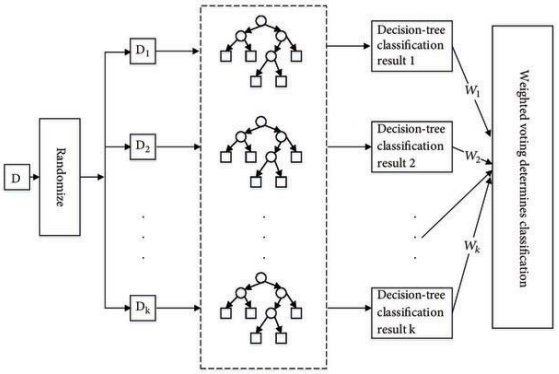

Fig.1. working of Random Forest

random state of 42 is then created. The model was then fitted using the fit method on the training set. Following that, we use the predict method to generate predictions on the test set and the accuracy score function to determine the model's accuracy. Lastly, we print out the model's accuracy.

VI. WORKING

A. Data Collection

The HR data that will be used in this study was obtained from the Kaggle website. It is a data set of several employee characteristics that is examined yearly and ultimately aids in forecasting turnover. It has 15000 records and 10 different qualities.

B. Data Description

The dataset's properties, descriptions, and potential values are listed below.

1) The following list shows the attributes of the dataset, description and possible values.
2) Employer evaluation of employee performance (0-1).
3) number of projects the person has finished.
4) The average number of hours a worker works each month.
5) Employee tenure as measured by the length of time they have worked there.
6) Any workplace accident either involved an employee or not.
7) Whether the person received a promotion
8) Which department does the individual work for?
9) Pay of the employee (Low, Medium and High)
10) Has the employee left the company?

C. Data Preprocessing

Employee turnover prediction studies frequently involve data pre-processing since the data sets frequently include missing entries, varied levels of noise, and significant variances in magnitude for each component. To provide results that were as meaningful as possible, the following data pre processing methods were employed:

1) Imputation of Missing Values: To ensure that all algorithms could handle missing values, they were imputed. Yet, other algorithms, like logistic regression, could deal with miss- ing variables automatically without imputation. The missing values were imputed using the data type of the missing values in order to limit the comparison complexity. The median value of the entire elements is used to replace the missing entries for numerical data types. For categorical data, the mode value of the entire items was used to replace any missing entries.
2) Conversion of Data Types and Feature Selection: Converting categorical variables to numerical representation is a crucial step in the preparation of data. Categorical variables cannot be used directly by some algorithms, including logistic regression, neural networks, and K-nearest neighbour. Conse- quently, if a categorical feature has a wide range of different values, this conversion may greatly expand the dimensions of the feature. In this study, label encoding was used to convert data using Python's Scikit-learn package. By choosing perti- nent attributes, the feature selection methods are frequently utilised to further enhance the classifier's predicting skills. If the data dimensionality is high, dimensionality reduction techniques like principal component analysis are also used. No feature selection nor dimensionality reduction were utilised in an effort to limit the complexity of the results analysis and the interpretation of HR data that was required.
3) Scaling Features: feature scaling is used to narrow the range of characteristics and harmonise different feature scales. scaling the inputs is advised to provide good outcomes while using random forest and logistic regression .

D. Analyzing Exploratory Data

Starting with the descriptive analysis ,here correlation matrix is used to determine the correlation coefficient between the attributes. There are two sorts of employees: those who left and those who stayed. By grouping the data according to the left attribute, we got to the conclusion that those who left had lower levels of satisfaction, lower rates of promotion, lower salaries, and also put in more hours than those who stayed.We have utilised a heat map to show the correlation matrix in order to determine the correlation coefficient between two variables. The creation of numerous graphs and data exploration were the next steps in the exploratory analysis

process. We began by conducting an analysis on a single variable, or a uni variate. One of the most fundamental methods of data analysis is the univariate approach.

