Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)
IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

Evaluation of Machine Learning Models for Employee Churn Prediction

Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari
Department of Computer Science & Engineering
National Institute of Technology Raipur, India

***Abstract*— Employees are the valuable assets of any organization. But if they quit jobs unexpectedly, it may incur huge cost to any organization. Because new hiring will consume not only money and time but also the freshly hired employees take time to make the respective organization profitable. Hence in this paper we try to build a model which will predict employee churn rate based on HR analytics dataset obtained from Kaggle website. To show the relation between attributes, the correlation matrix and heatmap is generated. In the experimental part, the histogram is generated, which shows the contrast between left employees vs. salary, department, satisfaction level, etc. For prediction purpose, we use 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. This paper proposes the reasons which optimize the employee attrition in any organization.**

Keywords—Turnover;JobSatisfaction;Attrition;Organization; employee retention strategy

I. INTRODUCTION

Employee Attrition [1] is a reduction in manpower in any organization where employees may voluntarily leave the organization or may be retired. Employee turnover is the number of existing employees replaced by new employees for a specific period. A high attrition causes high employee turnover in any organization. This in turn causes huge expenditure on human resource, by contributing towards new recruitment, training and development of the freshly appointed employees, also the performance management. Again, attrition [2] which are of voluntary is unavoidable. Hence, by improving employee morale and providing a desirable working environment, we can certainly reduce this problem significantly.

The rate of attrition is defined as the recruitment and termination criteria of the company. An employee can leave the job for various reason. Here, the 'Turnover' and 'Attrition' are the business terminologies that always conflicts each other. There are various kinds of 'turnover' in an organization. Lowering in number of employee is mainly considered as the 'attrition'. To analyze the manpower data and other measurements that are necessary for manpower planning these terminologies can be interchangeably used. When an employee leaves the company both attrition and turnover happens. Turnover, be that as it may occur because of various work activities, for example, release, termination, abdication or occupation relinquishment. Attrition happens when an employee resigns or when the organization eliminates his occupation. The real contrast between the two is that when turnover happens, the organization looks for somebody to supplant the employee. In instances of attrition, the business leaves the opportunity unfilled or wipes out that employment job**.**

This paper is organized as follows. The next section describes the related works done in the past and the motivation regarding this analysis. Section 3 will describe different machine learning algorithms used in this paper and their significance. Section 4 describes about the data set and also shows the statistical information using the data set. Section 5 contains the detailed experimental results using the machine learning algorithms using the mentioned data set, which will be followed by the conclusion section.

II. LITERATURE SURVEY

Middle level officers are more likely to leave, may be due to some disagreement with their senior officer as proposed by [3]. They observed major factors that influenced employee abandonment from the firm. The two rules are moderately derived by him. Some set of questions are asked with the both parties and depending upon their answers he concluded some facts based on workload, objectives, carrier opportunity and firm management. Human resource management [4] endeavors on basically termination rates and dismissal rates but actual content of them are enormously different. The previous model shows that, there are several distinct levels of attrition and turnover. Some research dictates that the consequences of dismissal and termination rates are at organizational level. Allen & Meyer (1990) [5] described the three-basic entity for the negative side of the turnover.

Regulating officer will more probable leave from the organization because of a contention with the higher administration than a representative who is in struggle with his prompt director. He recognized the determinant figures that influence employee acceptance [5] without protest from the organization. Two arrangements of information social occasion techniques were directed. An equivalent number of representative and officer respondents were solicited to answer a set from polls that were ordered by workload, objectives, identity, professional success, and hierarchical administration. The after-effects of the two information gathering methods

demonstrated that the most noteworthy component that adds to employee rejection [6] is money related compensation.

III. MACHINE LEARNING ALGORITHMS

There are various kind of machine learning techniques available to learn from the given data which is called train data. When new or unseen data arises the learned model analyses and predict desired class. In our experiment we have used the HR Analytic data set to apply various machine learning algorithms to predict the chances of employees to quit the job. The machine learning algorithms for predicting the same are described below.

A. K-Nearest Neighbour

k-NN classifier [7] is known as lazy learner in machine learning community. It never learns from the data and do not build any models. Rather, it finds out the examples from the train dataset which are closest to the unknown example. Based on the neighbor examples it will predict the new example. The value of 'k' determines the no. of closest data points or examples to be selected from the training example.

B. Supprt Vector Machine

A Support Vector Machine [8] 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 categorizes the data set in to two classes. When an unknown data comes it predicts which side/class it belongs to. The margin between the hyperplane and the support vectors are as large as possible to reduce the error in classification.

