ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20210204181812

Original scientific paper

Churn Prediction of Employees Using Machine Learning Techniques

Nilasha Bandyopadhyay*, Anil Jadhav

Abstract: Employees are considered as the most valuable assets of any organization. Various policies have been introduced by the HR professionals to create a good working environment for them, but still, the rate of employees quitting the Technology Industry is quite high. Often the reason behind their early attrition could be due to company-related or personal issues, such as No satisfaction at the workplace, Fewer opportunities for learning, Undue Workload, Less Encouragement, and many others. This paper aims in discussing a structured way for predicting the churn rate of the employees by implementing various Classification techniques like SVM, Random Forest classifier, and Naives Bayes classifier. The performance of the classifiers was compared using metrics like Confusion Matrix, Recall, False Positive Rate, and Accuracy to determine the best model for the churn prediction. We found that among the models, the Random Forest classifier proved to be the best model for IT employee churn prediction. A Correlation Matrix was generated in the form of a heatmap to identify the important features that might impact the attrition rate.

Keywords: attrition; churn rate; classification techniques; confusion matrix; feature selection

1 INTRODUCTION

"Attrition" is not a new term for us anymore, as it has become an unavoidable situation in any business or organization, where staff and employees tend to leave due to their personal and professional circumstances. Further, this can cause a huge impact on any organization's growth curve if it is not given any attention, soon [1]. The major battle of employee attrition is right now being fought by the Technology Industries in India. Analysis from LinkedIn shows us that the software industry suffers from the highest turnover rates, which is about 13.2% compared to retail, entertainment, and professional industries. As per Maren Hogan, a talent acquisition expert, following points needs attention:

1. One-third of the new joiners quit, after six months in an organization.
2. After a week of working in a company, few decide on whether they want to continue staying there for the long term or not.
3. Also, a third of heads in companies having more than 100 employees are searching for new job opportunities [2].

Today's Millennial crowd in organizations is often identified as "job-hoppers", as they frequently change or quit their jobs to get to the next step of their career, as compared to the past generations. Rather than staying loyal to one company they often tend to search for better opportunities so that they can keep up in the era of digital progression. If we dig deep then we can find a distinct set of challenges faced by them like industry, proper recognition, communication, ethnicity, age, gender, etc. that drive the employees to leave a particular organization. Challenges faced by talent-hiring consultants are, sorting out the appropriate candidates through resumes and conversation, who will become the asset of the organization, and then if a person quits they need to repeat the entire hiring process. Every time hiring new talent and training them in current technologies involves a great amount of cost to the organization. Apart from this tangible expense, a fair amount of time we need to give the newly employed person to become a productive member of the project [3].

The Human Resource department of any organization generates a plethora of data related to employee's leave, promotion cycle, rewards, wages, various evaluations, conflicts, policies, and benefits. As a researcher, our work is to identify the correct parameters or areas where the employees face issues regularly at their workplace.

In this data-driven study, we will try to analyze the employee's data using some classification techniques and will provide quality insights and suggestions, so that the organization can retain them as well as develop them before it's too late. As HR professionals or managers our main focus should always be on an individual or certain groups of employees, especially towards their specific needs or their situation, then only it can further help an organization to grow more without losing good employees.

1.1 Research Objective

In our study, we will analyze the data of the Technology Professionals, especially their challenges that they face directly or indirectly at their workplace.

Main objectives of this study are:

1. To identify those challenges or input variables that have a huge impact on the employee's intention to leave the organization.
2. To accurately predict which employee will leave the organization in the next few years, using classification models.

2 LITERATURE REVIEW

An evidence-based study by Janet et al. (2017) has combined the already published scholarly reviewed literature on HR Analytics and has concentrated on answering major questions on HR Analytics, how it works, its outcome, and why there is a need for HR Analytics to flourish? They have stated that the interest of people in analytics in the HR domain for the past few years has gradually increased [4].

Later, the authors concluded that the inclusion of HR Analytics in various organizations is very low and proofs on this topic are scattered, hence suggested areas for future research. Many firms or departments say Marketing, Finance, Supply Chain Management organizations today draw insights from the huge data collected from the employees so that they can stay in this competition. The Human Resource department generates massive amounts of data on employee turnover, Return on Investments, and Cost per hire, but somewhere they still face a harder time relating these data with the organization's performance. They should create reports on past performances, administrative tasks, and generate compliance reports to understand the employee's contribution to the organization [5].

HR applications followed by today's organization can act as a mediator between planned HR practices in an organization and the positive outcomes of employees. Hence, Innocenti et al. (2012) have proposed a model that uses survey data that has been collected from over 6000 employees working in almost 37 Italian organizations, and the outcome variables are employee commitment and their job satisfaction. By using the maximum likelihood estimation method and calculating the correlations between different variables it was reported that, there is always a positive effect of experienced HR practices on both affective engagement towards organization and job satisfaction factors [6].

Line managers are considered the assets of that particular organization, so it's necessary to keep them engaged so that they can add value to any organization. Few semi-structured interviews were performed by Sana et al. (2016), to understand the experience and perceptions of the line managers on the level of support and help provided by the HR professionals of their organization [7]. Further, they have stated that the line managers have raised concerns and have suggested ideas for improving few areas like perceptions regarding policies, workload, inadequate training, and HR practices, which we need to pay attention to during any research on the factors related to employee attrition or turnover.

