International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Implementation of Machine Learning Model for Employee Retention Prediction

Gaurav Thakre[1], Prajot Wankhede[1], Sargam Patle[1], Shrikant Joshi[1], Prof. Alok Chauhan[2]

[1]BE Student, Department of Information Technology, Rajiv Gandhi College of Engineering & Research, Nagpur, Maharashtra, India

[2]Assistant Professor, Department of Information Technology, Rajiv Gandhi College of Engineering & Research, Nagpur, Maharashtra, India

Article Info
Volume 8, Issue 3
Page Number : 503-508

Publication Issue
May-June-2021

Article History
Accepted : 15 May 2021
Published : 30 May 2021

ABSTRACT

Employees are regarded as the organization's backbone. The personnel who work for it determine the success or failure of a company. When trained, competent, and experienced personnel depart for greater opportunities, businesses must deal with the resulting challenges. The purpose of the research was to determine the level of employee unhappiness and the reasons why they would choose to change jobs. Once the dissatisfaction factor(s) of employees has/have been discovered, businesses can take appropriate action, which may aid in lowering the turnover rate. In this research, we attempt to develop a system that can forecast employee attrition using data from the Kaggle website's Employee dataset. To visualise the relationships between the attributes, we created a heatmap. We employed four different machine learning methods for prediction, including KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Decision Tree, and Random Forest. The reasons for employee attrition in any firm are discussed in this study.

Keywords : Attrition Rate, HR, Classifier, Preprocessing, Employment Features

I. INTRODUCTION

Employee attrition refers to the loss of manpower in any organisation as a result of employees leaving [1]. Any organization's most precious asset is its employees. It is vital to determine whether or not the employees are unsatisfied with their jobs or if there are any other reasons for their departure. Employees are keen to jump from one business to another these days in search of better chances. However, if they leave their jobs unexpectedly, the organisation may suffer a significant loss. New hiring costs money and time, and it takes time for newly employed people to contribute to the profitability of the company. One of the most difficult issues that many firms confront is retaining qualified and diligent staff. As a result, we

may greatly lessen this problem by increasing employee satisfaction and providing a desired working environment [1]. Employees leave organisations for a variety of reasons. Some of these reasons include a better-paying job outside the organisation, a bad relationship with the boss, pursuing higher education, relocating for family reasons, being fired from the organisation, and being dismissed from the company. Dissatisfaction, pay that isn't up to par, strained relationships with coworkers, a terrible working environment, a lack of opportunities for advancement, overtime, and a heavy task, to name a few. To address this problem, we created a system that analyses reasons for employee churn using data from employees. Employees who have finished their probation period are eligible to apply for this position. Because they are not a confirmed employee before their assigned probation time, it is difficult to foresee their discontent elements if the employee has recently joined the organisation.

This method can forecast which people are likely to depart a business and for what reason, allowing managers to take a variety of remedial actions to guarantee that people stay and attrition is reduced. Motivating employees, exposing employees to fresh responsibilities, getting frequent input from employees, and so on are some of the employee retention tactics to control attrition. SVM (Support Vector Machine), KNN (K-Nearest Neighbor), Decision Tree, and Random Forest were some of the machine learning methods we used. For a better comprehension of the insights, a graphical representation is also supplied.

II. LITERATURE REVIEW

Nagadevara et al. (2008) evaluated the impact of withdrawal practises such as tardiness and non-attendance, job content, residency, and socio-economics on worker turnover in the Indian software industry, which is still in its early stages. The use of five prescient data mining procedures (artificial neural networks, logistic regression, classification and regression trees, classification trees (C5.0), and discriminant analysis) on a sample data of 150 employees in a large software organisation was the most remarkable aspect of this study. The findings of the study show that there is a link between withdrawal behaviours and employee turnover. The findings of this study generated a few questions that should be investigated further in the future. To begin, more study may explicitly collect data on statistic components across a large sample of businesses in order to investigate the relationship between statistic factors and turnover. Second, large-scale data on characteristics associated with turnover from previous academic studies can be gathered [1].