C. Naïve Bayes Classifier

Naive Bayes [9] is a popular classification technique which classifies examples based on the probability of chances that are likely to be occurred. It often performs very well for complex data set which are very hard to learn using the traditional learning algorithms.

D. Decision Tree

This is one of the popular learning techniques in machine learning. C 4.5 [10] is the benchmark learning algorithm in decision tree which is often compared with the new algorithms that are being developed. Here C 5.0 [11] learning algorithm is used which is an advanced version of traditional decision tree learning algorithms. The nodes along with the edges are the series of conditions and the leaves are the class labels.

E. Random Forest

It is one of the ensemble learning technique [12] which consists of several decision tree 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.

IV. DATA SET ANALYSIS

The 'HR Analytics' data set [13], obtained from Kaggle Website, is used in this paper for the experimental verification. This data set comprises ten attributes and 15000 tuples.

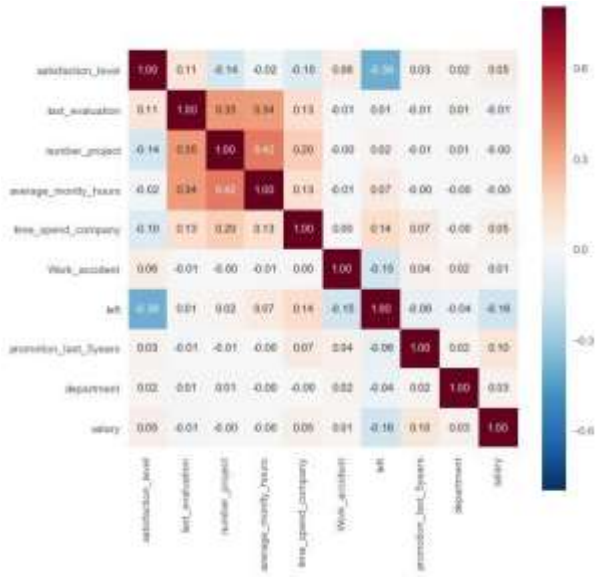

Fig.1. Correlation Matrix

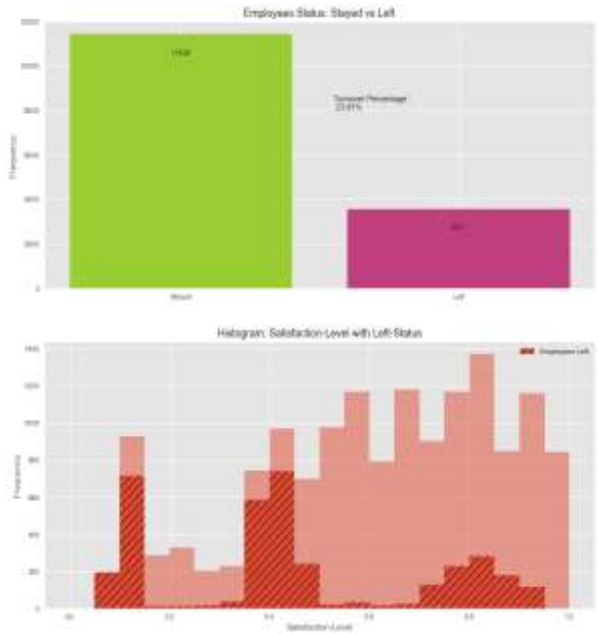

Fig. 2. Histogram of employee status and satisfaction level

The categorical values are converted to numeric values in order to make the classification algorithm more efficient. For example, categorical attribute 'salary' contains three values such as low, medium and high. Hence it is converted to 0, 1 and 2 respectively. The misspelled attributes are also corrected. Figure (1) represents the correlation matrix which helps to identify attributes with the strong or weak correlation.

Figure (2) represents the histogram of employee status and satisfaction level [14] . It can be seen from the figure that, there are three segments or behaviors.

- First bin in the histogram is empty. Second and third bins mostly contain people who left the company. These people must be unhappy for some reason as their satisfaction level is below 1.15, i.e. (1st Segment) 25.4% of the total employees who left.
- Then there is a peak around 0.4 (2nd Segment) i.e. 43.8% of total employees who left.
- Lastly, there is a chunk 0.7-0.95 (3rd segment) who left the company, i.e. 26.1 % of total employees who left. That's almost 1/4th of employees who left.