There have been several studies on identifying the parameters that play a role in job satisfaction of the employees and predicting the attrition rate. Many Data Analytics techniques and classification models have been used to predict turnover. In any organization, innovation can be seldom duplicated but once a group of productive employees leaves, that place cannot be replicated easily. So, to retain these employees and predict the turnover rate, a Neural Network, with a 10-fold Cross-Validation was designed for a small Midwest manufacturing company to a greater accuracy [8]. Among Layoffs, Discharges, Unavoidable separation, it was identified that voluntary separation from an organization always proves as the most difficult area because the particular organization loses its investment on talent to its competitors out there. On this same note, Fan et al. (2012) in their study, focused on why technology enterprises in Taiwan are unable to retain their talented employees and they have discussed ways so that the organizations can increase the competitiveness among themselves. Techniques like clustering analysis, hybrid artificial neural networks and other machine learning techniques were applied to forecast the patterns of employee's turnover rate [9]. Again, many Classification models have been used for prediction purposes, on a HR analytics dataset from Kaggle, an online community data site. Correlations between different attributes were evaluated by Sisodia et al. (2017) in their paper [10]. A comparison between different classifiers was drawn using parameters like Accuracy, Precision, True Positive Rate, F-Measure, and few others. Weighted TPR-TNR has been proposed as another performance metric to evaluate the performance of various classifiers, as it especially focuses on the imbalance ratio of any dataset and assigns different weights to TPR (Sensitivity) and TNR (Specificity), which are majorly considered while comparing ROC curve of any model. A mix of balanced and imbalanced datasets was used to evaluate the performance of 12 classifiers using the above metric [11].

To build and maintain a strong relationship between an organization and its employees, Hebbar et al. (2018), in their study initially implemented Logistic Regression on an IBM Employee Attrition dataset available in Kaggle just to get a basic idea, on which outcome group every individual falls [12]. Later on, a comparative study was done with SVM and Random Forest models, and determined the major characteristics of the dataset performing Exploratory Data Analysis and represented the data using different visualization.

With the same dataset (that has been used above), Synthetic Minority Oversampling Technique (SMOTE) was performed by Bhartiya et al. (2019) in their paper, to balance the imbalance dataset, because the count of the "Attrition" parameter with value 0 was greater than "Attrition" with a value of 1. The above technique is often used to generate synthetic data records for that class whose count is very less. Attributes like Gender, Education Field, and Performance Rate were visualized for Attrition parameters thus giving an idea on the relevant features. A comparison between the performance metrics of the classification models provided new insights on improving the work ethics [13].

With redundant data, predicting the correct features becomes a little challenging. So, a superior machine learning model or algorithm called XGBoost gives high accuracy in predicting the attrition rate with fewer running times. Jain et al. (2018) recommend XGBoost as a highly robust model, which easily handles noisy data in a huge dataset, and in their study, it gives an accuracy of about 90% on an online HR dataset [14]. Further, it suggests IT organizations to use this as a top priority, predictive model to identify those employees who are willing to leave in near future and their reasons behind that.

A very common issue that today's IT professionals face is stress disorders. Though organizations do offer a nice workplace environment and different activities or workshops to relieve this stress, still the risk increases among the employees. Various machine learning techniques like Boosting and Decision trees were implemented by Reddy et al. (2018) in their study, and have determined that data on family history of illness, gender and health benefits provided by employers plays an important role in evaluating this type of risks [15]. Ensemble method gave the highest degree of accuracy and precision compared to Random Forest. General characteristics like having peers to work with and the financial needs of the employees become critical factors for those who are working for a longer tenure in any business or organization. So, for the hospitality industry in the USA, Self et al. (2011) attempted a qualitative study on identifying various factors that might impact an employee's decision to stay back in a company. By analyzing the interview transcripts that were obtained after an in-depth process, four factors were identified: Strong Responsibility towards the company, Financial Requirement, Proper Job Description, and Peers at the workplace has a positive effect on employees [16].

One of the challenges that the big organizations are facing is, motivating their employees and investing in them for their further development. Understanding the importance of investing in employee development and its final results, is very much needed by the organization. A model proposed by Lee et al. (2003) gives us an interrelationship between perceived investments and other job attitudes and the employee’s plan to quit an organization. Factor analysis and Exploratory analysis were conducted for assessing the dimensionality and their insights, respectively. Results suggest that the more the employer spends resources on the development of their employees, the more they will be satisfied at their workplace, hence reducing the possibility of an employee quitting his or her job in that organization [17].

Burnett et al. (2019) propose a few topics on which one can use modern technology or tools to measure both employee engagement and the other HRM practices which can improve the same [18]. Different emotional states of employees affect their engagement at the workplace, either directly or indirectly. Further, they have pointed out that to improve on engagements we need to concentrate on three different levels: individual, team and organizational level and have suggested that with the real-time feedback from employees and rigorous research and analysis on the data will help the HRM department to understand the importance of employee engagement in their respective organization.

So, to stay in this competitive market, these technology industries need to continuously evolve in terms of skills and should be ready to embrace the ever-changing products and services. Even employees make themselves proficient in the new skills or technologies and try to search for better job opportunities outside. An analytics-driven approach can help organizations to overcome the situation. Combining the historic record of skills of each employee present in the HR database with the predictive models, Ramamurthy et al. (2015) have proposed an approach that evaluates a set of skills [19]. The algorithm in their study will provide a list of skills to some individuals, where they will fill in their target skills, helping business leaders to find potential candidates and will provide re-skilling offers to them.

One can go for Sentiment Analysis to determine the factors affecting employee retention, and organizations can use these models to understand the concepts of People Analytics. A conceptual study was done to identify key indicators to assess the human factors. Six important areas, like performance leadership, employee engagement, learning, workplace dynamics, and overall organizational development have used sentiment analysis to evaluate various insights. The Enron email corpus test case was incorporated to explain how we can predict the digital footprints. Further encourages implementing various data mining techniques or models to analyze the real-time data for predicting more accurate human factor patterns [20]. In addition to this, often interpersonal environment factors provide insights about employee development in any organization. Liu et al. (2019) in their study have concentrated on a state-owned enterprise in China, extracted the related features, and statistically analyzed the correlation between employee development in organizations and their interpersonal environment. The results of the predictive model prove that colleagues and classmates have a great impact on the growth of employees in their respective workplaces [21].

