Employee Attrition Prediction using Data Mining Techniques

[1]Jeel Sukhadiya, [1]Harshal Kapadia, [2]Prof. Mitchell D'silva

[1] Student, [2]Assistant Professor

[1-2]Department of Information Technology,

[1-2]Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai
jeelsukhadiya2@gmail.com, harshalkpd93@gmail.com,
mitchell.Dsilva@djsce.ac.in

Abstract

Employees are invaluable assets to an organization even just by their mere presence in it. When employees leave an organization, the situation has an overall impact on the organization in terms of loss in the resources, information or knowledge that the employees possess. To have an advantage over other organizations and to be on top in the market it is very much essential to reduce employee attrition. This paper discusses the prediction of employee attrition using various data mining techniques such as Random Forest, Support Vector Machines (SVM), Gradient Boosted Classifier and Logistic Regression. The dataset used for predicting and classifying the attrition is a fictional dataset created by IBM data scientists. The classification techniques used give good accuracies but amongst all the techniques implemented in this paper, Extreme Gradient Boosting proves to have a upper hand on the attrition prediction task.

.

Keywords: Data Mining, Random Forest, SVM, Logistic Regression, Gradient Boosted Classifier.

1. Introduction

In general terms, attrition is defined as a process of reducing the strength of a particular thing to reduce the effectiveness. In this paper the thing being referred to is an organization in which the attrition of employees is being deliberated over. For an organization, employees are an invaluable asset and employee attrition in particular affects the growth strategies and the resource balance of the organization. Reduced opportunities, seldom satisfaction with the job profile or the working environment and also the challenges faced with the management can in turn lead to high employee attrition rate. These problems also hamper the status of the organization and thus to find a solution to the rising rate of attrition, various data mining techniques have been implemented to predict the attrition rate. The techniques used are Logistic Regression, extreme Gradient Boosted Classifier (XGBoost), Support Vector Machines (SVM) and Random Forest.

2. Literature Review

In this section, along with the various contemporary works, the advantages and drawbacks of those works have been discussed. Data mining techniques are widely used in various sectors such as marketing, sales, customer relationship management. The works of Wang et al [1] provide a way of employing data mining techniques for customer relationship management whereas the work in [2] provides the use of data mining for Marketing, sales and customer relationship management. Other work done in [3] focuses on the use of data mining for customer relationship management. Thus, data mining is extensively used for multiple human resource (HR) related applications. In [4] different data mining

techniques are used for prediction of employee turnover based on historical and personal data of the employee. There are various applications of data mining ranging from prediction to classification of various parameters and features of HR related data under consideration. In [5], the early prediction of employee attrition is done where factors such as absenteeism, late-coming, disinterest of an employee are taken into consideration for predicting employee attrition. The factors and parameters considered are fairly limited and they do not provide a considerable ground to predict the attrition. The work proposed in [6] where employee churn prediction is done using Support Vector Machines and other data mining techniques also suffer from the major drawback of limited parameters and do not take into account the useful parameters for prediction as found in [5]. Thus with an impetus to accurately predict the attrition in employees and the reasons/factors responsible for the attrition we provide a detailed analysis of different approaches. The parameters which closely substantiate the build of attrition among the employees are explored and the numbers of parameters considered are greater than the previous approaches. Thus a detailed evaluation of different data mining techniques is done to accurately predict the attrition and the parameters leading to employee attrition.

3. Methodologies

In this section, various methods or techniques used in the paper to predict employee attrition have been discussed along with their respective diagrams. Also the working of the techniques mentioned have been stated.

3.1.1 Decision Tree and Random forest:

Decision Trees [7] are very popular amongst classifier algorithms due to their ease of interpretations and implementations. From the training data set, the algorithm builds a tree in which each node is an attribute and the branches represent the corresponding attribute values. A problem faced by decision trees is instability in which small changes in the input training samples may cause dramatically large changes in output classification rules. Because of this reason, random forests are used.

Random Forest [8] is an ensemble classifier which combines more than one algorithm to classify objects. For example, it can combine Naive Bayes, SVM and Decision Tree and then after combining them it takes final vote to classify the object. In random forest classifier, first it creates a set of decision trees from a subset of training data and then it summarizes the votes from different trees to decide the final class of the object under test.

