Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees

Rohit Hebbar A
ISE, Ramaiah Institute of Technology
Bangalore, India
rohithebbar@gmail.com

Sanath H Patil
ISE, Ramaiah Institute of Technology
Bangalore, India
sanath.sunny1@gmail.com

Rajeshwari S.B
ISE Department
Ramaiah Institute of Technology
Bangalore, India
rajeshwari.sb@msrit.edu

S S M Saqquaf
DIGITS
Indian Institute of Science
Bangalore, India
smsaqquaf@iisc.ac.in

***Abstract*— In most of the organizations, Employee Attrition has been one of the greatest concerns in today's world. The reason behind this can be due to personal or company related issues such as long- distance travelling, no work life balance, less salary hike, no job satisfaction etc. According to a study done by Businessdictonary, employee attrition results from resigning from their post, retirement, illness, or demise. Considering these issues, the project aims to find the employees who are most likely to attrite from the organization using pre-processing techniques such as exploratory data Analysis (EDA), feature selection techniques and utilizing various machine learning techniques such as Logistic Regression, Support Vector Machine (SVM) and Random Forest. According to which several programs can be incorporated by the organizations to minimize the attrition rate and also help in building and maintaining a robust relationship between the employees and the organization.**

***Keywords*— Attrition, Data Analysis, Feature Selection, Machine Learning, Classification.**

I. INTRODUCTION

Employee attrition has been one of the greatest concerns in most of the organizations. A lot of employees have been resigning from their organizations lately due to various factors. The reason behind the resignation involves a lot of personal as well as company related issues. This problem has been affecting many organizations and causing hindrance towards the development of the organization. By knowing the exact reason for an employee to attrite from an organization, the companies can develop a program to minimize the attrition rate which will help build a robust relationship between the organization and the employees.

In this study, we implement some of the well-known machine learning techniques for data classification such as Logistic Regression, SVM and Random Forest on HR Employee Attrition data set provided by IBM from Kaggle. The data set contains 2941 records with 34 features. We first implemented Logistic Regression to estimate the probability that an individual will fall into one outcome group or the other. In addition, to carry out the comparative study we proceeded to implement the common classification techniques such as Random Forest and SVM. Furthermore, we performed EDA to determine the main characteristics of the data set through visual representations using various graphs and plots.

II. LITERATURE SURVEY

In the research work [1], the author has taken the feedback from 60-current employees and 60 Ex-Employees and has decided on certain Key factors that would make an employee leave the Industry that they have been working on. This study was conducted to find out the main reasons behind an employee resigning the company and also to predict and control them.

This research was carried out in different BPO companies. This research was purely carried out in descriptive in nature. The main aim of this research was to reduce the attrition rate that had been increasing in BPO companies. For this research they have used structured questionnaire for collecting the data and they carried out test like Chi-Square Test, ANOVA and Percentage Analysis to predict whether an employee is leaving the company or not.

Remarks: In this paper, the only reason for the attrition rate has been due to resignation. They haven't considered attrition due to death, or due to illness of an employee or other personal reasons. And all the opinions from the employees has been considered true and valid. Also, this study is related to only one of the divisions in the company that is, call center division. Hence, it might not be applicable to all the other divisions in the company.

In this Paper [2], the author has used Logistic Regression Technique to predict whether an employee is going to resign from the company or not. The model is based on the demographic data that has been collected from the resigned employees. This is a real-life project that has been executed with their clients. They considered data from the current employees as well as the resigned employees. This paper also presents a model that will reduce the employee attrition rate

978-1-5386-2440-1/18/$31.00 ©2018 IEEE

and provide reasons as to why the employees had been resigning. The cost of the attrition rate can be reduced by meeting the demands of the employees. It is also a known fact that by retaining the best employees of an organization, their revenue can be increased also all the staffs and colleagues will be happy.