2.1 Research Gap

After reviewing the existing work, it was observed that many of the studies were following secondary data which is a HR analytics dataset available in an open-source dataset site, to predict employee turnover using Data Mining Techniques. The attributes that they have considered in their study are the generic parameters related to any employee who has already left the organization. Today, if we discuss with the IT professionals, we will get to know that they still face a set of challenges, both at their workplace and in their personal life which results in early attrition. This set of challenges often goes overlooked in this industry by the HR executives.

Every new employee who gets recruited might face a different set of challenges while working. So, analyzing the data of those employees who have already left the organization might not give us the features that apply to the new joiners. Rather, we need to interact with them frequently or take their feedback on a real-time basis, just to get the actual data related to their challenges, like Recognition, Challenging work, Scope of Development, Satisfaction Level, Unhealthy work ethics and Impact on them of their peers leaving an organization. For this reason, we are using primary data in our study that has been collected from employees working in various IT industries.

We need to concentrate more on discussing what they want for their betterment in this organization. Then start predicting who might leave within a couple of years, post this we can offer them proper opportunities. This will not only encourage the employees but will help the organization in retaining its talent.

3 RESEARCH METHODOLOGY

This study is focused on employees from a specific age group that is from 20 to 39 years old, who are considered to be the major contributors to the highest turnover of any organization. In this research, surveys were conducted to get the raw data from the employees, which is first pre-processed, and then analysis was done to derive meaningful insights.

A questionnaire consisting of 35 questions was circulated among 200 employees and the response rate was around 79%. Among these responses, 83 were male and 75 were female employees. 80% of these employees had working experience of 4 years or less, the remaining 20% had an experience that varies from more than 4 years to 13 years. This survey had combinations of few open and close-ended questions, which includes a Likert scale and few dichotomous answer types. This will help us understand the actual perception of the employees regarding the organization or employer.

The entire questionnaire was designed based on our detailed review of the previous work that has been done by other researchers in this topic and our discussions with a few experts who are involved in the technology industries. Further, these questions have been divided into 5 sections, like Individual Beliefs, Management and Team, Engagement and Encouragement, Talent Development, Organisation and Leadership, to get an overall idea of the employees towards different verticals of an organization.

We are implementing and analyzing a few classification models in R studio.

3.1 Input Data Set

The data collected includes 11 attributes for each employee. The target variable "Quit_in_2years" consists of three classes, they are: "Maybe", "No" and "Yes", thus our study is a multi-class classification. Tab. 1 gives us the details on the attributes that will be used in our study:

Table 1 Dataset Attributes

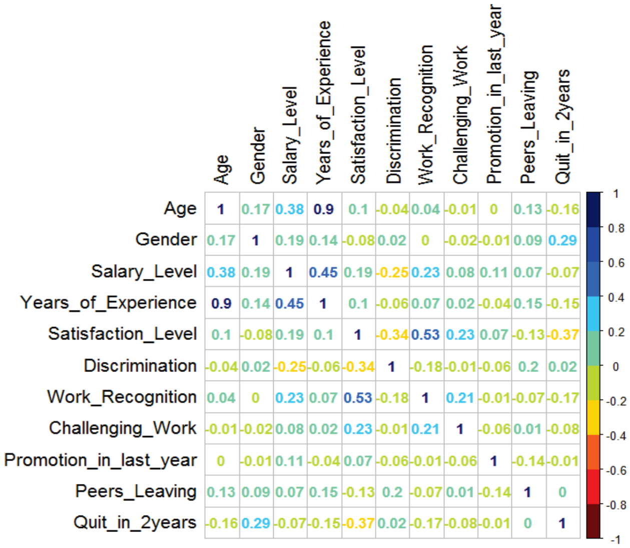
Sl. No.	Attributes	Data types	Description
1	Age	Numeric	Age of the employees.
2	Gender	Categorical	Gender of the employees
3	Salary_Level	Numeric	Salary window under which the employees fall.
4	Years_of_Experience	Numeric	For how many years that employee is associated with that organisation
5	Satisfaction_Level	Numeric	Degree of satisfaction of the employee at their workplace
6	Discrimination	Numeric	Any discrimination faced based on age, gender or ethnicity
7	Work_Recognition	Numeric	Been given proper recognition of their work in their team or not
8	Challenging_Work	Numeric	To what degree the Employees feel challenged with their daily work
9	Promotion_in_last_year	Numeric	Got promotion in last one year
10	Peers_Leaving	Categorical	To what degree does good colleagues or friends leaving organisation affects them
11	Quit_in_2years	Categorical	Their plan to leave the organisation in 2 years

3.2 Data Pre-Processing

Among the 11 attributes, "Gender", "Peers_Leaving" and the target variable, "Quit_in_2years" are categorical data types. So, to determine the impact of the above predictors on the target variable and evaluate the correlations among the attributes, the categorical fields were converted to numeric values. For example, "Female" was denoted by 1 and "Male" as 2. Under "Peers_Leaving" there were three categories, where "Yes" and "No" were given 1 and 0, respectively, while "Maybe" was denoted as 0.5. Similarly, the values of "Maybe", "Yes" and "No" for the target variable were denoted as 1, 2 and 3 respectively.

Though the null values in the dataset were really less, it was chosen to be replaced by the mean of the whole column rather than dropping the whole entry. To summarize the whole data, and to determine how close these variables have a linear relationship among themselves, we plotted a correlation matrix. This gives us an idea of identifying the features which have weak and strong dependencies.

