IBM Attrition Prediction Analysis: Factors That Can Influence the Attrition Rate

Rong Fan[1, *]

[1]Department of Data Science, University of Rochester, Rochester, NY, United States

*Corresponding author: rfan7@u.rochester.com

Abstract. This paper aims at solving the high level of attrition rate at IBM. Data scientists would like to measure the relationships between the attrition rate and other descriptive information such as gender, age, working experience and position, to predict the attrition status and look for commons on those who have left or will probably leave the corporation in the future. Based on those predictions, appropriate summaries of their characteristics can be sketched for the manager in IBM to provide better working conditions to lower the rate. The main prediction method in this paper will be tree-based including simple decision tree and other advanced techniques, and the prediction results will be shown as scatter plots. While XGBoost shows the highest performance in this study, it will be selected as the prediction method to predict the attrition status of IBM employees. With such a model, IBM data analysts will be able to do further research on the reasons for attrition and thus lower the attrition rate.

Keywords: Attrition; Classification prediction; XGBoost.

1. Introduction

Attrition has been a serious problem for large corporations in current society. Even for the territory of education, the attrition of teachers has been a great problem which can lead to higher burden of workload as a result of shortage of teachers, based on the qualitative study done by Wushishi [1]. As the basic measurement of the loss of employees, it can lead to implications in staffing, employee morale, project costs, loss of experience, and a general hindrance to organizational growth. According to Valier, high attrition indicates intrinsic problems within the company and can lead to higher hiring costs, lost opportunities and lowered productivity, while replacing them with new employees will cost one half to twice the employee's annual salary [2]. Therefore, by far great efforts have been put into the study of attrition to solve this internally significant problem. A main focus is that researchers would like to come up with a universal solution to this issue so that all the companies will be able to release the severe influences. However, since every corporation has distinct situations, it seems that it will be impossible to find that solution. According to Dobhal and Nigam's journal, there is no universal attrition management solution and each company has to motivation system on compatility between organizational and individual goals, which may vary from company to company and from industry to industry [3]. For the BRO sector in India, based on the study conducted by Mishra and Solanki, employee engagement is important and their positive attitudes towards the job is the most important contributor [4]. More importantly, organizations should create an environment that fosters ample growth opportunities, appreciation for the work accomplished and a friendly cooperative atmosphere that makes an employee feel connected in every respect to the organization, according to Goswami and Jha's research paper [5]. Thus, based on such persuasive studies, this paper mainly focus on IBM employees. Using their information and the attrition status as the reponse variable, prediction models will be built to estimate whether the employees will leave the company for the further study of solutions to attrition.

Basically this study will deal with the dataset containing information of IBM employees, in which the attrition will be selected as the final response variable that indicates whether the corresponding employee will quit the company. Since it is a binary variable, classification methods will be adopted to fit the adequate models, in which logistic regression, decision tree, random forest and XGBoost will be used for prediction and the one with the highest performance will be chosen as the final model. Since the XGBoost presents the highest performance in this study, IBM attrition can be predicted

using this model, and it can also be utilized in further study about the underlying factors that have served as the trigger for the employees leaving their positions.

The study will simply follow a common sequence: looking for a dataset initially, preprocessing the data by label encoding and deleting missing values, presenting the data with plots, building the models, and displaying the results as both statistics and graphs. Besides, the application of the model used will also be included because there might be better ones which can potentially improve the performance and accuracy of the prediction model.

2. Data and Methodology

2.1 Data

The data is derived from Kaggle (https://www.kaggle.com/). The basic distributions of the numeric variables are shown in histograms in Figure 1. As shown, most employees are middle-aged with relatively lower salaries, and most of them have been in current positions in IBM for relatively fewer years, while they are involving in their jobs and were satisfied to them.