The researchers investigated the elements impacting worker voluntary turnover in the North American former professional sales force of a Fortune 500 advanced manufacturing corporation in an exposition by Marjorie Laura Kane-Sellers (2007). The goal of investigating Voluntary Turnover was to have a greater comprehension of Human Resource Development strategies that could help employees stay longer. Over a 14-year longitudinal period, the central firm provided perceptions of the worker database for certain individuals from the expert specialised deals power. The first database had 21,271 discrete observations, each of which was identified by an unique company clock number [2]. Ibrahim et al. advocated constructing models with various processes, such as categorization for prediction, clustering for detection, and association for detection, to address a key issue of customer turnover identified with a firm, notably telecoms [3].

Choudhary et al. discuss the use of a logistic regression method based on employee data to construct a risk equation for predicting employee attrition. Later, this equation was used to determine the risk of attrition with the existing group of employees. Following the estimation, a high-risk cluster was identified in order to determine the causes, and an action plan was chosen to reduce the risk [4].

M. Singh et al. from IBM Watson conducted an investigation of staff attrition procedures and offered a structure for determining the reason for attrition as well as identifying probable attrition. They compared the difference between Expected Cost of Attrition Before the Retention Period (EACB) and Expected Cost of Attrition After the Retention Period (EACA) in order to calculate the cost of attrition and propose the employee's name for the retention process [5]. To predict employee performance, the authors of Q. A. Al-Radaideh et al. used decision trees (ID3 C4.5) and the Nave Bayes classifier. They discovered that job title was the most important factor, while age had no discernible effect [6].

Employees are more likely to depart according to O.Ali et al., possibly due to a disagreement with their superior officer. We noticed a number of significant factors influencing the company's worker turnover. From the two rules, he derives in a modest way. Some questions were asked of both parties, and he came to some conclusions based on their responses[7], such as workload, goals, and carrier. Human resource management, according to A. frederiksn et al., concentrates on termination rates and firing rates in general, although their real substance is radically different. According to the prior paradigm, there are various different degrees of recruiting and turnover. The dismissal and termination rates, according to certain research [8,] have institutional repercussions. Allen and Meyer (1990) defined the three-basic entity for the negative side of turnover in H.Ongori et al. A regulating officer is more likely to depart the company because of a disagreement with higher management than a delegate dealing with his immediate employer. He recognised the guiding principles that influence the organization's willingness to absorb workers without opposition [9]. Two systems of methodologies for social opportunity data are guided by V.V. Saradhi et al. An equal number of members and officers were invited to participate in a series of polls that were organised by workload, priorities, personality, professional success, and hierarchical management. The results of the two ways of data collection revealed that the most significant factor of employee rejection is monetary compensation [10].

Hossein et al. employed the CRISP-DM data mining methodology throughout their investigation. The main data mining tool utilised to build the model of classification where multiple rules of categorization were constructed was the decision tree. The model that had been created had been tested and found to be accurate. Several experiments were conducted utilising real data acquired from a variety of sources. The goal of the model is to forecast the quality of new applicants[11]. Amir Mohammad and colleagues used real-world data from a manufacturing to construct knowledge discovery stages. Many personnel qualities, such as age, technical capabilities, and job experience, are well-understood by them. They looked studied the relevance of software features using the Pearson Chi-Square test [12]. Rohit Punnoose and Pankaj Ajit et al investigated the application of the approach of Extreme Gradient Boosting (XGBoost), which is more dependable due to its formulation of regularisation. The HRIS data from a worldwide store is used to compare XGBoost to six commonly used supervised classifiers in order to show that it has a much higher reliability in predicting employee turnover [13]. Churn prediction, particularly customer churn prediction, has gotten a lot of interest from researchers. Verbeke et al., for example, propose a profit-centric measure of performance by calculating the greatest profit that can be generated in a retention effort by integrating the ideal customer percentage with the highest estimated likelihood of churning[14]. Coussement and Van den Poel investigated the topic of improving the quality of a decision support system for churn prediction. They looked at the impact of textual data on the churn prediction process. They discovered that incorporating unstructured, textual data into a traditional churn prediction algorithm improved predictive performance significantly [15]. Coussement and Van den Poel employ the SVM

approach to forecast client churn. Supporting vector machines produce good generalisation performance when applied to noisy marketing data, according to their findings[17].

Burez and Van den Poel are studying customer churn and class imbalances in order to forecast customer churn. Undersampling can improve prediction accuracy, according to the findings of a study [18].