For example, in Fig. 1, the darkest blue on the scale means there is a positive correlation among the attributes, whereas the dark red means a negative correlation. In the above figure, it can be observed that there is a stronger relationship between "Age" and "Years_of_Experience", again "Satisfaction_Level" and "Work Recognition has a positive correlation, with a coefficient of 0.53. The rest of the variables do not have a strong consistent relationship with each other. We observe that there are Negative Coefficients in the above matrix, this indicates that if the value of one attribute increases then the value of the other attribute will tend to decrease.



	Age	Gender	Salary_Level	Years_of_Experience	Satisfaction_Level	Discrimination	Work_Recognition	Challenging_Work	Promotion_in_last_year	Peers_Leaving	Quit_in_2years
Age	1	0.17	0.38	0.9	0.1	-0.04	0.04	-0.01	0	0.13	-0.16
Gender	0.17	1	0.19	0.14	-0.08	0.02	0	-0.02	-0.01	0.09	0.29
Salary_Level	0.38	0.19	1	0.45	0.19	-0.25	0.23	0.08	0.11	0.07	-0.07
Years_of_Experience	0.9	0.14	0.45	1	0.1	-0.06	0.07	0.02	-0.04	0.15	-0.15
Satisfaction_Level	0.1	-0.08	0.19	0.1	1	-0.34	0.53	0.23	0.07	-0.13	-0.37
Discrimination	-0.04	0.02	-0.25	-0.06	-0.34	1	-0.18	-0.01	-0.06	0.2	0.02
Work_Recognition	0.04	0	0.23	0.07	0.53	-0.18	1	0.21	-0.01	-0.07	-0.17
Challenging_Work	-0.01	-0.02	0.08	0.02	0.23	-0.01	0.21	1	-0.06	0.01	-0.08
Promotion_in_last_year	0	-0.01	0.11	-0.04	0.07	-0.06	-0.01	-0.06	1	-0.14	-0.01
Peers_Leaving	0.13	0.09	0.07	0.15	-0.13	0.2	-0.07	0.01	-0.14	1	0
Quit_in_2years	-0.16	0.29	-0.07	-0.15	-0.37	0.02	-0.17	-0.08	-0.01	0	1

Figure 1 Correlation Matrix of the attributes

3.3 Feature Selection and Ranking

This approach helps in recognizing the correct features in any dataset, where we can easily differentiate the features that play a significant role in predicting employee's intention to leave in the next 2 years, from the other features. Further, it will help in building a reliable model, with greater accuracy. Here, an R package known as "caret", is being used which will automatically give us a report on the importance and relevance of the attributes in our dataset and will help in ranking those features.

So for the feature selection process, RFE (Recursive Feature Elimination) is chosen, which is majorly used with SVM to continuously build a model and simultaneously remove those features that have low weights and discover the optimal number of features. The algorithm is configured to explore all possible subsets of the attributes. Next, to specify ranks to the feature by importance, a method known as LVQ (Linear Vector Quantization) was used, which is a form of ANN (Artificial Neural Network) algorithm and allows us to choose the training instances and learn what those instances should look like.

In Fig. 2, we have ranked all the features as per the target classes. So, it can be inferred that among all the 11 features, "Satisfaction_Level", "Salary_Level", "Work_Recognition", "Gender" and "Challenging_Work" are the top 5 challenges that have a huge effect on the target variable, that is, "Quit_in_2 years". Whereas, "Promotion_in_last_year" and "Peers_Leaving" have the least impact on the employee's decision on leaving the organization in the future.

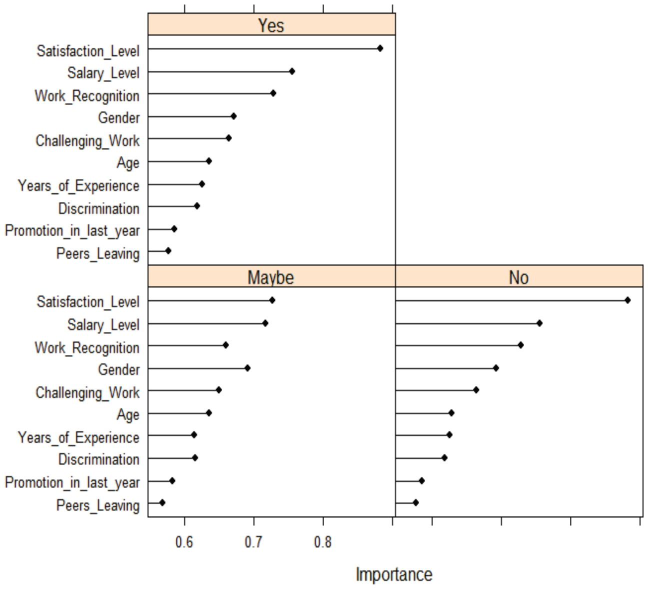


Figure 2 Ranks of the attributes

4 MODELS AND IMPLEMENTATION

In this research, three classification models were used to predict, whether a particular employee will leave the organization or not, based on the challenges he or she is facing currently at the workplace. Here, the classifiers that we are going to implement are the Random Forest classifier, SVM (Support Vector Machine), and Naive Bayes.

As per our research framework, after preprocessing the data, it was split into two parts, that is, train and test dataset in the 70:30 ratio. Trained our classification models by passing the training dataset and then evaluated the most efficient model by predicting the target value using the test dataset.

Our study on analyzing the performance metrics of the models has been bifurcated into two cases.
Case 1: Includes all the three classes of the Target Variable.
Case 2: Here we are including only two classes, that is, "Yes" and "No" of the Target Variable.

56 Employees who are still in the dilemma of whether they will leave their organization or not might affect the accuracy of the model. Hence, we removed them in Case 2 and analyzed the performance metrics.

4.1 Support Vector Machine

It comes under supervised learning techniques, majorly used for classification of data but is often implemented for regression problem statements. In this technique, the data points are separated from each other by a line or a hyperplane, and this division between the two sides categorizes the whole data sets into two or more classes. The space between the two classes is also known as margin, and this should be as large as possible so that we can reduce the error while classification. Package "e1071" is used for the implementation of the said model.

