International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Machine Learning Based Employee Attrition Prediction using logistic regression

S.V.KISHORE BABU[1], K INDUMATHY[2], B SUNIL KUMAR[3]

Assistant professor [1],Assistant Professor[2],Assistant Professor[3]

CSE Department, Sri *Mittapalli College* of *Engineering*, *Guntur*, *Andhra Pradesh*-522233

Abstract- In any industry, attrition is a big problem, whether it is about employee attrition of an organization or customer attrition of an e-commerce site. If we can accurately predict which customer or employee will leave their current company or organization, then it will save much time, effort, and cost of the employer and help them to hire or acquire substitutes in advance, and it would not create a problem in the ongoing progress of an organization. In this chapter, a comparative analysis between various machine learning approaches such as Naïve Bayes, SVM, decision tree, random forest, and logistic regression is presented. The presented result will help us in identifying the behavior of employees who can be attired over the next time. Experimental results reveal that the logistic regression approach can reach up to 86% accuracy over other machine learning approaches.

1. INTRODUCTION

Today attrition is one of the major problems faced by industry across the world. It is the most burning issue for the industry, and high attrition rates lead to many issues in the boundary of the organization like losing the talents and knowledge, cost related to training and administration, and recruitment. It is observed that many attributes lead to the attrition of an employee. Which includes working environment, job satisfaction, employer's behavior, job timing, and most important is salary or incentives. Also, the prediction model plays an essential role in finding the behavior of employees.Timely delivery of any service or product is the primary goal of any organization in recent days due to high competition in industries. If a talented employee leaves unexpectedly, the company is not able to complete the task at defined times. It may become the reason for the loss of that company. Therefore, companies are interested in knowing the employee's attrition. They can make a proper substitute or arrangements earlier. There may be various reasons for employee attrition, which include less salary, job satisfaction, personal reasons, or environmental issues if the employer terminates an employee for any reason. It is known as involuntary attrition (Kaur &

Vijay, 2016). On the other hand, voluntary attrition is known as the left of an employee by their side. This kind of attrition is a loss for the company if he or she is a talented employee. In the present scenario, everyone wants a higher salary and job security. Therefore, employees leave jobs immediately if they got a better chance in other places. In the recent era of computer science, machine learning approaches play an important role in employee attrition prediction. These approaches provide predictions based on historical information of the employee, such as age, experience, education, last promotion, and so on. Based on the prediction results HR department have prior knowledge about employee attrition. The HR department also has preplanned recruiting employees as a substitute for the employee who is interested in leaving in the coming days. Various researches have also studied the performance of different machine learning approaches (Ajit, 2016; Sikaroudi et al., 2015). Kaur et al. (Kaur & Vijay, 2016) have discussed various reasons or factors that are involved in employee attrition. They have also investigated that talented employee replacement is a time-consuming and challenging task. It is also a significant factor in loss in business. Compensation is one solution to decreasing the attrition rate. Moncaarz et al. (Moncarz et al., 2009) have discussed how attrition can be decreased by providing better compensation. Punnoose and Ajit (Ajit, 2016) have provided a comparative analysis of various machine learning approaches for employee turnover. Tree-based approaches are also used to predict employee attrition (Alao& Adeyemo, 2013). Jantan et al. (Jantan et al., 2010) have compared tree-based methods with other traditional machine learning approaches. Radaideh and Nagi (Al-Radaideh& Al Nagi, 2012) uses the decision tree for employee attrition prediction. In their work, they have found that job title is an essential feature of attrition, whereas age is not a very important feature. Saradhi (Saradhi&Palshikar, 2011) uses various machine learning approaches for employee attrition prediction. They have taken a database of 1575 records with 25 features of employee and applied various classification approaches to predict attrition. They have shown that SVM has higher accuracy, which is 84.12%. Due to confidentiality and noisy HR data, sometimes prediction has higher accuracy. It is difficult to generalized predictions for different organizations and employee roles (Zhao et al., 2018). \

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

2. PREVIOUS STUDIES

presented accuracy as a primary evaluation standard for attrition prediction. Various machine learning approaches are used and evaluated in different datasets. It is challenging to conclude that which model is best for attrition prediction. The rate of employee attrition is always less than the employee who stays in the organization. Therefore, datasets are always imbalanced. Accuracy measures are not reliable for imbalanced datasets (Sexton et al., 2005; Sikaroudi et al., 2015; Tzeng et al., 2004). So that it is desired to have an accurate model to enhance the prediction accuracy of the models. Which provides better results to employers. Based on the accurate prediction results employers and HR department know the behavior of their employees.