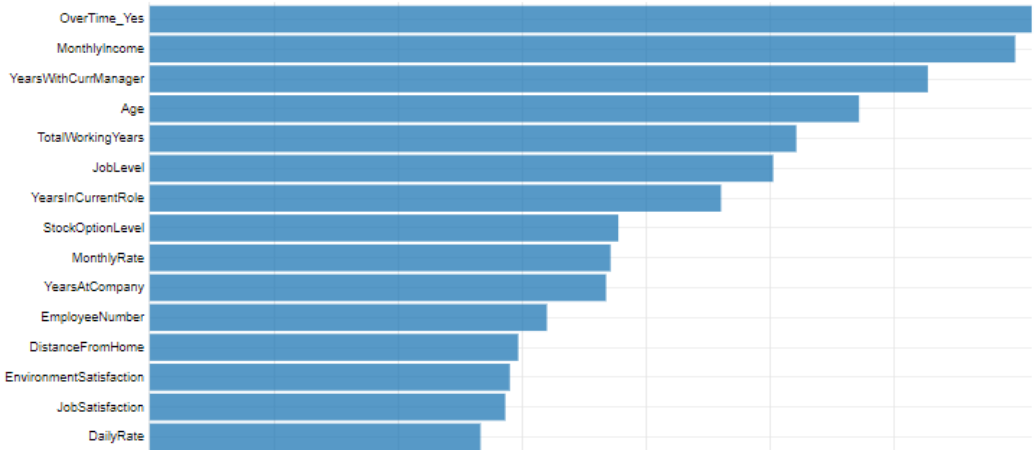

Fig 1. Feature Importance according to Random Forest model

Fig.1 represents the top 15 factors that lead to attrition of the employees according to the random forest model. The feature of employee working overtime is seen as the most important factor which leads to attrition among the employees.

3.1.2 Extreme Gradient boosting:

Gradient boosting or XGBoost [9] is a very popular machine learning algorithm and is extensively used in projects for structured or tabular data. It is a technique for regression and classification in which a prediction model is produced. The implementation of this algorithm is very fast and scalable. In gradient boosting, classification is done sequentially and the new predictors learn from the mistakes that are made by the previous predictors. As it learns from the mistakes done previously, it takes less time to classify and give the results. Another important thing to remember is that it is very important to provide the stopping criteria i.e. when the classifier or predictor should stop classifying so that there is no over-fitting.

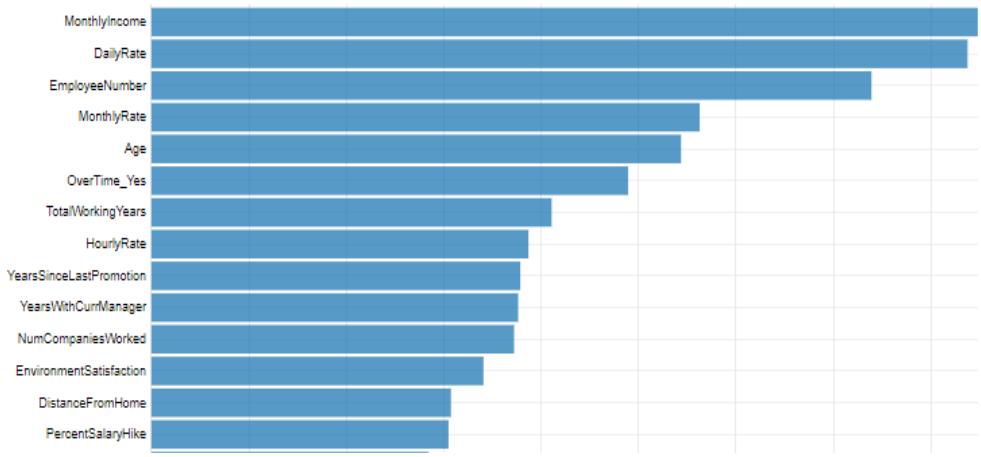

Fig 2. Feature Importance according to Extreme Gradient Boosting model

Fig. 2 represents the top 15 factors that lead to attrition among the employees according to the extreme gradient boosting model. The feature pertaining to the monthly income of employee is found to be a major factor that leads to attrition of the employees. A situation of Employee working hard and earning less on monthly basis can lead to attrition among the employees.

3.1.3 Support Vector Machines:

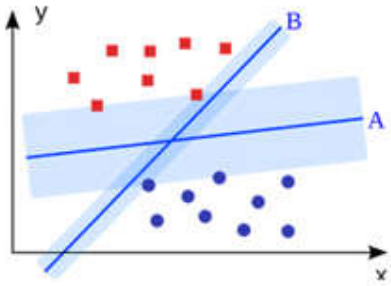

Fig 3. Support Vector Machines [10]

A Support Vector Machine (SVM) [10-11] is a classifier which is defined by a separating plane or a separating line. In 2D space, this algorithm generates an output which is a line that divides the data into two parts with different classes on the either side of the line (say for example class 0 and class 1). The implementation of the SVM algorithms is done using a kernel in practice. In linear SVM, the plane learns by transforming the problem using linear algebra. Optimization procedure is used to solve the SVM model.