Tab. 2 gives us the Confusion Matrix of SVM, which includes all the classes of the Target variable whereas, Tab. 3 represents the Confusion Matrix for only two classes.

4.1.1 Case 1: Including all the three classes of the Target Variable

Table 2 Confusion Matrix of Test Dataset

		Actual		
		Maybe	No	Yes
Predicted	Maybe	9	2	8
	No	1	6	0
	Yes	5	3	14

4.1.2 Case 2: Including only two classes of the Target Variable (without "Maybe")

Table 3 Confusion Matrix of Test Dataset

		Actual	
		Yes	No
Predicted	Yes	14	5
	No	4	8

4.2 Naive Bayes Classifier

The crux of this classification method is based on the famous Bayes Theorem. It assumes that a particular feature or attribute in a class is independent of the existence of any other feature. The model is easy to build and is particularly useful if we have a huge dataset. With its simplicity in the model, Naive Bayes can outperform other sophisticated classification models for multi-class prediction. Below Tab. 4 and Tab. 5 are the Confusion Matrices for the above model for two different cases that we are considering in our study.

4.2.1 Case 1: Including all the three classes of the Target Variable

Table 4 Confusion Matrix of Test Dataset

		Actual		
		Maybe	No	Yes
Predicted	Maybe	10	2	6
	No	0	7	0
	Yes	5	2	16

4.2.2 Case 2: Including only two classes of the Target Variable (without "Maybe")

Table 5 Confusion Matrix of Test Dataset

		Actual	
		Yes	No
Predicted	Yes	16	5
	No	2	8

4.3 Random Forest Classifier

This model is an ensemble tree-based learning technique. Rather than using a single decision tree for classification of the data, it uses a set of decision trees that randomly selects subsets of data and train the model. Voting will be performed on the predictions from each of these trees and finally, the best solution will be selected. This method helps reduce the overfitting by averaging the results, as compared to traditional decision trees.

For the implementation of the classifier, a package called "randomForest" is used in our study. We can observe the values of predicted and actual instances from Tab. 6 and Tab. 7.

4.3.1 Case 1: Including all the three classes of the Target Variable

Table 6 Confusion Matrix of Test Dataset

		Actual		
		Maybe	No	Yes
Predicted	Maybe	12	1	5
	No	0	7	2
	Yes	3	3	15

4.3.2 Case 2: Including only two classes of the Target Variable (without "Maybe")

Table 7 Confusion Matrix of Test Dataset

		Actual	
		Yes	No
Predicted	Yes	15	4
	No	3	9

5 RESULTS AND DISCUSSION

So, to choose the best classifier for this study, we are comparing the existing performance metrics, say Model's Accuracy, Recall, Specificity, Precision, F-Measure, Area Under Curve (AUC) and another metrics that we are considering is Weighted TPR-TNR. For comparing the results of multi-class classification we are using the Macro Average Method for parameters like Recall (Sensitivity), Specificity, Precision, and F-Measure. This method helps in determining the performance of the overall system. As our data is a balanced dataset, we are using this method to calculate the average of the values that we obtained for each class.

As per our problem statement, we are mainly concerned with the people leaving the organization, thus to acquire complete knowledge to overcome this, parameters like Recall and AUC play a huge role along with the accuracy of the models.

5.1 Case 1: Comparing the performance metrics for all the Target Variable classes

Table 8 Final Results of the Classifiers

Metrics	SVM	Naive Bayes	Random Forest
Accuracy	60.42%	68.75%	70.83%
Recall or TPR or Sensitivity	0.5939	0.6768	0.7061
FNR	0.4061	0.3232	0.2939
TNR or Specificity	0.7874	0.8294	0.8445
FPR	0.2125	0.1705	0.1555
Precision	0.6558	0.7504	0.7196
F-Measure	0.6108	0.6983	0.70833
Weighted TPR-TNR	0.6584	0.7277	0.7522
AUC	61.64%	67.31%	73.48%

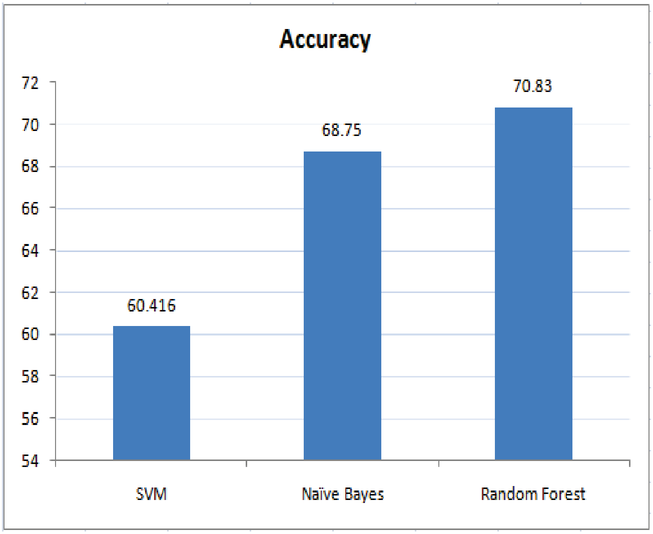


Figure 3 Accuracy of all the Models

From the above graph, we observe that the Random Forest classifier has achieved a far better prediction accuracy of 70.83% when compared to other classifiers.

Simultaneously, one must look for the Recall and Precision value apart from the model's accuracy. From Tab. 8 we can see that for Random Forest classifier the Recall value has increased but the Precision value is slightly less than Naive Bayes. Values of weighted TPR-TNR are the highest in the case of Random Forest than the other two models.