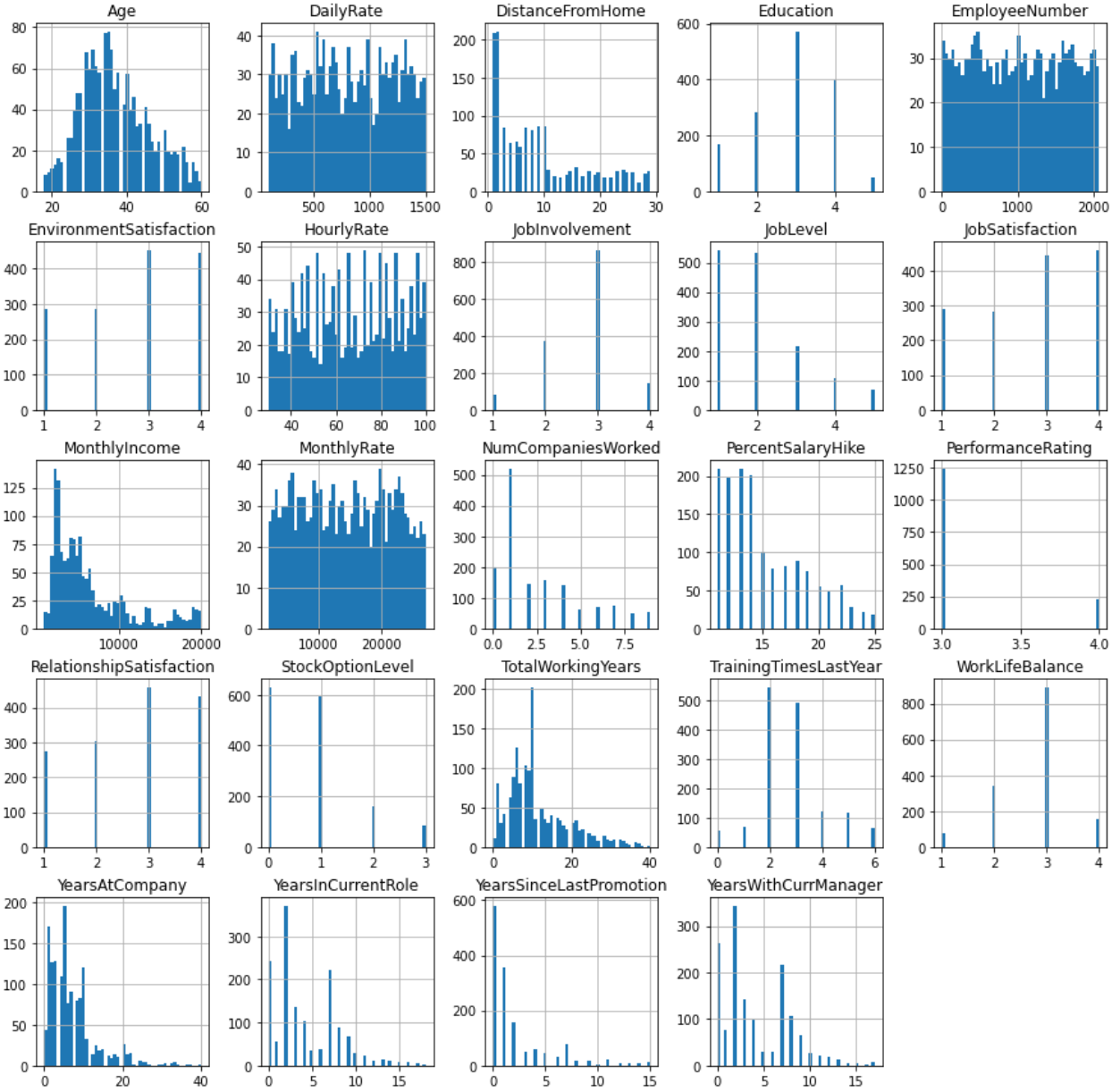

Fig. 1 Histograms for Numeric Variables

The mean, meaximum and mnimum values are shown in table 1. Similar to what has been shown in the histograms, most employees are middle-aged and have stayed in IBM for fewer years. Moreover, although the highest month income is 19999 and the average is appproximately 6502, the minimum of 1009 indicates that there are still a large number of employees who are receiving lower salaries. On the contrary, when evaluating their performances, most employees receive a grade of 3 and some of them receive 4.