Remarks**:** The motive behind this paper is to help the organizations to predict the employee attrition rate in real time and also take the necessary steps to prevent it, or to keep the required employees in advance. Also, the accuracy of logistic regression algorithm was 80% for the employees who separated from the company and 62% for the employees who stayed in the organization.

III. MODELLING AND IMPLEMENTATION

In the research done, models were built using machine learning techniques to predict whether an employee is going to attrite from an organization or not. The models that have been implemented include Logistic Regression, Random Forest and SVM. Comparison and performance analysis of the model are also performed to find the best performing model.

Table I. Condition for Calculating Confusion Matrix

Condition		True Condition	
Predicted Condition	Predicted condition Positive	Condition Positive	Condition Negative
		True Positive	*False Positive*
	Predicted Condition Negative	*False Negative*	*True Negative*

Formulae to Calculate Metrices:

$$\mathit{Accuracy} = \sum \mathit{TruePositive} + \sum \mathit{TrueNegative} / \sum \mathit{TotalPopulation}$$

$$\mathit{Sensitivity} = \sum \mathit{TruePositive} / \sum \mathit{ConditionPositive}$$

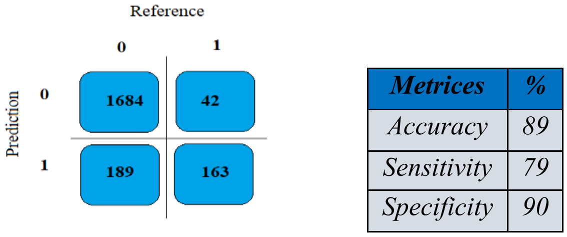
$$\mathit{Specificity} = \sum \mathit{TrueNegative} / \sum \mathit{ConditionNegative}$$

A. Logistic Regression

For predicting the values for binary variables Linear Regression is not suitable for prediction as it will predict values outside the acceptance range i.e values exceeding the range between 0 and 1.

Logistic Regression is used when the variable to be classified is in 0 or 1 state. Where, 1 is 'success' and 0 is 'failure'. In this technique the probability of individual variables affecting the final variable is estimated.

Table II. Resultant Metrics for Logistic Regression

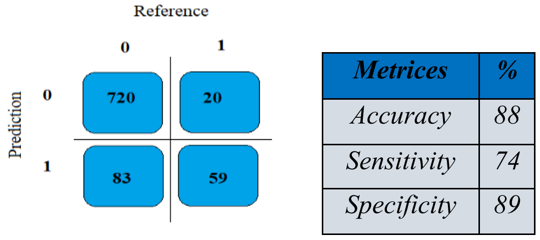


		Reference	
		0	1
Prediction	0	1684	42
	1	189	163

Metrices	%
Accuracy	*89*
Sensitivity	*79*
Specificity	*90*

Fig 1. Confusion Matrix: Train Data Set

Table III. Resultant Metrics for Logistic Regression



		Reference	
		0	1
Prediction	0	720	20
	1	83	59

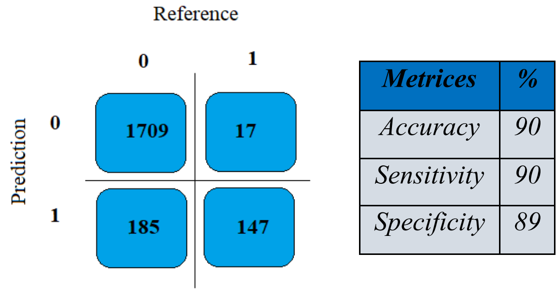
Metrices	%
Accuracy	*88*
Sensitivity	*74*
Specificity	*89*

Fig 2. Confusion Matrix: Test Data Set

B. Random Forest

The advantage of using Random Forest Algorithm is that it works good for both classification and regression on a large dataset. It is a collection of multiple decision trees and aids in determining only those features that are affecting the final variable.