ROC curve is a trade-off between sensitivity and specificity, where the curve of a perfect classifier should have the highest Recall (True Positive Rate) with the lowest False Positive Rate. So, to summarize the performance of the classifiers we take the calculated area under the ROC curve into consideration, which is also known as AUC. So, the higher the AUC, the greater will be the accuracy of the model.

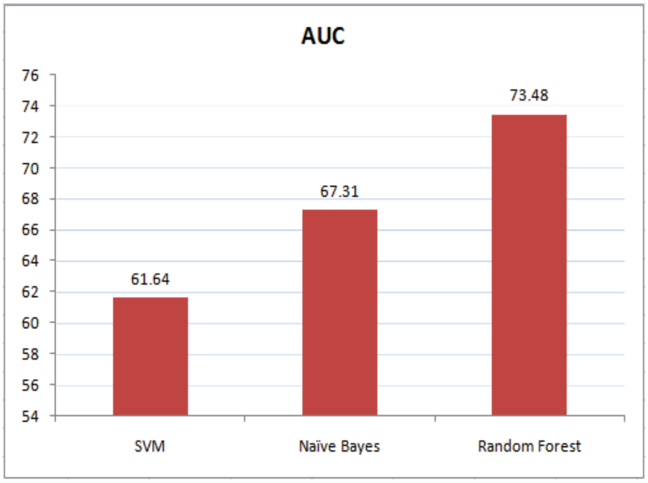


Figure 4 Comparing the AUC of the Models

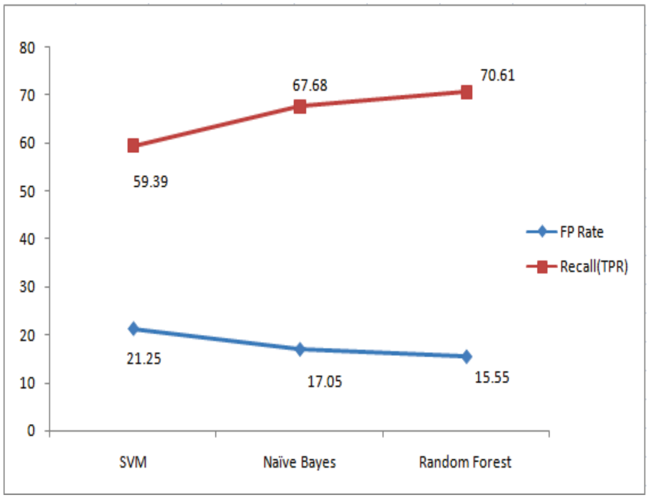


Figure 5 Comparison of TPR and FPR

With the highest AUC and lowest False positive rate, the Random Forest classifier stands out from the rest of the models.

5.2 Case 2: Comparing the performance metrics for only two Target Variable classes (without "Maybe")

Compared to Case 1, it can be observed that the accuracy of each model has increased by quite a percentage after excluding those employees who still had some difficulty in deciding on leaving their organization in the next two years. Both Naive Bayes classifier and Random Forest classifier have obtained an accuracy of 77.42%. Recall and F-Measure value of Naives Bayes is greater than the other two models, whereas if we observe Tab. 9, we can state that the Precision and weighted TPR-TNR for Random Forest classifier has increased, compared to SVM and Naive Bayes.

Table 9 Final results of the Classifiers

Metrics	SVM	Naive Bayes	Random Forest
Accuracy	70.97%	77.42%	77.42%
Recall or TPR or Sensitivity	0.7778	0.8889	0.8333
FNR	0.2222	0.1111	0.1667
TNR or Specificity	0.6154	0.6154	0.6923
FPR	0.3846	0.3846	0.3077
Precision	0.7368	0.7619	0.7895
F-Measure	0.7568	0.8205	0.8108
Weighted TPR-TNR	0.6834	0.7301	0.7513
AUC	69.66%	75.21%	76.28%

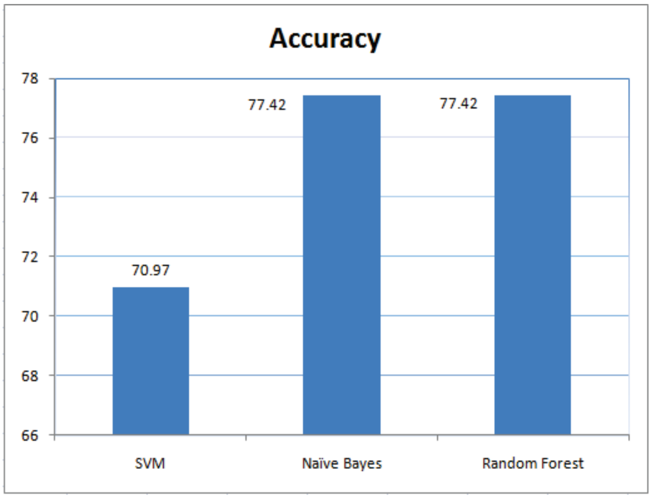


Figure 6 Accuracy of all the Models

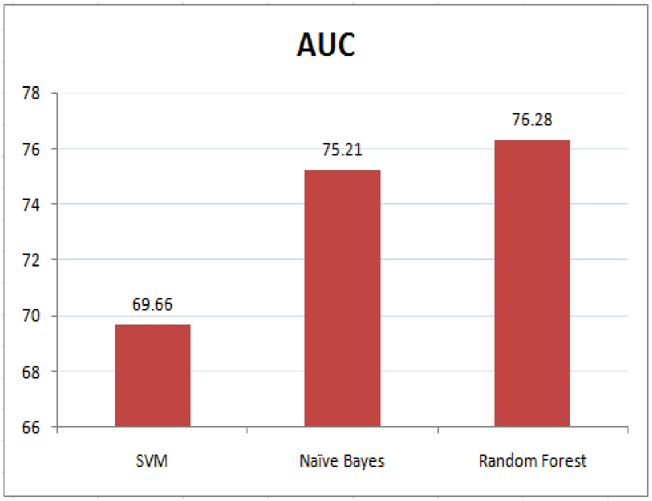


Figure 7 Comparing the AUC of the Models

Considering the area under the ROC curve we can see from the above figure, that the Random Forest classifier still has a lead of 6.62% from SVM and 1.07% from Naive Bayes classifier.