Table 1. Descriptive Statistics for Variables

Variable	Mean	Maximum	Minimum
Age	36.9238	60	18
DailyRate	802.4857	1499	102
DistanceFromHome	9.1925	29	1
Education	2.9129	5	1
EmployeeNumber	1024.8653	2068	1
EnvironmentSatisfaction	2.7218	4	1
HourlyRate	65.8912	100	30
JobInvolvement	2.7299	4	1
JobLevel	2.0639	5	1
JobSatisfaction	2.7286	4	1
MonthlyIncome	6502.9313	19999	1009
MonthlyRate	14313.1034	26999	2094
NumCompaniesWorked	2.6932	9	0
PercentSalaryHike	15.2095	25	11
PerformanceRating	3.1537	4	3
RelationshipSatisfaction	2.7122	4	1
StockOptionLevel	0.7939	3	0
TotalWorkingYears	11.2796	40	0
TrainingTimesLastYear	2.7993	6	0
WorkLifeBalance	2.7612	4	1
YearsAtCompany	7.0082	40	0
YearsInCurrentRole	4.2293	18	0
YearsSinceLastPromotion	2.1878	15	0
YearsWithCurrentManager	4.1231	17	0

Since the response variable in this research is the attrition, for the categorical variables, the relationship of them with attrition is plotted as shown in Figures 2 to 6. In these bar plots, sketchy relationships between the attrition and other variables can be visually seen. Obviously more employees are leaving the corporation while some factors are influencing them while making decisions. For instance, for employees with distinct frequencies of business travels, the attrition rates also differ. According to the bars, employees who never traveled or traveled less are more likely to leave their positions compared to those who traveled frequently. Similarly, married employees and those who travelled less are also leaving more frequently. With these brief recognitions, model construction becomes simpler, but data analysts have to come up with further analysis for characteristics of people who have left and preventing measurements to lower the attrition rate in IBM. Also, label encoding has to be done to turn the categorical information into numbers to be used in the prediction models.