Table IV. Resultant Metrics for Random Forest

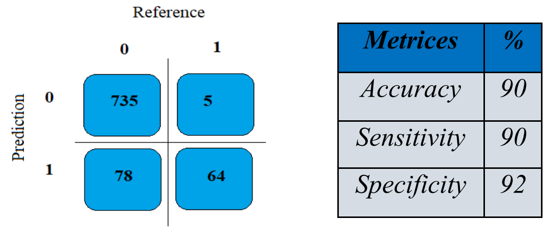


		Reference	
		0	1
Prediction	0	1709	17
	1	185	147

Metrices	%
Accuracy	*90*
Sensitivity	*90*
Specificity	*89*

Fig 3. Confusion Matrix: Train Data Set

Table V. Resultant Metrics for Random Forest



		Reference	
		0	1
Prediction	0	735	5
	1	78	64

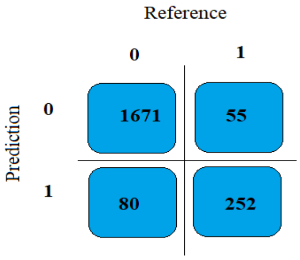
Metrices	%
Accuracy	*90*
Sensitivity	*90*
Specificity	*92*

Fig 4. Confusion Matrix: Test Data Set

C. Support Vector Machine

SVM is a machine learning technique which is used for data classification. In this technique we plot the data points on the hyperplane and estimate the distance of the data points from the separator.

Table VI. Resultant Metrics for SVM

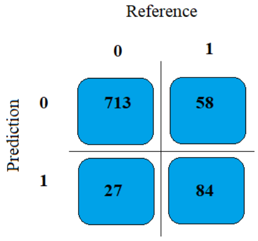


		Reference	
		0	1
Prediction	0	1671	55
	1	80	252

Metrices	*%*
Accuracy	*93*
Sensitivity	*82*
Specificity	*95*

Fig 5. Confusion Matrix: Train Data Set

Table VII. Resultant Metrics for SVM



		Reference	
		0	1
Prediction	0	713	58
	1	27	84

Metrices	*%*
Accuracy	*90*
Sensitivity	*60*
Specificity	*96*

Fig 6. Confusion Matrix: Test Data Set

IV. Result and Discussion

A. Model Comparison

Table VIII. Train Sample

Statistical model	*Sensitivity (%)*	*Specificity (%)*	*Accuracy (%)*
Losgistic Regression	*79*	*90*	*89*
Random Forest	*89*	*90*	*90*
SVM	*82*	*95*	*93*

Table IX. Test Sample

Statistical model	*Sensitivity (%)*	*Specificity (%)*	*Accuracy (%)*
Losgistic Regression	*74*	*89*	*88*
Random Forest	*90*	*92*	*90*
SVM	*60*	*96*	*90*

Based on above analysis shown in table VII and VIII, following are the recommendation and conclusions drawn:

- As per the accuracy SVM is coming out to be the best performing model on the train sample. But, on the test sample the accuracy has reduced by 3%. Hence, we can conclude that this model is not the best performing model.
- Logistic regression is performing better on the training sample but with respect to sensitivity and accuracy there is a considerable drop on the test sample.
- Random Forest is giving good and consistent performance across both training and validation data.

Keeping in mind the objective of our project where we are more focused on correctly predicting the employees who are most likely to attrite from the organization, by maximising the True Positive Rate (Sensitivity). Random Forest is coming out to be the best performing.

B. Logistic Regression

To predict the employee attrition rate using logistic regression, we first split the data into train and test samples respectively. When the model was built on the train data sample, an accuracy of 89% was observed. Also, to show the trade-off between the sensitivity and the specificity, ROC (Receiver Operating characteristics) curve is plotted and to determine whether the model is predicting the classes to the best of its ability AUC (Area Under Curve) is also plotted. ROC and AUC can be used to analyse the performance of the model for different cut-off values.