So, it can be stated that with the lowest False Positive Rate and highest AUC, Random Forest Classifier proves to be a good model in this case as well.

Adding to this, as our study focuses more on predicting employees who might leave in near future, we should never forget the False Negatives in this case. That is, those employees who are planning to leave but the model somehow does not predict them correctly. We need to identify these False Negatives and should find ways to reduce this.

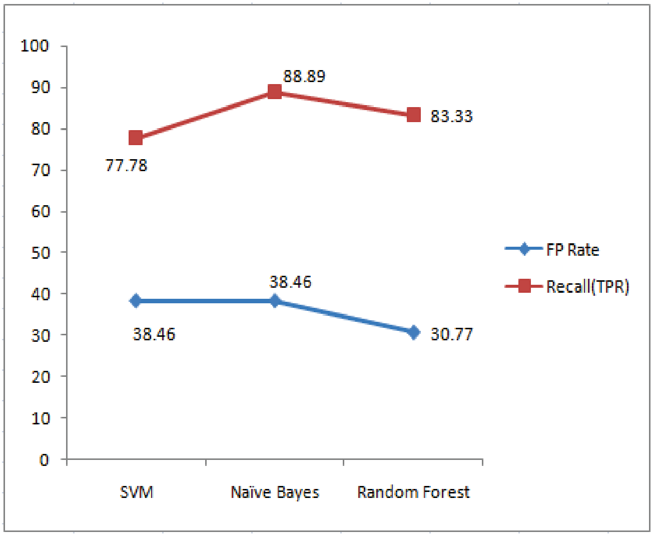


Figure 8 Comparison of TPR and FPR

6 CONCLUSION

As per our discussion above, employees leaving organization has a major impact on the development of these technology organizations. Often the challenges or issues faced by the employees at the workplace or in their personal life have a great impact on their early attrition from the organization.

In our study we have identified the important factors that affect the employees, resulting in future attrition. To help with the analysis, data were collected from professionals working in IT industries. Majority of the attributes we considered did not have a significant correlation with each other. Further to get the top features that have a positive impact on employees, a method called RFE (Recursive Feature Elimination) was chosen for variable selection. This method helped in removing redundant and less important variables and highlighted features which has more impact on the target attribute. In addition to this, LVQ (Linear Vector Quantization) was introduced to rank all the attributes as per their importance.

Secondly, our goal was to accurately predict those employees who are planning to leave the organization in the next 2 years, using a few classification models. Techniques like SVM, Naive Bayes, and Random Forest classifiers have been implemented in this study. So, to analyze the pattern we bifurcated our analysis into two cases. In the first case, we considered all the Target Variable classes but for the second case, we removed those employees who still had their doubts about leaving a particular organization in near future. Observing the results, we conclude that the models implemented in Case 2 gave good accuracy as compared to Case 1. The most efficient model in our study was the Random Forest classifier giving us the highest accuracy and Recall value when compared to the other models.

Apart from getting a good raise and promotion, there have been other kinds of challenges faced by today's talent, which the HR executives or managers of the project need to take care. In the future direction, this study can be further extended, by including attributes, like Scope of Development, Views on workload distribution, Career goal discussion and Issues on unhealthy work ethics.

Organizing frequent feedback or a one to one interview on the organization policies can help HR understand the expectations. In our study we had a limited data size of 158 entries, it is suggested that with more data points and features we can achieve higher accuracy from these models.

Notice

This paper was presented at IC2ST-2021 – International Conference on Convergence of Smart Technologies. This conference was organized in Pune, India by Aspire Research Foundation, January 9-10, 2021. The paper will not be published anywhere else.

7 REFERENCES

[1] Basu Mallick, C. (May 2020). *What is Employee Attrition? Definition, Attrition Rate, Factors and Reduction Best Practices*. https://hr.toolbox.com/articles/what-is-attrition-complete-guide

[2] (June 2019). *Tech industry battles highest attrition rate in the world - and it's costly.* https://www.viglobal.com/2018/06/13/tech-industry-battles-highest-attrition-rate-in-the-world-and-its-costly/

[3] Yadav, S., Jain, A., & Singh, D. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. *2018 IEEE 8th International Advance Computing Conference (IACC)*, 349-354. https://doi.org/10.1109/IADCC.2018.8692137

[4] Marler, J. H. & Boudreau, J. (2017). An evidence-based review of HR Analytics. *The International Journal of Human Resource Management, 28*, 3-26. https://doi.org/10.1080/09585192.2016.1244699

[5] Harris, J., Craig, E., & Light, D. (2011). Talent and analytics: new approaches, higher ROI. *Journal of Business Strategy, 32*, 4-13. https://doi.org/10.1108/02756661111180087

[6] Innocenti, L. & Peluso, A., M. & Pilati, M. (2012). The Interplay between HR Practices and Perceived Behavioural Integrity in Determining Positive Employee Outcomes. *Journal of Change Management*, 12. https://doi.org/10.1080/14697017.2012.728763

[7] Anwaar, S., Nadeem, A., & Hassan, M. (2016). Critical assessment of the impact of HR strategies on employees' performance. *Cogent Business & Management, 3*. https://doi.org/10.1080/23311975.2016.1245939

[8] Sexton, R., McMurtrey, S., Michalopoulos, J., & Smith, A.M. (2005). Employee turnover: a neural network solution. *Comput. Oper. Res., 32*, 2635-2651. https://doi.org/10.1016/j.cor.2004.06.022

[9] Fan, C., Fan, P., Chan, T., & Chang, S. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Syst. Appl., 39*, 8844-8851. https://doi.org/10.1016/j.eswa.2012.02.005

[10] Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017). Evaluation of machine learning models for employee churn prediction. *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 1016-1020. https://doi.org/10.1109/ICICI.2017.8365293

[11] Jadhav, A. S. (2020). A novel weighted TPR-TNR measure to assess performance of the classifiers. *Expert Syst. Appl., 152*, 113391. https://doi.org/10.1016/j.eswa.2020.113391

[12] Hebbar, A., Sanath, P., Rajeshwari, S., & Saqquaf, S. (2018). Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees, 934-938.