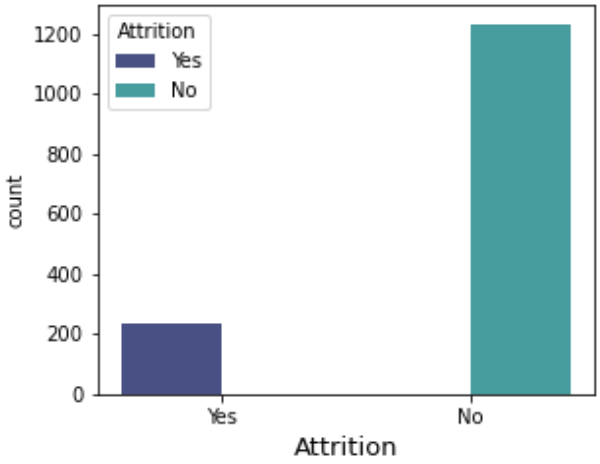

Fig. 2 Attrition

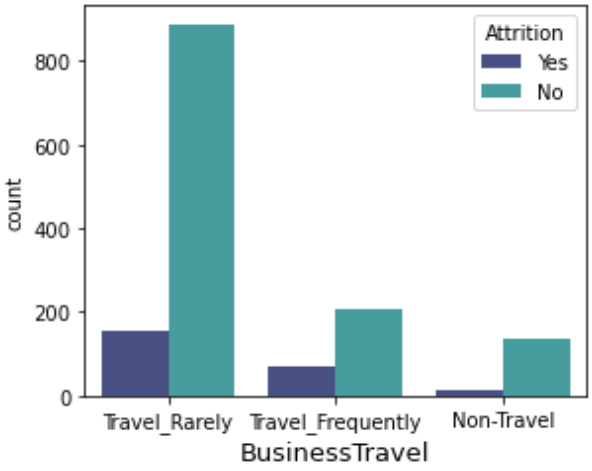

Fig. 3 Attrition with Business Travel

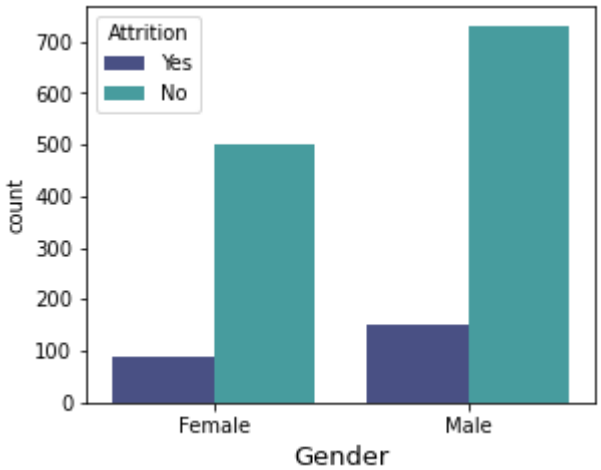

Fig. 4 Attrition with Gender

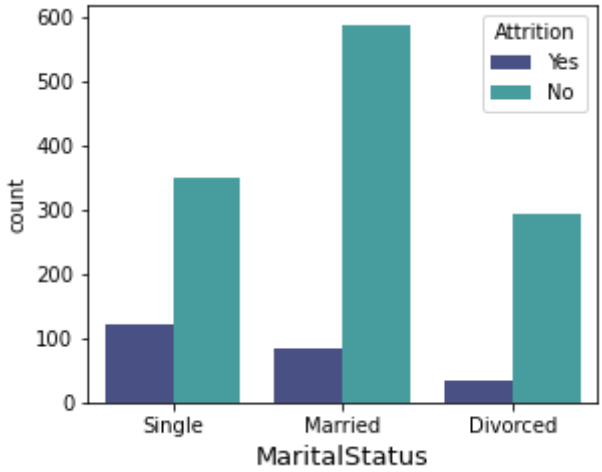

Fig. 5 Attrition with Marital Status

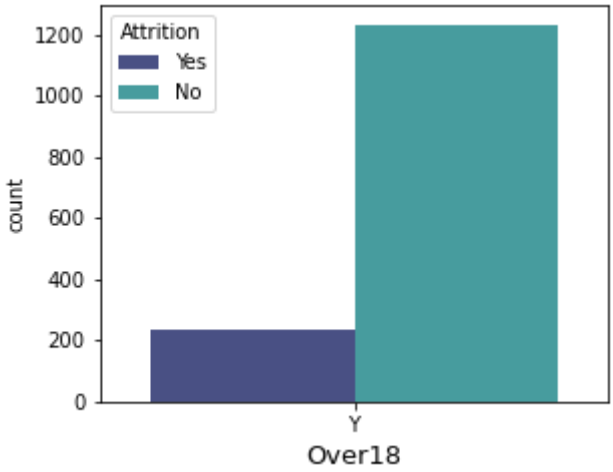

Fig. 6 Attrition with Over 18

2.2 Methodology

The main methodology adopted in this research is classification model. Since there are multiple kinds of models, four of them are chosen and tested to see their final performance in predicting the attrition status: decision tree, random forest, logistic regression and XGBoost. The response variable is the binomial attrition status and the independent ones are everything about every employee. There are personal information such as age, gender, marital status, education and over18, which is a binomial variable indicating whether the employee is over 18 years of age. Besides, there are information related to their positions such as department, income and job role, and there are information about the employee's past working performance such as total years in IBM, in current position and with the current manager. All these pieces of information gather to form a whole employee and some of them must have impacted the employee's mind for him/her to determine whether to leave IBM.

Logistic regression is the simplest method among the four models. It predicts the result by estimating a probability of the result being 1. The predicted result becomes 1 if the probability is greater than 0.5 and 0 otherwise. The prediction function is written as:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}} \tag{1}$$

where μ is the location paramater that makes p(μ) equal to 0.5 and s is the scale parameter that determines how the curve of the function spreads. The sigmoid curve of the prediction function is also sketched in Figure 7.

