People Leaving the Job – An Approach for Prediction Using Machine Learning
Ishaan Ballal¹, Shlok Kavathekar², Shubham Janwe³, Pratik Shete⁴, Prof. Nivedita Bhirud⁵
¹²³(Students, Computer Engineering Department, Vishwakarma Institute of Information Technology, Pune)
⁴⁵(Professor, Computer Engineering Department, Vishwakarma Institute of Information Technology, Pune)

Abstract: Organizations face huge costs resulting from employee turnover. It may incur a high cost such as training expenses and the time it takes from when an employee starts to when they become a productive member. When a productive employee quits, the organization loses new product ideas, great project management, or customer relationships. So there is a need for the system that understands the key variables that influence the employee attrition rate using data mining.

This paper discusses different approaches used by the researchers for the prediction of employee attrition rate in addition to an abstract view of the system that we are going to implement for Employee Attrition Rate Prediction using data mining.

Keywords—Data Mining, Employee Attrition Rate Prediction, Identifying Reason, Employee Turnover.

I. INTRODUCTION

Information technology (IT) industry in India has been facing a universal issue of high attrition in the past few years in HR Analytics [1][6], resulting in monetary and knowledge-based losses to the companies. HR Analytics is the need of the company so that they can spend the money on the right employees rather than spending on the wrong ones. With the help of HR Analytics, HR Managers can take numerous decisions on investment in employees to get excellent outcomes that benefit the stakeholders and customers. Human Resource departments of the organizations can predict the employee attrition [1][3][4] and the reason for an employee leaving an organization.

Many factors lead to dissatisfaction in an employee, like long working hours, peer pressure, job location, job role, traveling time, office space, amenities in the office, perks, and many more reasons could be a factor for employee attrition.

Every year a lot of companies hire several employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. These programs aim to increase the effectiveness of their employees.

Sometimes the employee may not have any problem in the company but others may offer a better profile with a better pay package. So, the employee may be willing to leave. So the HR department needs to understand the employee satisfaction level [6].

II. LITERATURE SURVEY

Rohit Punnoose et al.[2] explore the application of Extreme Gradient Boosting (XGBoost) [3][5] technique which is more robust for prediction of Employee turnover [2][8] because of its regularization formulation. The data from HR Information Systems (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.

Jeel Sukhadiya et al.[3] presented a detailed comparison and analysis of different methodologies used for forecasting employee attrition rates. The methods such as Random Forest, Support Vector Machines (SVM), Gradient Boosted Classifier and Logistic Regression is used for predicting and classifying the attrition is a fictional dataset created by IBM data scientists.


Yue Zhao et al.[5] demonstrated and assessed Supervised machine learning methods for the prediction of employee turnover within an organization. The experiments are performed on simulated human resources datasets using a decision tree method; a random forest method; a gradient boosting trees method; an extreme gradient boosting method; a logistic regression method; support vector machines; neural networks; linear discriminant analysis; a Naïve Bayes method; and a K-nearest neighbour method.

Dilip Singh Sisodia et al.[6] build a model that will predict employee churn rate based on the HR analytics dataset obtained from the Kaggle website. The author generated a heat map to show the relation between attributes, and the correlation matrix. In the experimental part, the histogram is generated, which shows the contrast between left employees vs. salary, department, satisfaction level, etc. For prediction purposes, five different machine learning algorithms such as linear support vector machine, C 5.0 Decision Tree classifier, Random Forest, k-nearest neighbor, and Naïve Bayes classifier are used. This paper proposes the reasons which optimize employee attrition in any organization.

Heng Zhang et al.[7] adopt the supervised learning technique for employee turnover prediction. The approach from [8] helps human resource personnel to understand psychological climate and supports decisions in resolving employment turnover by selecting desirable applicants who have a high probability of staying longer in an organization. Using the 12 dimensions of retention applied in generating association rules and naïve Bayes classifiers, the researchers develop a custom application that can support the organization’s decision in relation to the hiring process.