Figure (3) represents the histogram of comparison among the departments with respect to employee attrition and turnover. From the first graph of this figure tell that, Turnover rate (percentage of churned employees) of department of 'R & D' and 'Management' are the lowest. When it comes to employee churn, 'HR' department is having the highest with 29% and rest of the departments are roughly the same. Again, from the second and third graph of this figure it can be seen that highly paid employees are less likely to leave the job which makes sense.

From Figure (4), only 5 % of the employees who got promoted churned away, as compared to 24% of employees who were not promoted and left the company. Hence 'promotion' is also an important factor of retaining employees.

Also, it can be seen from Figure (4) that, employees with projects '2', '6' and '7' left the company most which represents most of the employees quit if the work load increases. However, there is an inconsistency where number of projects is equal to 2, which may be due to the nature of the job in project or other. Again, employees, who are there in the company for a long time are less likely to quit as compared to the ones who have joined recently, which can be seen from the graph 'time spent' from Figure (4).

From Figure (5), employees with lower 'Last-Evaluation' left the company. But it is a little ambiguous why employees with high 'Last-Evaluation' left while employees with 0.6-0.8 evaluation score tend to stay. Employees with heavy working hours also tend to leave the company, which is obvious.

Using all these attributes and precondition basically employee churn rate can be generalized to the following criteria:

- Employees with no promotion in last 5 years tend to leave.
- Employees having high salary but no promotion and high working hours also tend to leave.
- Employees getting promotion but not hike in salary also tend to leave.