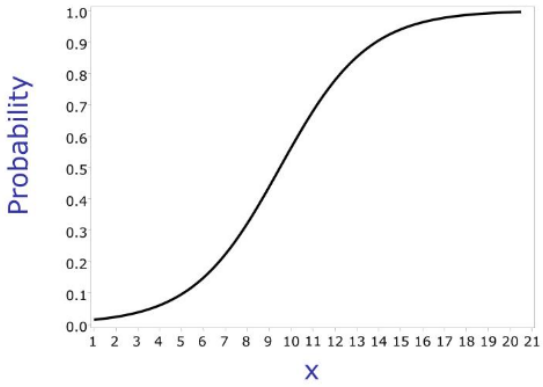

Fig. 7 Sigmoid Curve of Logistic Regression

In decision tree, a tree is built to classify the employees into two groups and each node in the tree represents a point at which a decision is made. Starting from the root node, several decisions are made to determine the classification result. This method uses the ID3 algorithm, Iterative Dichotomiser 3, to iteratively dichotomise the features into two or more groups at each step. Based on Sakkaf's explanation, ID3 algorithm uses a top-down greedy approach to build a decision tree, which means that it starts from the root node and at each iteration, it selects the best feature at present to build the node [6]. Further, decision tree can be used in both classfication and regression situations because it contains the CART algorithm, Classification And Regression Tree, to split the dataset into a decision tree. According to Arif's articles, the CART algorithm computes Gini Impurity to determine the predictors for each node. In classification problems, it calculates the statistics by subtracting the squares of the probabilities of the targets being 1 and 0, while in regression problems, it uses the least aquare approach to generate the residual sum of squares. By comparing these statistics, the algorithm finally comes up with the best predictor for each node and constructs the decision tree [7, 8].

Random forest is an advanced supervised learning model based on decision tree. It trains multiple trees in the fitting process and select the result chosen by most trees. In other words, every tree takes advantage of different decision methods to get the results, which can make the final result more precise compared to that gained from only one tree.

XGBoost, which stands for extreme gradient boosting, is the advanced version of gradient boosting. Gradient boosting is similar to random forest because it also builds several trees to get the result, but the difference is how they build the trees. Random forest builds parallel trees and derive the result by selecting the most common ones existed in the trees, while gradient boosting sets the result at first and builds new trees to achieve the result and reduce error. Using much more powerful computing power, XGBoost improves efficiency by building the trees parallelly instead of sequentially.

The models are evaluated using accuracy scores and mean square error and they are both derived from the fitted model using testing data as the input. Seven tenths of the data is used as training data to fit the model with train test split, and the rest is used to test the performances and gain the visualizations and the statistics.

3. Results

3.1 Test Results

The most important following step is to meticulously observe the results in which employees who have left will probably present unique characteristics. Therefore, data visualizations are significantly required in later analysis. Since every employee has multiple characteristics including their age, gender and working time, data analysts will usually conduct PCA, principal component analysis to decrease the dimension of the data to present them in 2-dimensional scatter plots. The actual test data is spread as shown in Figure 8 while the predicted ones are presented in Figures 9, 10, 11 and 12.