In another study, Tsai and Chen used association rules to select important features before applying neural networks and Decision Tree to predict customer churn in a telecommunications company. To examine their results, they utilise four performance measurements that are identical to ours: accuracy, accuracy, recall, and F-measurement. Cushioning and colleagues [19]. Construct the Generalized Additive Models (GAM) approach, which allows the model to fit non-linear data that is complex. Other studies have used well-known data mining approaches to assess customer churn [20]. In comparison, there are only a few studies in the literature that look at employee turnover estimation and research. Saradhi and Palshikar investigate employee churn prediction using naive Bayes, Logistic Regression, Decision Trees, and Random Forest approaches [21]. To the best of our knowledge, the most recent research is a report by KaneSellers on the database of a Fortune 500 North American industrial automation manufacturer's trained sales team. The method of logistic regression [22] is Kane-Sellers' primary method.

III.IMPLEMENTATION

This use case takes HR data and applies machine learning models to forecast which employees are more likely to resign based on certain characteristics. An organisation may use such a model to predict employee turnover and develop a strategy to mitigate such a costly problem. An Excel file containing information about 1470 employees serves as the input dataset. There are traits / features for each employee, such as age, employee role, daily rate, job happiness, years at the company, years in present function, and so on, in addition to whether the person departed or not (attrition).

The steps we will go through are:

- Data preprocessing
- Data analysis
- Model training
- Model validation
- Model predictions
- Visualization of results

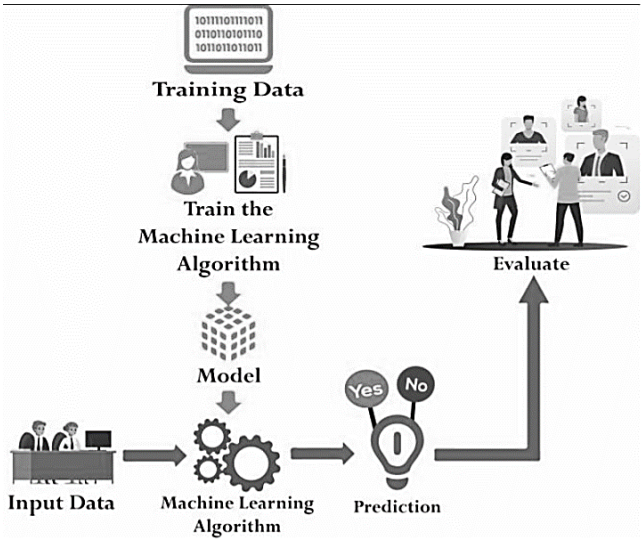

Figure 1. System Architecture

A variety of machine learning algorithms are used in the proposed system. To create the model, we start with the employee dataset, which contains all of the employees' past and present records, and then we perform data preparation (Data Preprocessing is that step in which the data gets transformed, or encoded, to bring it to such a state that the machine can easily analyse it). The dataset has been separated into two parts: training data and testing data. The majority of the data is used for training, whereas only a tiny percentage is used for testing (Train: 70 percent , Test: 30 percent). The goal of training is to get as close to 100% accuracy in predictions as possible. The test data is used to determine how effectively the machine can anticipate new answers and to validate the behaviour of the machine learning model.

After that, we built the model using a variety of machine learning algorithms. Following the creation of the model, the user can provide the system with new input data. Furthermore, the user has the option of selecting an algorithm and checking the results. The system produces two types of output: one is a graphical representation, and the other is a polar form, such as a 'Yes' or 'No' format. Following the evaluation of the results, the system also provides the cause for attrition.

IV. RESULT AND EVALUATION

Various factors, such as department, gender, overtime, business travel, and so on, are included in the given dataset. Based on these variables, a model was created using a variety of machine learning techniques to predict whether or not employees will leave the company. To calculate the accuracy of each method, the predicted values are compared to test values. The table below summarizes numerous aspects so that we may quickly determine which algorithm is suitable for our situation. We may deduce from the table that Random Forest has the maximum accuracy on the HR Employee-Attrition dataset, whereas Decision Tree has the lowest accuracy.