The aim of this chapter is to provide a comparative analysis of different machine learning approaches for employee attrition prediction. Here we have significantly enhanced the training process to solve the imbalanced class problem.

3. METHODOLOGY

In this chapter, various supervised machine learning approaches are used. This section provides a general description of these approaches.

It is a probabilistic approach that uses Bayes theorem to predict the posterior probability of a class. It computes the posterior probability of an event based on the prior knowledge of the related feature(Rane & Kumar, 2018). Equation 1 shows the computation of posterior probability.

$$P\left(\frac{Cl}{x}\right) = \frac{P\left(\frac{x}{Cl}\right).P(Cl)}{P(x)}$$

Support Vector Machine

It is a non-probabilistic supervised machine learning approaches used for classification and regression. It was initially proposed by Vapnik and cooper in 1995(Cortes &Vapnik, 1995). It assigns a new data member to one of two possible classes. It defines a hyperplane that separates n-dimensional data into two classes.

Logistic Regression

It is a traditional classification approach used to assign observations to a particular class. It transforms its output using a logistic sigmoid function to return a probability value that can be mapped to a class. It is a widely used classifier that is easy to implement and works well on linearly separable classes (Raschka, 2015).

Random Forest

It consists of many individual decision trees that are used to train data (Ho, 1995). Each tree in a random forest spits out a class prediction, and the class with the most votes becomes the model's prediction. It uses an ensemble approach that provides an improvement over basic tree structure. It combines various weak learners to form an active learner. Ensemble methods are based on the divide and conquer approach to improve performance.

Decision Tree It is a supervised learning approach that builds a classification model in a tree structure. It makes classification rules by using the top-down approach. It makes sequences of rules which is used to determine the class of a new observation. It uses post pruning methods to handle overfitting problems (Morgan &Sonquist, 1963). Dataset and Tools In this work, we used a publicly available dataset of HR details. This dataset is a simulated dataset that is created by IBM Watson Analytics (McKinley Stacker, 2015). This dataset contained standard HR features such as attrition, age, gender, education, last promotion, job title, and so on. This dataset contained 1470 employee records with 38 features. In this dataset 237 employee has "yes" attrition category while 1233 employee was "no" attrition category. Here all non-numeric values were assigned numerical values. The data conversion was performed using label encoding via the Scikit-learn package in Python (Pedregosa et al., 2011). Furthermore, Python is used in this work to train and evaluate various machine learning approaches. The correlation of different features is a heatmap in figure 1. Figure 1 shows how different features of the HR dataset are correlated. It also shows poorly correlated features and highly correlated features. Experiment Design In this section, the results of various machine learning approaches are illustrated. All the approaches are evaluated on the precision, recall, accuracy, and AUC (Area Under ROC Curve). Various performance parameters are described below