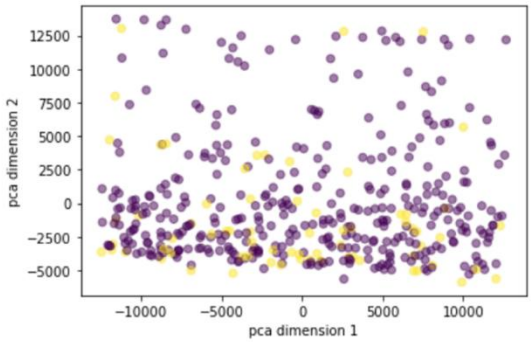

Fig. 8 Actual Distribution

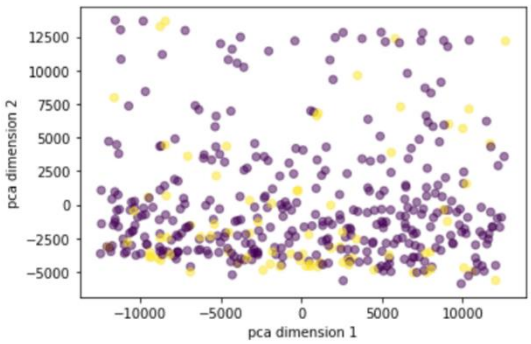

Fig. 9 Logistic Regression Distribution

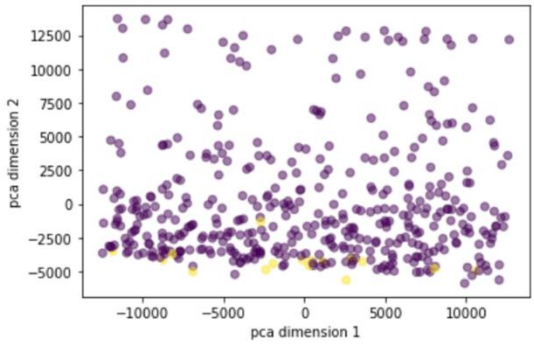

Fig. 10 Decision Tree Distribution

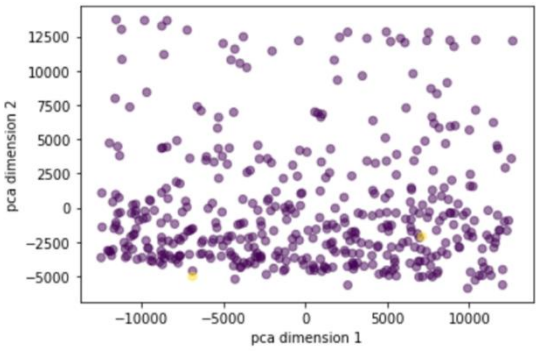

Fig. 11 Random Forest Distribution

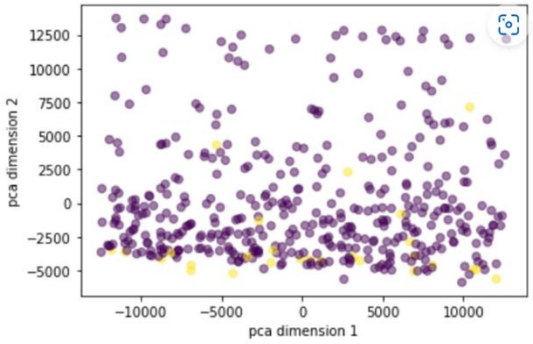

Fig. 12 XGBoost Distribution

According to the distributions, logistic regression implies the worst prediction functionality because almost all the observations are classified into the same class although the model has a relatively high accuracy. Compared to that, random forest improves a little by turning some of the classifications correct. In contrast, XGBoost and decision tree proves higher performance, but decision tree has great amounts of mistakenly classified predictions. Therefore, although the ratio of the employees who has left with those who has not looks better in decision tree, its accuracy becomes lower due to these errors. Therefore, the XGBoost model will be selected as the fittest predictor for employee attrition. The next step for IBM is to do further research on those correct predictions to find out the exact reasons for why they have left the company, hence lowering the attrition rate in the future by providing better service for its employees, which will also improve their working efficiency.

After visualizing the results in plots, it came out that XGBoost may be the best model to predict the attrition for IBM. This can be tested by doing model selection using different statistics including accuracy score and mean square error, which former measures how much the model matches the reality and latter measures how much error the model is creating. The estimation results are shown in table 2. The classification models predict the attrition of the IBM employees. By doing train test split before constructing the models, test statistics such as accuracy scores and mean square error scores can be derived through calculations. Accordingly, the XGBoost model shows the highest performance in this analysis with an accuracy score of 87.53% and an MSE score of 0.1247, corresponding to its scattered distribution shown before in the plots. This implies that the XGBoost model will be selected in later studies of attrition of IBM employees. With the prediction model, the corporation can have stronger power to prevent its employees from leaving their positions.

Table 2. Test Statistics for Four Models

Model	Accuracy	MSE
Decision Tree	76.64%	0.2336
Random Forest	85.49%	0.1542
Logistic Regression	84.58%	0.1542
XGBoost	87.53%	0.1247

3.2 XGBoost Application

In terms of XGBoost itself, it is a powerful machine learning tool for data scientists to derive the expected results that they want. It improves its performance by creating more trees than simple random forest using less resources. However, it has inevitable drawbacks while fitting the model. For this attrition problem, it works well on categorical data with binomial targets of 0's and 1's, while for other regression problems, it may collapse. As it creates trees as predictors, it does not provide a fixed formula for the dependent variables. Based on Mavuduru's publication, when large amounts of data appear, there can be observations which exceed what training data had provided [9]. For instance, if the training data ranges from 0 to 100 while the real-world application needs a prediction from 1000, the XGBoost predictor generally does not work at all. Besides, for categorial data, XGBoost also displays some drawbacks if compared to CatBoost and LightGBM according to Przybyla's opinions [10]. The most different characteristic of XGBoost from the others is that it needs preprocessing, which requires data scientists' more time and effort, because it only receives numeric values as inputs. Thus, for categorical inputs, label encoding is needed and this is a time consuming process because researchers have to memorize the actual representations beneath the meaningless labels, and analysis after model construction becomes also complicated due to these labels. Accordingly, these also consume more training time with bigger datasets. Nevertheless, accepting all these disadvantages, XGBoost offers overfitting-preventing functionality by adding in-put lasso regression and ridge regression regularization. Therefore, for the IBM attrition dataset used in this paper, XGBoost is indisputably the best predictor, and the performance may even improve by tuning inputting parameters.