[13] Bhartiya, N., Jannu, S., Shukla, P., & Chapaneri, R. (2019). Employee Attrition Prediction Using Classification Models. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 1-6. https://doi.org/10.1109/I2CT45611.2019.9033784

[14] Jain, R. & Nayyar, A. (2018). Predicting Employee Attrition using XGBoost Machine Learning Approach. *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, 113-120. https://doi.org/10.1109/SYSMART.2018.8746940

[15] Reddy, U. S., Thota, A., & Dharun, A. (2018). Machine Learning Techniques for Stress Prediction in Working Employees, 1-4. https://doi.org/10.1109/ICCIC.2018.8782395

[16] Self, J. & Dewald, B. (2011). Why Do Employees Stay? A Qualitative Exploration of Employee Tenure. *International Journal of Hospitality & Tourism Administration, 12*, 60-72. https://doi.org/10.1080/15256480.2011.540982

[17] Lee, C. H. & Bruvold, N. T. (2003). Creating value for employees: investment in employee development. *The International Journal of Human Resource Management, 14*(6), 981-1000. https://doi.org/10.1080/0958519032000106173

[18] Burnett, J. & Lisk, T. C. (2019). The Future of Employee Engagement: Real-Time Monitoring and Digital Tools for Engaging a Workforce. *International Studies of Management & Organization, 49*, 108-119. https://doi.org/10.1080/00208825.2019.1565097

[19] Ramamurthy, K., Singh, M., Davis, M., Kevern, J.A., Klein, U., & Peran, M. (2015). Identifying Employees for Re-skilling using an Analytics-Based Approach. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 345-354. https://doi.org/10.1109/ICDMW.2015.206

[20] Gelbard, R., Ramon-Gonen, R., Carmeli, A., Bittmann, R., & Talyansky, R. (2018). Sentiment analysis in organizational work: Towards an ontology of people analytics. *Expert Syst. J. Knowl. Eng., 35*. https://doi.org/10.1111/exsy.12289

[21] Liu, J., Li, J., Wang, T., & He, R. (2019). Will Your Classmates and Colleagues Affect Your Development in the Workplace: Predicting Employees' Growth Based on Interpersonal Environment, 71-78. https://doi.org/10.1109/BigDataService.2019.00016

[22] Levenson, A. (2018). Using workforce analytics to improve strategy execution. *Human Resource Management, 57*, 685-700. https://doi.org/10.1002/hrm.21850

[23] Thite, M. (2010). All that Glitters is not Gold: Employee Retention in Offshored Indian Information Technology Enabled Services. *Journal of Organizational Computing and Electronic Commerce, 20*, 7-22. https://doi.org/10.1080/10919390903482390

[24] Srivastava, D. & Tiwari, P. (2020). An analysis report to reduce the employee attrition within organizations. *Journal of Discrete Mathematical Sciences and Cryptography, 23*, 337-348. https://doi.org/10.1080/09720529.2020.1721874

[25] Aliyu, O. & Nyadzayo, M., (2016). Reducing employee turnover intention: a customer relationship management perspective. *Journal of Strategic Marketing*, 1-17. https://doi.org/10.1080/0965254X.2016.1195864

[26] Schiemann, W. A., Seibert, J. H., & Blankenship, M. H. (2018). Putting human capital analytics to work: Predicting and driving business success. *Human Resource Management, 57*, 795-807. https://doi.org/10.1002/hrm.21843

[27] Robinson, M. (2018). Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management, 57*, 739-750. https://doi.org/10.1002/hrm.21852

[28] Hendrick, R. Z. & Raspiller, E. E. (2011). Predicting Employee Retention through Preemployment Assessment. *Community College Journal of Research and Practice, 35*, 895-908. https://doi.org/10.1080/10668920802421561

[29] Book, L., Gatling, A., & Kim, J. (2019). The effects of leadership satisfaction on employee engagement, loyalty, and retention in the hospitality industry. *Journal of Human Resources in Hospitality & Tourism*, 18, 1-26 https://doi.org/10.1080/15332845.2019.1599787

[30] Kryscynski, D., Reeves, C., Stice-Lusvardi, R., Ulrich, M., & Russell, G. (2018). Analytical abilities and the performance of HR professionals. *Human Resource Management, 57*, 715-738. https://doi.org/10.1002/hrm.21854

[31] Wei, D., Kush, R., & Wagman, M. (2015). Optigrow: People Analytics for Job Transfers, 535-542. https://doi.org/10.1109/BigDataCongress.2015.84

Authors' contacts:

Nilasha Bandyopadhyay, Student
(Corresponding author)
Symbiosis Centre for Information Technology, Pune
Plot No: 15, Rajiv Gandhi Infotech Park, MIDC, Hinjewadi, Phase 1, Pune,
Maharashtra 411057, India
8458075463, nilasha.bandyopadhyay@associates.scit.edu

Dr. **Anil Jadhav,** Professor
Symbiosis Centre for Information Technology, Pune
Plot No: 15, Rajiv Gandhi Infotech Park, MIDC, Hinjewadi, Phase 1, Pune,
Maharashtra 411057, India
9764294698, anil@scit.edu