3.1.4 Logistic Regression:

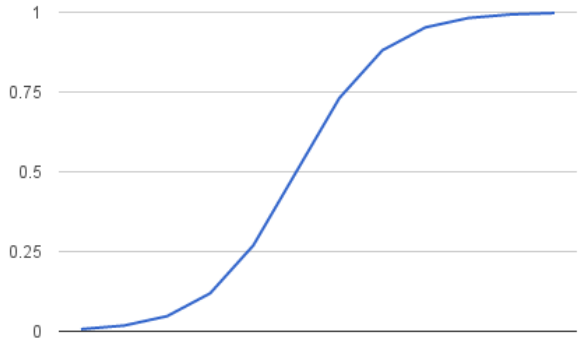

Fig 4. Logistic function [13]

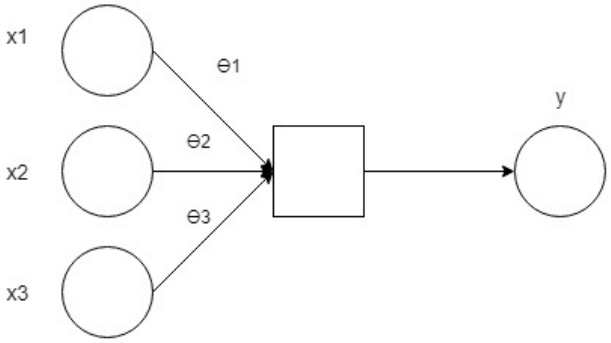

Fig 5. Logistic regression model

Logistic regression [12-13] is a statistical method which is used to analyse a dataset which consists of one or more independent variables that determine an outcome. The outcome has only two values i.e. either yes or no or other binary outcomes like 1 or 0, true or false. It was developed to estimate the likeliness of a binary response based on more than one independent feature. Logistic regression permits anyone to say that the existence of a risk factor enhances the probability of a given case. Like every other regression analyses, logistic regression is also a predictive analysis. In general, it is used for the description of data and to delineate the relationship between a dependent variable or feature and more than one nominal or ordinal independent variables.

4. Experimental Setup

This section discusses the flow of implementation of the project starting from the first step i.e. what the dataset is and what all features it contains. After that any preprocessing on the dataset is mentioned along with the evaluation criteria. The evaluation criteria states what criteria is used to assess the prediction of different techniques. The cloud service that is used for implementing the various algorithms is also mentioned.

4.1 Dataset:

To determine the most important features which lead to employee attrition, the IBM HR Analytics Employee Attrition dataset has been considered. The IBM HR analytics dataset is a fictional dataset created by IBM data scientists. The dataset takes into account 34 types of factors that may or may not account for attrition among employees. Some of the factors include Overtime of the employee, Job Satisfaction, Hourly Rate etc. The factors are not just restricted to the office work, but other factors such as work-life balance and marital status of the person are also taken into account. These factors help in the prediction of the attrition by using various classifiers on the given dataset. Thus, various factors are assessed and the factors strongly leading to employee attrition are discovered and analysed.

Table 1. Features in the dataset

AGE	JobInvolvement	RelationshipSatisfaction
Business Travel	JobLevel	StandardHours
Daily Rate	JobRole	StockOptionLevel
Department	JobSatisfaction	TotalWorkingYears
DistanceFromHome	MaritalStatus	TrainingTimesLastYear
Education	MonthlyIncome	WorkLifeBalance
EducationField	MonthlyRate	YearsAtCompany
EmployeeCount	NumCompaniesWorked	YearsInCurrentRole
EmployeeNumber	Over18	YearsSinceLastPromotion
EnvironmentSatisfaction	OverTime	YearsWithCurrManager
Gender	PercentSalaryHike	
HourlyRate	PerformanceRating	

4.2 Data Pre-Processing:

The dataset that is used needs to be pre-processed because of the presence of redundant attributes in it. Initially, data cleaning operation is performed where the redundant factors are determined and are not considered for the prediction of attrition in the employees. These redundant factors include Standard Hours, Employee count, Over18 which are either having the same values for all the employees or are completely unrelated to the prediction task. As part of the exploratory data analysis, the categorical factors are splitted

and are assigned values as 0 and 1 based on whether the factor is present or not. These assigned values assist in further classification based on that particular factor. Detailed analysis of the model leads to an uncovering of a class imbalance problem in the dataset. In most of the cases attrition was found to be absent, leading to class imbalance. Different statistical techniques like oversampling or undersampling have been used to treat the imbalances in data. In this project we have used an oversampling technique known as Synthetic Minority Over-sampling Technique (SMOTE) [16] to treat the imbalance that is present in the dataset.