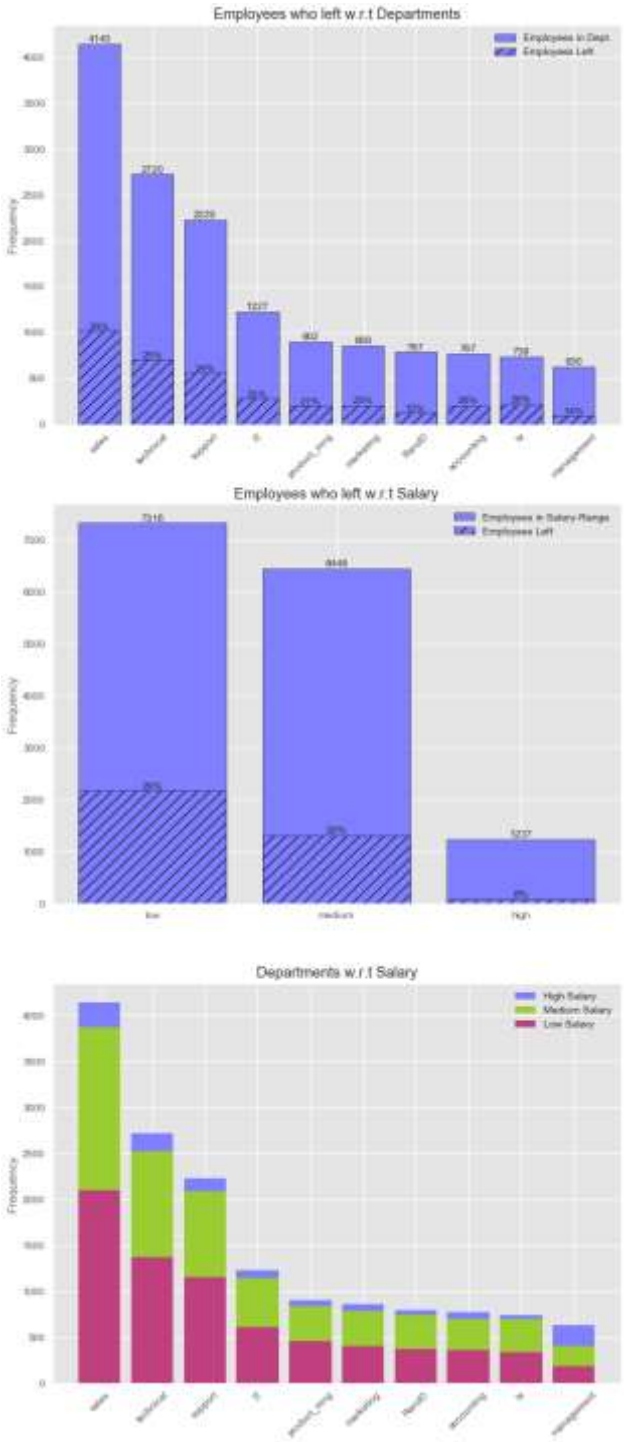

Fig. 3. Histogram of department and salary

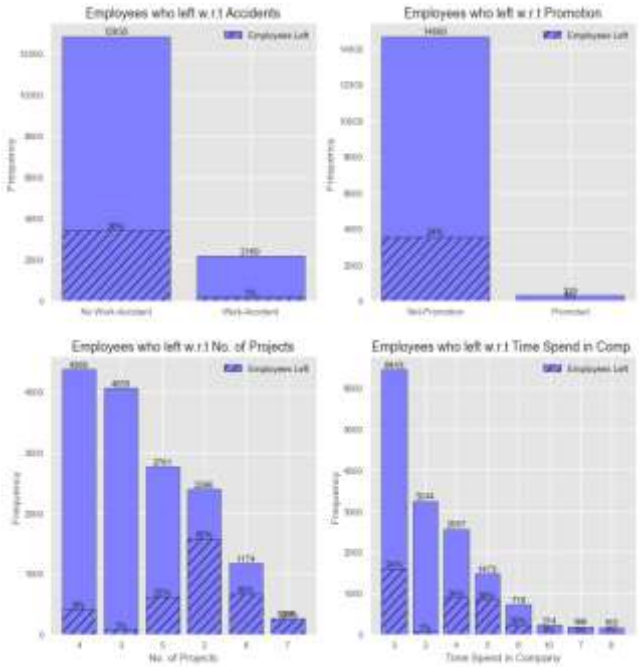

Fig. 4. Histogram of accident, promotion, project and time spend

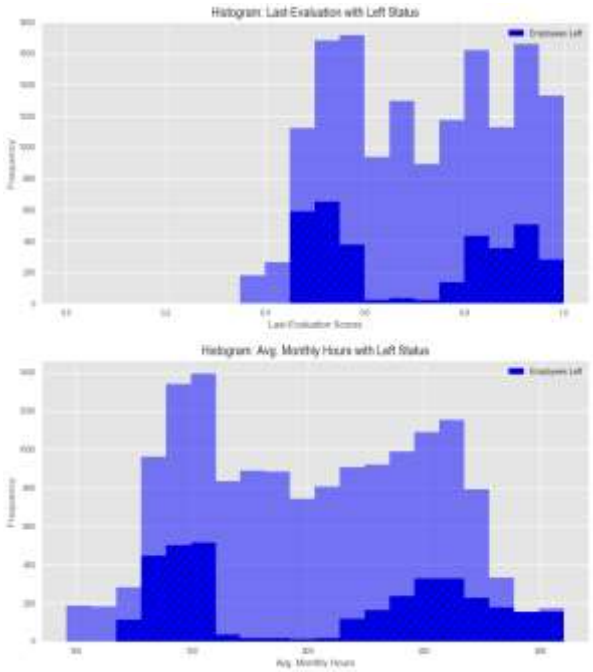

Fig. 5. Histogram of last evaluation and average monthly hours

V. EXPERIMENTAL RESULSTS AND EVALUAITON

In the data set mentioned above, the attributes like employee's salary, current position, promotion etc. is given. Based on these values the learning algorithm [15] will predict whether the employee will quit the organization or not. The predicted value is compared with the actual value in the database.

For evaluating the experimental results, 'Confusion Matrix' is used which is a common evaluation criterion for any classification model. Using this the parameters like Accuracy, Precision, Recall and F-Measures are used and the corresponding values obtained through experiment is displayed in Table 1 with respect to different learning techniques.

It can be seen from Table 1 that, Random Forest classifier gives the highest accuracy on the given HR analytics data set while, LSVM classifier gives the lowest accuracy for the same data set. The same result is also displayed in Figure 6 in terms of graphs. From Figure 6 we can conclude that Random Forest again gives the highest precision i.e. true positive rate. Likewise, Random forest also performing well for other measures like Sensitivity or Recall, F-Measure, Specificity, FPR and FNR as compared to other classifiers. LSVM classifier gives the lowest sensitivity and Naïve Bayes gives the lowest F-Measure for the given data set.