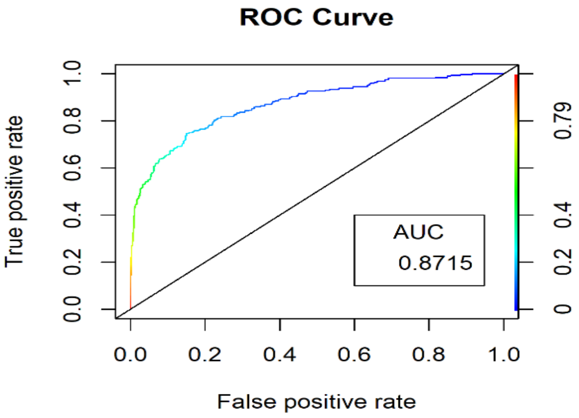


Fig 7. AUC-ROC Curve for Train Data Set

The trade-off between the True positive rate (Sensitivity) and the False positive rate (Specificity) can be observed in Fig 7. An ideal situation for a ROC curve would be [0,0], [0,1], [1,1] for 100% accuracy but in reality, this will not be the case for any data set.

An optimal line is drawn in the plot having intercept=0 and slope=1 this is the case when all the employees are leaving the organization. If the curve is below this benchmark line then the model is not performing well.

Furthermore, the AUC is found to determine the performance of the model on the training sample which shows 87.15%.

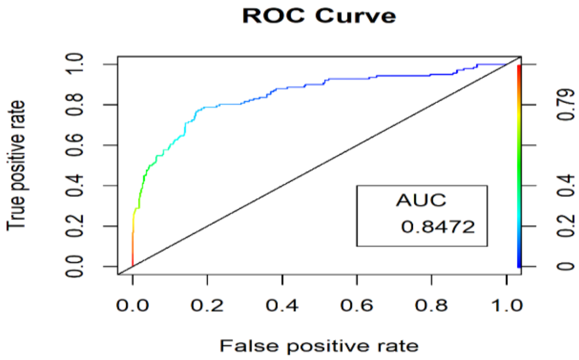


Fig 8. AUC-ROC Curve for Test Data Set

By looking at the ROC and AUC plot (Fig 8) on the validation sample, the performance of the model on this sample is 84.72 %. When the model was built on the test data sample, an accuracy of 90% was observed.

C. *Random Forest*

To predict the employee attrition rate using Random Forest, we first split the data into train and test samples respectively. Furthermore, the Boruta package is used for variable selection to enhance model building. When the model was built on the train data sample, an accuracy of 90% was observed.

Also, to show the trade-off between the sensitivity and the specificity ROC (Receiver Operating characteristics) curve is plotted and to determine the model is predicting the classes to the best of its ability AUC is also plotted. ROC and AUC can be used to analyze the performance of the model for different cut-off values.

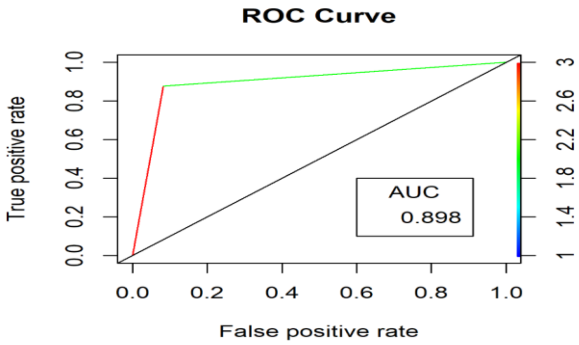


Fig 9. AUC-ROC Curve for Train Data Set

By looking at the ROC and AUC plot (Fig 9) on the train sample, the performance of the model observed on this sample is 89.8 %.

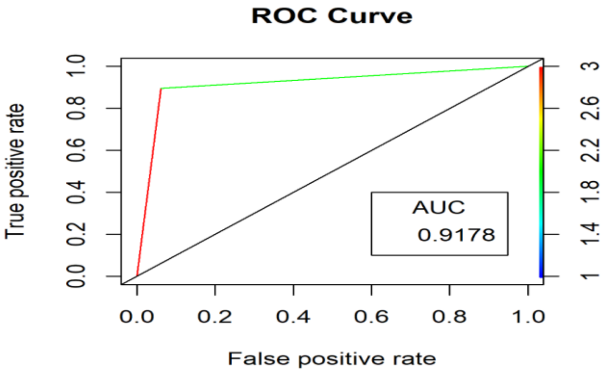


Fig 10. AUC-ROC Curve for Test Data Set

By looking at the ROC and AUC plot (Fig 10) on the train sample, the performance of the model on this sample is 91.78 %. When the model was built on the train data sample, an accuracy of 90% was observed.