4. Conclusions

Since attrition has been a serious problem for IBM, data analysts would like to estimate whether their employees are leaving IBM to avoid further increases in the attrition rate and keep the employees in their positions. With the data collected recording the employees' basic information, classification models can be built to estimate their attrition status of them, and by looking into the results, analysts will be able to figure out the reasons why some of the employees have left and provide better working conditions to prevent others from resignation. This paper takes advantage of four different classification models and selects XGBoost as the best predictor according to its highest accuracy and lowest MSE. Based on this study, XGBoost can be selected as the final predictor for the attrition status of IBM employees. With this model, IBM data analysts will eventually be able to investigate why some of the employees chose to leave the company and provide advice for the corporation to provide better working conditions to avoid such a high attrition rate. Besides, this can also improve working efficiency, which will probably lead to higher profit for IBM.

The next step for data analysts should be figuring out the specific things in common to lower the currently increasing attrition rate. For instance, if they find out that the employees who have left are those who are getting older and still did not promote from their initial positions, the company may probably pay more attention to these employees. If they have contributed to the company, they can likely earn a promotion.

Last but not least, the only concern left is whether there is a better model to solve the problem and whether the performance can be enhanced by subtly modifying the parameters, so further research is still required to seek better performance. Besides XGBoost, there might be more suitable machine learning algorithms to solve this problem. As has been mentioned in this article, XGBoost required the data analysts more time to complete the preprocessing, therefore potentially lowering the efficiency and accuracy. Instead, CatBoost can be tried because it does not need label encoding, so users can see the variables more clearly in their original names, but not labels. The change in the method is likely to improve the performance of the prediction model.

References

[1] Wushishi A.A., Fooi, F. S., Basri, R. et al. A qualitative study on the effects of teacher attrition. International Journal of Education and Literacy Studies, Australian International Academic, vol.2, no.1, pp.11-16, 2014.

[2] Valier, K., What is Attrition rate and what does it mean for your company? Jun 7, 2022, Oct 23, 2022. https://factorialhr.com/blog/attrition-rate/#attrition

[3] Dobhal, R., Dr. Nigam, A., Employee attrition and employee satisfaction: A study of H.R., performance appraisal & training practices in defense PSUs in India. IOSR Journal of Business and Management, vol. 20, pp. 1-27, 2018.

[4] Mishra, P., Solanki, N., A study of the factors leading to attrition in employees of BPO Industry with special focus on attitude towards job and employee engagement. Barkatullah University, vol. 6, no. 2, pp. 158-165, 2018.

[5] Goswami, B. K., Jha, S., Attrition issues and retention challenges of employees. International Journal of Scientific & Engineering Research, vol.3, no. 4, pp. 1-6, 2012.

[6] Sakkaf, Y., Decision trees for classification: Id3 Algorithm explained. Mar 31, 2020, Oct 23 2022. https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1.

[7] Arif, R., Classification in decision tree - a step by step cart (classification and regression tree). Apr 5, 2020, Oct 23, 2022. https://medium.com/analytics-vidhya/classification-in-decision-tree-a-step-by-step-cart-classification-and-regression-tree-8e5f5228b11e.

[8] Arif, R., Regrssion in decision tree - a step by step cart (classification and regression tree). May 2, 2020, Oct 23, 2022. https://medium.com/@arifromadhan19/regrssion-in-decision-tree-a-step-by-step-cart-classification-and-regression-tree-196c6ac9711e.

[9] Mavuduru, A., Why XGBoost can't solve all your problems. Nov 10, 2020, Oct 23, 2022. https://towardsdatascience.com/why-xgboost-cant-solve-all-your-problems-b5003a62d12a.

[10] Przybyla M., Stop using XGBoost. Jun 25, 2021, Oct 23, 2022. https://towardsdatascience.com/stop-using-xgboost-660ed6718845.