TABLE I. RESULTS OF DIFFERENT CLASSIFIERS

Attribute/ models	KNN	LSVM	Naïve Bayes	Decision Tree	Random Forest
Accuracy	0.9600	0.7821	0.7918	0.9768	0.9897
Precision	0.9674	0.9457	0.8081	0.9796	0.9981
Sensitivity or Recall or True positive rate	0.9799	0.8032	0.9085	0.9898	0.9886
F-measure	0.9736	0.8686	0.8554	0.9846	0.9933
Specificity or True negative rate	0.8997	0.5979	0.5463	0.9368	0.9936
False positive rate	0.1003	0.4021	0.4537	0.0632	0.0064
False negative rate	0.0201	0.1968	0.0915	0.0102	0.0114

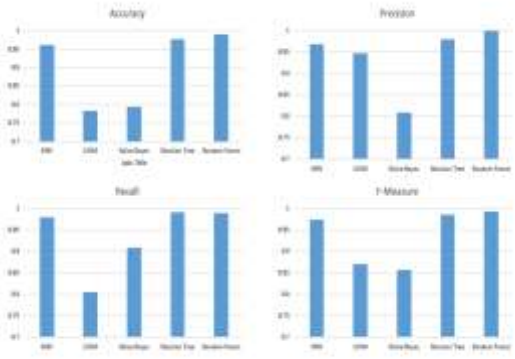

Fig. 6. Performance Comparison of Different Classifiers

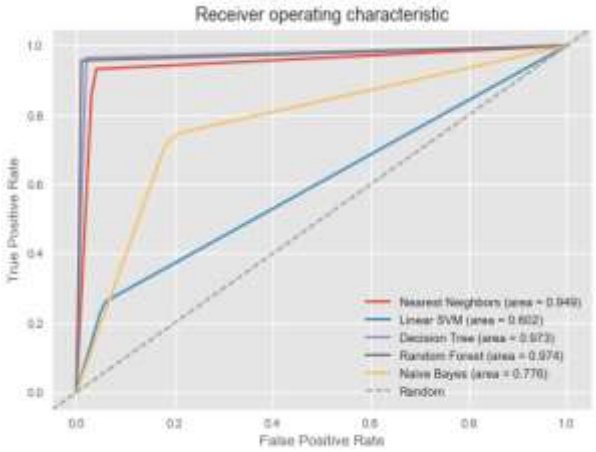

Fig. 7. ROC Curve for different classifiers

From Table 1, using false positive rate (FPR) and true positive rate (TPR) we have plotted the receiver operating characteristic (ROC) curve in Figure 7 for different classifiers. The higher the area under the curve, greater is the accuracy of the classifiers.

VI. CONCLUSION

This paper find out which machine learning algorithm is performing well in predicting the employees, those are likely to quit the respective organization based on their working details and environments. From the experimental results, Random Forest is clearly outperformed all other classifiers as obtained in evaluation criteria. This study might help organization what are the factors causing the employee leaving the organization and can take appropriate steps to minimize that. This study requires further exploration to minimize the prediction error rate.

REFERENCES

[1] R. Punnoose and P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 9, 2016.

[2] D. Alao and A. B. Adeyemo, "Analyzing employee attrition using decision tree algorithms," *Comput. Inf. Syst. Dev. Informatics Allied Res. J.*, vol. 4, 2013.

[3] O. Ali and N. Z. Munauwarah, "Factors affecting employee turnover in organization/Nur Zuhan Munauwarah Omar Ali," 2017.

[4] A. Frederiksen, "Job Satisfaction and Employee Turnover: A firm-level perspective," *Ger. J. Hum. Resour. Manag.*, vol. 31, no. 2, pp. 132–161, 2017.

[5] H. Ongori, "A review of the literature on employee turnover," 2007.

[6] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1999–2006, 2011.

[7] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[8] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[9] K. P. Murphy, "Naive Bayes classifiers Generative classifiers," *Bernoulli*, vol. 4701, no. October, pp. 1–8, 2006.

[10] D. Gupta, D. S. Kohli, and R. Jindal, "Taxonomy of tree based classification algorithm," in *Computer and Communication Technology (ICCCT), 2011 2nd International Conference on*, 2011, pp. 33–40.

[11] A. S. Galathiya, A. P. Ganatra, and C. K. Bhensdadia, "Classification with an improved decision tree algorithm," *Int. J. Comput. Appl.*, vol. 46, no. 23, pp. 1–6, 2012.

[12] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 5, pp. 1–35, 1999.

[13] Kaggle, "HR Analytic Data set." [Online]. Available: https://www.kaggle.com/ludobenistant/hr-analytics.

[14] K. L. Latha, "A study on employee attrition and retention in manufacturing industries," *BVIMSRs J. Manag. Res.*, vol. 5, no. 1, pp. 1–23, 2013.

[15] A. Rawat and A. Choubey, "A Survey on Classification Techniques in Internet Environment," *IJSRSET*, vol. 2, no. 3, 2016.