Table 1. Results of Different Classifier

Attributes/Model	KNN	SVM	Decision Tree	Random Forest
Accuracy	0.8639	0.8684	0.8163	0.8843
Precision	0.8196	0.8364	0.8224	0.8723
Sensitivity or Recall or True Positive Rate	0.8621	0.8697	0.8165	0.8852
F-Measure	0.8403	0.8520	0.8193	0.8786
Specificity or True Negative Rate	0.9921	0.9790	0.8897	0.9895
False Positive Rate	0.0079	0.021	0.1103	0.0105
False Negative Rate	0.1379	0.1303	0.1835	0.1148

V. CONCLUSIONS

This research investigates which machine learning method is the most effective at predicting which employees are most likely to depart a company. As a result of the results, we may conclude that Random Forest outperforms the other classifiers. It has been discovered that both external and internal variables contribute to employee attrition. This study may aid organisations in understanding the elements that contribute to employee attrition so that appropriate action can be taken to reduce attrition rates.

VI. REFERENCES

[1] Nagadevara, Vishnuprasad, Vasanthi Srinivasan, and Reimara Valk. "Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques." Research & Practice in Human Resource Management 16.2 (2008).

[2] Kane-Sellers, Marjorie Laura. Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis. Texas A&M University, 2007.

[3] Mitkees, Ibrahim MM, Sherif M. Badr, and Ahmed Ibrahim Bahgat ElSeddawy. "Customer churn prediction model using data mining techniques." 2017 13th International Computer Engineering Conference (ICENCO). IEEE, 2017.

[4] Khare, Rupesh, et al. "Employee attrition risk assessment using logistic regression analysis." IIMA International Conference on Advanced Data Analytics, Business Analytics. 2015.

[5] Singh, Moninder, et al. "An analytics approach for proactively combating voluntary attrition of employees." 2012 IEEE 12th International Conference on Data Mining Workshops. IEEE, 2012.

[6] Al-Radaideh, Qasem A., and Eman Al Nagi. "Using data mining techniques to build a classification model for predicting employees performance." International Journal of

Advanced Computer Science and Applications 3.2 (2012).

[7] Ali, Omar, and Nur Zuhan Munauwarah. "Factors affecting employee turnover in organization/Nur Zuhan Munauwarah Omar Ali." (2017).

[8] Frederiksen, Anders. "Job Satisfaction and Employee Turnover: A firm-level perspective." German Journal of Human Resource Management 31.2 (2017): 132-161.

[9] Ongori, Henry. "A review of the literature on employee turnover." (2007).

[10] Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." Expert Systems with Applications 38.3 (2011): 1999-2006.

[11] Alizadeh, Hossein, and B. Minaei Bidgoli. "Introducing A Hybrid Data Mining Model to Evaluate Customer Loyalty." Engineering, Technology & Applied Science Research 6.6 (2016): 1235-1240.

[12] Devi, P. Saranya, and B. Umadevi. "A Novel Approach to Control the Employee‟s Attrition Rate of an Organization." (2018).

[13] Ajit, Pankaj. "Prediction of employee turnover in organizations using machine learning algorithms." algorithms 4.5 (2016): C5.

[14] Verbeke, Wouter, et al. "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach." European Journal of Operational Research 218.1 (2012): 211-229.

[15] Coussement, Kristof, and Dirk Van den Poel. "Integrating the voice of customers through call center emails into a decision support system for churn prediction." Information & Management 45.3 (2008): 164-174.

[16] Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." Expert systems with applications 23.2 (2002): 103-112.

[17] Coussement, Kristof, and Dirk Van den Poel. "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques." Expert systems with applications 34.1 (2008): 313-327.

[18] Burez, Jonathan, and Dirk Van den Poel. "Handling class imbalance in customer churn prediction." Expert Systems with Applications 36.3 (2009): 4626-4636.

[19] Tsai, Chih-Fong, and Mao-Yuan Chen. "Variable selection by association rules for customer churn prediction of multimedia on demand." Expert Systems with Applications 37.3 (2010): 2006-2015.

[20] Coussement, Kristof, Dries F. Benoit, and Dirk Van den Poel. "Improved marketing decision making in a customer churn prediction context using generalized additive models." Expert Systems with Applications 37.3 (2010): 2132-2143.

[21] Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." Expert Systems with Applications 38.3 (2011): 1999-2006.

[22] Kane-Sellers, Marjorie Laura. Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis. Texas A&M University, 2007.

Cite this article as :

Gaurav Thakre, Prajot Wankhede, Sargam Patle, Shrikant Joshi, Prof. Alok Chauhan, "Implementation of Machine Learning Model for Employee Retention Prediction", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 3, pp. 503-508, May-June 2021. Journal URL : https://ijsrst.com/IJSRST2183122