4.3 Evaluation criteria:

For classification problems, Area Under Curve (AUC) can be trusted for performance measurement. It is considered as one of the most important evaluation metric for assessing any classification model's performance. This paper uses Area Under Curve (AUC) [14-15] as the evaluation criteria for the above mentioned models. Different models are implemented, tested and then evaluated based on the AUC score.

4.4 System Specification:

All the models are trained using the data from the IBM HR Analytics dataset as mentioned above. A split of 80-20 is used to divide the dataset into training data and test data. The different models are implemented and trained on Google Colab cloud service with Tesla K80 GPU.

5. Results

Area under the curve is the chosen evaluation metric. The table below shows the area under curve of different models:

Table 2. Comparison of AUC of different models

Models	Area Under Curve
SVM	0.72202
Random Forest	0.812168
Logistic Regression	0.831726
eXtreme Gradient Boosting	**0.845960**
Ensemble Average	0.855127

In the above table, different models are compared according to the chosen evaluation parameter i.e. AUC. SVM provides considerably low results as compared to other models which can be considered as a drawback that SVM is not so competent in tasks involving multi-class problems. Random forest and Logistic regression provides competent results but the drawback of Random forest is that it is based on bootstrap sampling which fails to take into account the rare personality/feature traits which are necessary to be taken into account for classification problems. Logistic regression is a good predictive model but the disadvantage of using logistic regression is that it works on independent variables which always possess the risk of over-fitting the model. extreme Gradient Boosting is seen to be providing the best result amongst the various chosen models with an AUC of 0.845960. Gradient boosting trees build sequentially. So because of that each new tree helps to correct the errors made by the previous trees. In this way, the iterations go forward and the errors made in the previous iterations are mitigated in the new one. Although the Gradient Boosted trees are prone to over-fitting, the parameters such as the shrinkage, depth of the trees and the number of trees can be fine-tuned to avoid the over-fitting of trees. Thus, Extreme Gradient Boosting proves to have a upper hand on the attrition prediction task as compared to other methods.

6. Conclusion

Thus, a detailed comparison and analysis of different methodologies has been made pertaining to the employee attrition problem. Extreme Gradient boosting method comes out to be the best method suitable for the problem whereas Logistic Regression and Random forest also prove to be competent enough to predict the attrition amongst the employees. Support Vector Machines lack behind by a significant amount as compared to other predictive models. The top 5 factors highly responsible for employee attrition are discovered which include overtime, monthly income, daily rate, age, and total working years.

References

[1] Rygielski, C., Wang, J., & Yen, D. C. (2002). Data mining techniques for customer relationship management. Technology in Society, 24(4), 483–502

[2] M.J.A. Berry, G.S. Linoff, Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management, Second ed.Wiley, New York, 2004.

[3] E.W.T. Ngai, L. Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: a literature review and classification, Expert Systems with Applications 36 (2) (2009) 2592–2602.

[4] Esmaieeli Sikaroudi, Amir & , RouzbehGhousi & Esmaieeli Sikaroudi, Ali. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). Journal of Industrial and Systems Engineering.

[5] Nagadevara, Vishnuprasad. (2018). Early Prediction of Employee Attrition in Software Companies-Application of Data Mining Techniques.

[6] Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38, 1999–2006.

[7] Prashant Gupta (18th May 2017)Decision Trees in Machine Learning [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [Accessed: 10-Oct-2018].

[8] Synced (25th Oct 2017) How Random Forest Algorithm Works in Machine Learning [Online]. Available: https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674 [Accessed: 12-Oct-2018].

[9] Jason Brownlee (17thAugust 2016) A Gentle Introduction to XGBoost for Applied Machine Learning [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ [Accessed: 14-Oct-2018].

[10] Savan Patel (3rd May 2017) SVM(Support Vector Machine) – Theory [Online]. Available: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 [Accessed: 16-Oct-2018].

[11] Rohith Gandhi (7th June) Support Vector Machine – Introduction to Machine Learning Algorithms [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [Accessed: 16-Oct-2018]

[12] Rohith Gandhi (28th May) Introduction to Machine Learning Algortihms: Logistic Regression [Online]. Available: https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36 [Accessed: 18-Oct-2018].

[13] Jason Brownlee (1st April 2016) Logistic Regression for Machine Learning [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/ [Accessed: 18-Oct-2017].

[14] Wuhan and Hubei, "The Problem of Area Under Curve", IEEE International Conference on Information Science and Technology, China (2012) March 23-25.

[15] Jocelyn D'Souza (15th March) Let's learn about AUC ROC Curve! [Online]. Available: https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152 [Accessed: 21-Oct-2018].

[16] Jason Brownlee (19th August 2015) 8 Tactics to Combat Imbalanced Classes in Machine Learning Dataset [Online]. Available: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ [Accessed: 19-Oct-2018].