Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
DOI: https://doi.org/10.17762/msea.v70i2.2469

Prediction of Employee Turnover based on Machine Learning Models

Rahul Chauhan
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number:1767 - 1775
Publication Issue:
Vol 70 No. 2 (2021)

Abstract
As a result of the fact that the procedure of decision making constitutes a vital component in the management of a company, the personnel of that company are seen as a valuable kind of asset by the latter. Therefore, the procedure of employing them in the first place by making the appropriate choices is generally recognised as a well-known obstacle by administrative authorities. Employee turnover may be a time-consuming and difficult process because recruiting new workers requires not only additional time but also a significant amount of financial expenditure. In addition to this, there are a number of additional elements that play a role in the selection and hiring of a qualified applicant, who in turn would provide economic returns for an organisation. In this research, I propose building a model to predict employee turnover rate using data from three datasets acquired from the Kaggle repository and a subset of their features. To analyse staff traits and forecast turnover and churn rate, the work summarised here employs machine learning approaches and pre-processing techniques. Logistic regression, AdaBoost, XGBoost, KNN, decision tress, and Naive Bayes are only some of the machine learning algorithms tried out on extracted datasets in the report's implementation experiments. Evaluating qualities against evaluation parameters like accuracy and precision factors follows thorough research and training of selected attributes.

Article History
Article Received*: 05 September 2021*
Revised*: 09 October 2021*
Accepted*: 22 November 2021*
Publication*: 26 December 2021*

Keywords: —Churn rate, Employee attrition, Employee retention strategy, Employment Features, Machine learning algorithms

Introduction

The term "employee attrition" sometimes goes by the name "churn rate," and it generally refers to the process by which an organization experiences a loss of manpower and human employment as a result of employees leaving the organization without giving prior notice or retiring from their positions. This constant churning through staff and workers causes attrition over a predetermined length of time, which can lead to a drain on human resources [1]. This expense was ultimately incurred as a result of a newly implemented recruitment process and the necessary monetary investment in the training and development of newly employed personnel. In spite of this, the attrition rate, also known as the churn rate, is strongly dependent on the termination criteria of the organization as well as the reason that employee gives for leaving their position. Because of this, concepts such as abdication, occupational relinquishment, and termination are utilized in order to examine and quantify the amount of manpower that was involved.

A company's ability to invest in and choose employees and resources that are capable of competing in the market is normally the responsibility of the Human Resources (HR) department, which forms the backbone of any organization. This procedure, which is more commonly referred to as the "hiring process," requires the investment of both time and money from the company. Additionally, the corporation suffers a loss if a valuable resource departs the organization. This incidence of

unexpected loss is typically referred to as "employee attrition" and has a tendency to have a direct impact towards the organization's ability to carry out its daily operations [2].
In addition to an investment of both time and money, the employee also has to be managed by some higher authority who can direct them throughout their service period until they get familiar with their work. This can help them learn their job more quickly. As a result, one could argue that it is impossible to consciously prevent the process of attrition. Therefore, this problem of employee morale needs to be solved in order to achieve a desirable working environment, address the issue of employee turnover and churn rate, and significantly reduce the number of employees who leave their jobs [3].
Various HR departments have been able to judge the historical patterns of an employee and predict the attrition rate in a company by using computer-aided technologies such as data mining and machine learning. This has been made possible thanks to the widespread adoption of these technologies. On the other hand, up until now, this process of attrition has been performed manually by subject matter specialists according to gender, cause of termination, etc. Therefore, in light of the aforementioned, the purpose of the proposed research is to develop a model based on machine learning that is capable of predicting the probability of such attrition occurring in an organization on the basis of certain factors.
In order for the model to be able to optimally forecast and support the domain experts with dependable outcomes, it is vital to assess the aspects that have been described above. In order to accomplish this goal, I proposed developing a model that would be able to incorporate machine learning-based methods like logistic regression, KNN, and decision trees in order to achieve real-time analysis of employee turnover.
The following is a condensed version of the primary contributions that this work has made:

- The application of machine learning algorithms, along with several pre-processing stages, in order to generate an accurate prediction of the customer turnover rate
- Utilization of three datasets in order to conduct attribute feature analyses and counteract the effects of the noisy dataset
- The production of outcomes that are optimized by utilizing the appropriate algorithms
- Assessment of the job by the application of hyper parameters including the confusion matrix

Related Works

The loss of brilliant personnel is one of the most significant challenges that businesses confront today. Therefore, in order to get around this problem, Sarah and her colleagues developed a model that was published in [4]. In this model, Sarah used machine learning algorithms to figure out which factors were responsible for the attrition rate. These factors were identified, and then models based on machine learning, specifically KNN and SVM, were trained. In a different strategy that was carried out by D. Vanden [5], the authors provided a voice-based attrition detection method that was utilized in contact centres to support the decision that was made about attrition detection. The authors of this study built a model that was based on textual formats, and the attrition variables were entered manually by department personnel.
In a different piece of work that Chiu and colleagues [6] produced, he offered a method for attrition detection that involved the use of a data mining methodology. In this study, the model was developed based on telecommunication elements including subscriber data and the length of the call's time period. Singh et al. [7] wanted to focus on how employee turnover affected the normal operation of a company and identify the major component that could predict how often this problem occurred in the future. Their research was published in the journal Management Science. He

provided specific numbers that were derived from a variety of reports and discussed how the productivity of an organization was reliant on its resources as well as its workforce.