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

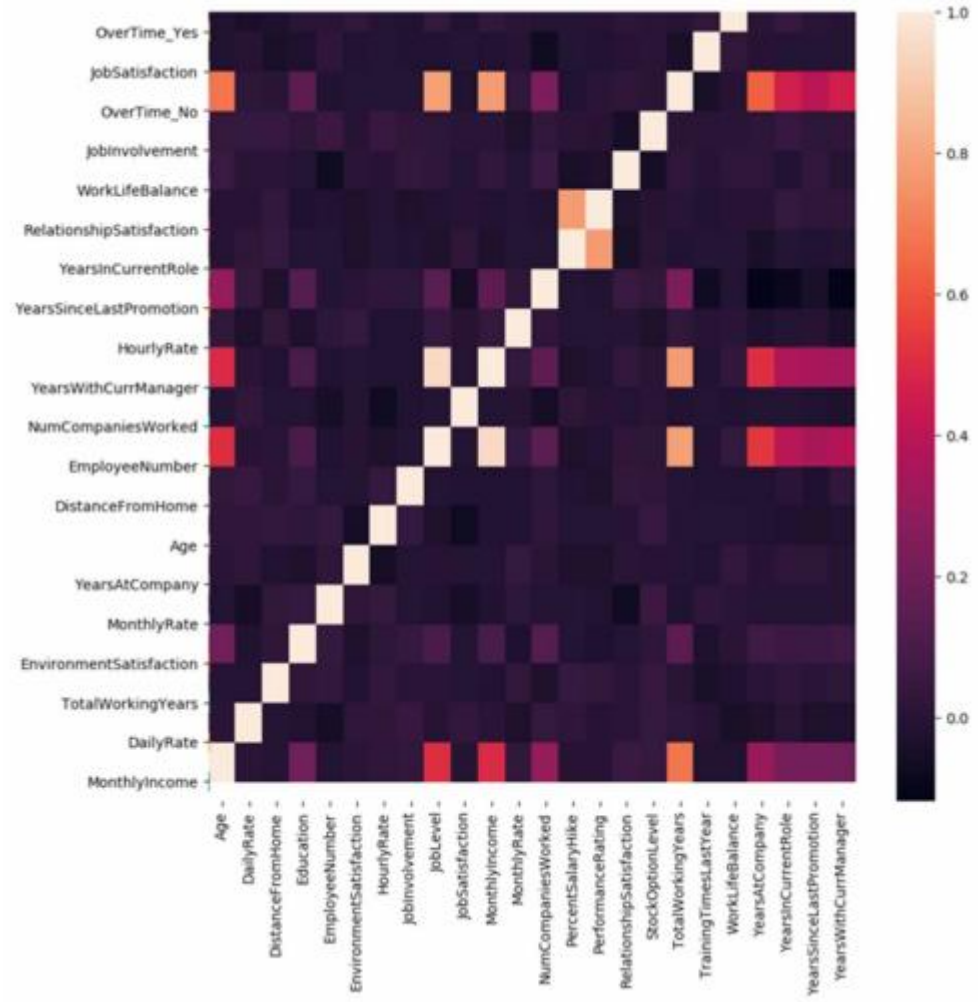
$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$F-\text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

(Alduayj& Rajpoot, 2018; Powers, 2011).

Figure 1. Correlation between various features



International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4. CONCLUSION

Employee attrition has been identified as a significant problem for any organization. High performance and talented employee attrition are considered as a loss for that organization. Finding a substitute for that employee is a time-consuming task. In this work, the performance of various machine learning approaches is evaluated on the HR dataset. Here five different approaches are compared. Based on the accuracy measurement, logistic regression well performed for this dataset. It has higher precision, recall, and accuracy. The result of the attrition prediction will be helpful for an organization to reduce the attrition rate of their company.

REFERENCES

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. Algorithms, 4(5), C5. Al-Radaideh, Q. A., & Al Nagi, E. (2012).

Using data mining techniques to build a classification model for predicting employees performance. International Journal of Advanced Computer Science and Applications, 3(2).

Alao, D., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. Computing, Information Systems, Development Informatics and Allied Research Journal, 4. Alduayj, S. S., & Rajpoot, K. (2018). Predicting Employee Attrition using Machine Learning. 2018

International Conference on Innovations in Information Technology (IIT), 93–98. 10.1109/INNOVATIONS.2018.8605976 Cortes, C., &Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. doi:10.1007/ BF00994018 Ho, T. K. (1995).

Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278–282. Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010).

Human talent prediction in HRM using C4. 5 classification algorithm. International Journal on Computer Science and Engineering, 2(8), 2526–2534. Kaur, S., & Vijay, M. R. (2016).

Job satisfaction-A major factor behind attrition of retention in retail industry. Imperial Journal of Interdisciplinary Research, 2(8), 993–996.