Since the accuracy and performance of this model is comparatively better than the other models built, Random Forest is considered as the best performing model.

V. CONCLUSION

Employee attrition has been one of the greatest concerns for most of the organizations in today’s world. This problem has been causing a hindrance towards the development of the organization as the organizations would have invested both time and money, two of the most valuable assets, towards the employee’s.

We seek to analyze the employee data and predict the future attrition’s and also learn the key factors affecting the same. By looking at the results of this study, we can conclude that various machine learning techniques can be used to demonstrate and build reliable models for comparison of the attrition rate.

Our goal is not just limited to only finding the employee attrition but also the variables which are affecting the attrition variable. By using EDA (Exploratory Data Analysis) the main characteristics of the employee data can be analyzed through visual representation such as plots and graphs.

Also, variable selection is very important for model building as unimportant and redundant variables will result in overfitting and slow computation. Hence Boruta package, which is an all-relevant variable selection technique, is used for variable selection for Random Forest model to enhance model building. By, using various machine learning techniques we can predict the employees who are most likely to attrite from the organization.

As a future direction, a comprehensive and universal program can be developed which can be inculcated by the organizations to minimize attrition rates and which will not only help for development of the organization, but also help build a robust relationship of the employees with the organization

REFERENCES

[1] V. VIJA Y ANAND, R. SARA V ANASUDHAN & R. VIJESH. (March 30,31,2012**)**. *Employee Attrition - A Pragmatic Study with reference to BPO Industry*. Tamil Nadu, India: IEEE-International Conference on Advances in Engineering, Science and Management.

[2] Rupesh Khare, Dimple Kaloya, Chandan Kumar Choudhary, Gauri Gupta. (January 8-9, 2011). *Employee Attrition Risk Assessment using Logistic Regression Analysis*. Ahmedabad: Indian Institute of Management.

[3] N. Kasap et al. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications. 41(2).

[4] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[5] Al-Radaideh, A., Al-Nagi, E.: 'Using data mining techniques to build a classification model for predicting employees performance', Int. J. Adv. Comput. Sci. Appl., 2012, 3, (2), pp. 144–151

[6] Modi, M., Patel, S.: 'An evaluation of filter and wrapper methods for feature selection in classification', Int. J. Eng. Dev. Res., 2014, 2, (2), pp. 1730–1733

[7] Mitchell, M.: 'Generative and discriminative classifiers: Naive Bayes and logistic regression', 'Machine learning' (McGraw-Hill, New York, USA, 1997). Copyright © 2015

[8] Gwendolyn M Combs, Rachel Clapp-Smith, Sucheta Nadkarni, Managing BPO service workers in India: Examining hope on performance outcomes. Human Resource Management (2010), Volume: 49, Issue: 3, Pages: 457-476 - ISSN: 00904848.

[9] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1999–2006, 2011.

[10] Shradha Prakash & Rahul Chowdhury, *Managing Attrition in BPO - A win-win model to satisfy employer and the employee*, Thursday, October 28, 2010.

[11] Neeraj Pandey, Gagandeep Kaur, *Factors influencing employee attrition in Indian ITeS call centres*, International Journal of Indian Culture and Business Management, Volume 4, Number 4/2011, Pages 419-435.

[12] X. Lin, F. Yang, L. Zhou, P. Yin, H. Kong, W. Xing, X. Lu, L. Jia, Q. Wang, and G. Xu, "A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information," Journal of chromatography B, vol. 910, pp. 149–155, 2012.

[13] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.