Amir Mohammad Esmaieeli Sikaroudi, et al, [9] implemented knowledge discovery steps on real data of a manufacturing plant. They chew over many characteristics of employees such as age, technical skills, and work experience. They used to find out the importance of data features is measured by the Pearson Chi-Square test.

John M. Kirimi and Christopher Moturi et al, [10] proposed infrastructure for employee performance forecasting that enables the human resource professionals to refocus on human capability criteria and thereby enhance the performance appraisal process of its human capital.
Ancheta et al.[12] make use of rule-based classification Data Mining technique to extract knowledge significant for predicting training needs of newly-hired faculty members in order to devise the necessary development programs. They used the Cross-Industry Standard Process for Data Mining (CRISP-DM) is discovering significant models needed for predictive analysis and demonstrated the required professional training to prepare faculty members to perform their tasks effectively

A. Data mining Technique Used for Prediction of Employee Rate

a. Naïve Bayes Algorithm [4][5][6][8]

Naïve’s Bayes is a probability-based classification theorem which is based on the Bayes Theorem. It can be used to predict the outcome of an occurring event with independent conditions. Naïve Bayes is an eager classifier and fast executing algorithm used for real-time predictions and also has a higher success rate compared to other algorithms.

b. Tree Learning Algorithm [4][5]

The decision tree [6] is a conventional algorithm used for performing classifications based on the decisions made in one stage. This provides a tree-structured representation of the decision sets. This classification based on the decision tree let us predict the qualitative response without creating the dummy variables also the class proportion among the training region which falls into a particular region can be predicted using this algorithm.


This is a regression analysis that is used to set a statistical process for estimating the relationship between the dependent variable and one or more independent variables. We can analyze how the dependent variable is affected when one of the independent variables is changed and by fixing the other independent variables. This technique is used to predict the qualitative response.

d. Extreme Gradient Boosting (XGB)[3][5]

Extreme Gradient Boosting is a tree-based method that was introduced in 2014 by Chen. It is also commonly referred to as XGBoost. It is a scalable and accurate implementation of gradient boosted trees, explicitly designed for optimizing the computational speed and model performance. Compared to gradient boosting, XGBoost utilizes a regularization term to reduce the overfitting effect, yielding a better prediction and much faster computational run times.

e. Random Forest [6]

Random Forest, it is one of the ensemble learning technique which consists of several decision trees rather than a single decision tree for classification. While classifying all the trees in the random forest gives a class to an unknown example and the class having maximum votes will be assigned to the unknown example.

f. Support Vector Machine [3][5]

A Support Vector Machine is a kind of classification technique, where the data points are separated by a line in case of linear SVM, and a hyperplane in case of non-linear SVM. The separation is chosen in such a way that; the two sides of the hyperplane categorize the data set into two classes. When an unknown data comes it predicts which side/class it belongs to. The margin between the hyperplane and the support vectors is as large as possible to reduce the error in classification. Support Vector Machine.

III. PROPOSED SYSTEM

We are developing a system that helps to predict employee attrition rate and understand the key variables that influence the employee attrition rate using data mining.

Figure 1: System Architecture

Here we are using random forest algorithm to build a prediction model for identifying the various reason for employee turnover. The data mining algorithm is performing well in predicting the employees; those are likely to quit the respective organization based on their working details and environments.

IV. CONCLUSION

The growth level, as well as market perception, mainly depends on the strength of the employees. Now a day due to increased population and people with high capacity makes great success for any firm. But the prime issues which are normally addressed in any organization are only the attrition.

This paper finds a different data mining algorithm used for performing well in predicting the employees, those are likely to quit the respective organization. The paper contains the survey of various classification algorithms like logistic regression, LDA, SVM, KNN, Random Forests to predict the probability of attrition of any new employee.
REFERENCES