VII. CONCLUSION

The issue of employee turnover is addressed in this study. This study includes estimates of the time frame and likelihood of leaving in addition to a prediction of whether an employee will leave the company.The corporation can take action by dealing with salary concerns and other things to keep these potentially leaving personnel. An organisation may fail if its personnel turnover rate is too high. Administrative staff members might be unaware of the reasons behind this turnover. In conclusion, it is now more crucial than ever for businesses all over the world to predict employee turnover using machine

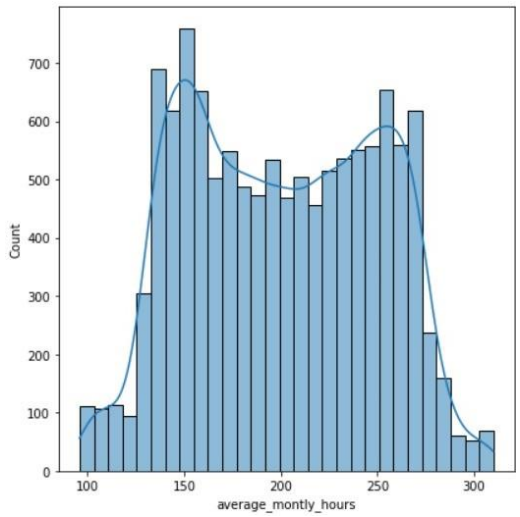

Fig. 2. Average monthly working hours

learning approaches. The most important characteristics and factors that influence employee turnover have been successfully identified using machine learning approaches like random forest and logistic regression. Organizations can take specific efforts to enhance staff retention and lower turnover by recognising these characteristics and causes. It is crucial to remember that the selection of technique depends on the particular situation and data set. Before choosing which method to utilise for predicting employee turnover, organisations should assess the effectiveness of both random forest and logistic regression. To get the most accurate projections, businesses should also make sure the data they utilise is reliable, thorough, and current. In general, applying machine learning algorithms to forecast employee turnover is a useful tool for businesses to efficiently manage their human resources. To improve the precision of employee turnover forecasts, future research can investigate other machine learning techniques or integrate various methodologies. Another potential area for research is examining the efficacy of staff retention initiatives in light of the projections.

VIII. RESULTS

The algorithms used were fruitful as they brought about high accuracy rates in their predictions. Using the Logistic Regression algorithm, we obtained an accuracy of 82.9%. The Random Forest algorithm provided an accuracy of 98.08%.

REFERENCES

[1] Al-Radaideh, Q.A., Al Nagi, E.: Using data mining techniques to build a classification model for predicting employees performance. Int. J. Adv. Comput. Sci. Appl. 3, 144–151

[2] Nagadevara, V., Srinivasan, V., Valk, R.: Establishing a link between employee turnover and withdrawal behaviours: application of data mining techniques. Res. Pract. Hum. Resour. Manag. 16, 81–97

[3] RohitPunnoose ”Prediction of Employee Turnover in Organizations using Machine Learning Algorithms A case for Extreme Gradient Boosting”, PhD dissertation Ph.d candidate XLRI – Xavier School of Management Jamshedpur, India Pankaj Ajit

[4] B. Holtom, T. Mitchell, T. Lee, and M. Eberly, “Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future”, Academy of Management Annals, 2: 231-274

[5] N.Silpa , “Study on Reasons of Attrition and Strategies for Employee Re- tention”, Annamacharya P.G college of Management Studies, Rajampet, Andhra Pradesh, India, December 2

[6] A. Liaw and M. Wiener, “Classification and regression by randomForest”, R news, 2(3)

[7] N.Silpa , “Study on Reasons of Attrition and Strategies for Employee Re- tention”, Annamacharya P.G college of Management Studies, Rajampet, Andhra Pradesh, India, December 2.

[8] B. Holtom, T. Mitchell, T. Lee, and M. Eberly, “Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future”, Academy of Management Annal. 9