Another study that highlighted the elements of employee retention was mentioned in [8] by Zhou.et.al. The author of this study stressed how the overall productivity of an employee could be increased, and this study was cited as demonstrating the features of employee retention. In addition to this, the author provided a comprehensive analysis of the myriad of unfavourable circumstances that led to the dismissal of an employee.

The authors of [9] set out to construct a model with the intention of predicting the key reasons that employees leave their jobs. In order to accomplish this goal, they utilized six different machine learning algorithms and assessed each approach against fundamental assessment parameters in an effort to reach the highest possible level of accuracy. In a study that was quite similar to this one [10], the authors used the IBM-HR dataset to predict employee turnover. The methods of feature selection and feature extraction were used because many of the features contained in the dataset were not organized. PCA, which came later, was the method that ultimately succeeded this one in reducing the dimensionality of the feature.

Logistic regression was the ML technique that the authors of [11] successfully adopted in order to successfully implement prediction of employee attrition. In order to accomplish this, they obtained the dataset from the Kaggle repository and set up a relationship between the input characteristics and the output vectors. The accuracy of the model was improved to its maximum potential, and it is now being used in a variety of different businesses. However, a random forest was used as the basis for the process of selecting features, and logistic regression was applied in the later stages of the modelling process in order to predict the final output of the model. Within the field of machine learning, this study received a lot of praise and admiration.

In a separate piece of research referred to as [12], the authors executed their idea on a dataset with the help of several machine learning algorithms, such as decision trees and KNN, in order to make predictions about the turnover rate of employees. They validated their work by employing a technique known as 10-fold cross validation, and after that, they divided the dataset 70:30 between being used for training and testing purposes correspondingly. However, the accuracy of the results they obtained was lower when compared to the accuracy of the results obtained by other research works; as a consequence, their contribution was only applicable to the pre-processing stage and could not be used in any further stages.

In a distinct piece of work that the authors contributed to [13], the researchers assessed the methods and constraints of article [12] and, as a solution to the problem of employee attrition, they later employed new sets of classifiers. However, the authors of this study also used a 70:30 split of the dataset for training and testing purposes, which was very similar to the approach taken in the work proposed in [12].

The authors of [14] made a similar suggestion, proposing that the churn rate of employees may be predicted by showing the different processes involved in the framework. They described the various stages of the model by using a feature selection method, which was then integrated with strategies for data reduction. Following the first stage of the implementation, which involved training the model through the use of machine learning methods such as logistic regression, the third stage of the implementation proposed to predict the model through the use of confidence analysis. When compared to other methods, it was seen that this system provided the highest level of precision possible. Classification trees and random forests were used in an experiment that was part of a research paper that was published in [15] and intended to estimate the employee churn rate. The authors of this study applied the principles of Pearson correlation to the pre-processing of the data,

which involved removing undesirable characteristics. Comparing the various algorithms that were employed to achieve optimum accuracy was another contribution of this work.

System Design

The specific characteristics chosen from the dataset will have a significant impact on how the model is implemented. The performance and accuracy of prediction are both impacted and influenced by this particular feature that was chosen. The following is a condensed summary of the model's implementation in its entirety:

A Dataset

The dataset that was utilized to put into action the work that was presented was taken from the Kaggle repository. In order to attain an enhanced level of efficiency in the system model and to construct an experimental verification, a total of three datasets are utilized. However, in order to maximize the effectiveness of the system, the characters from the dataset are initially transformed into their corresponding numerical values. The following is a rundown of the datasets that were utilized:

- This dataset provides information on employee terminations that have taken place over the past decade, and as a result, it provides information about employees who are now working for the firm. This dataset was created with the primary purpose of analysing employee terminations and determining the likelihood of such events occurring in the future based on the data that was provided. The dataset has a total of 18 columns and 49653 rows, each of which contains an employee's identifying information, such as their city name, job title, employee ID, and gender.
- HR Dataset: This dataset is used to obtain insights into an employee working in a company in order to reduce the likelihood of employee turnover in the foreseeable future. The dataset has a total of ten columns, each of which contains selected attribute information such as the most recent evaluation, number of promotions received in the past five years, number of projects completed, sales, and compensation.

B Pre-Processing

After completing this process, the next one is to clean the raw data that was retrieved from the repository. This involves removing any missing numbers and achieving successful results as a result. In the first stage of processing, known as pre-processing, one form of data is converted to another form, and the integration process is used to smooth out the numerical values. As a consequence of this, the received data is now condensed to a greater extent and, if necessary, formatted in order to generate useful results.

C Analysis of Dataset

During the phase of data analysis, the categorical values received from the dataset are transformed to their appropriate numerical values. This is done to boost the overall efficiency of the classification algorithms. The graphic below illustrates a correlation matrix that is constructed in order to determine the link between the features of a dataset. In the beginning, all of the categorical characteristics such as "salary" that contain values such as low, medium, and high are converted to numerical values such as 0, 1, and 2 respectively. Consequently, in this kind of circumstance, a correlation matrix is helpful in identifying the attributes that depict strong and weak correlation.

D Workflow

The first step in putting the suggested model into action is to collect raw data from three different repositories. Each of these repositories contains a variety of information on employees that might be used to determine the turnover rate. The following stage of the process involves the data going through a phase known as pre-processing, during which time any unnecessary characteristics are removed and the pertinent features are worked on. Following this, the EDA process is carried out, in which the data is visually analysed through the use of a correlation matrix in order to gain insights on the numerical values contained within the dataset. After the EDA process has been completed, the dataset is partitioned into a train-test phase, and the appropriate machine learning algorithms are applied to the data in order to learn from it. After each algorithm has been implemented, its accuracy score will be measured in order to determine which algorithm is the best at predicting the future.

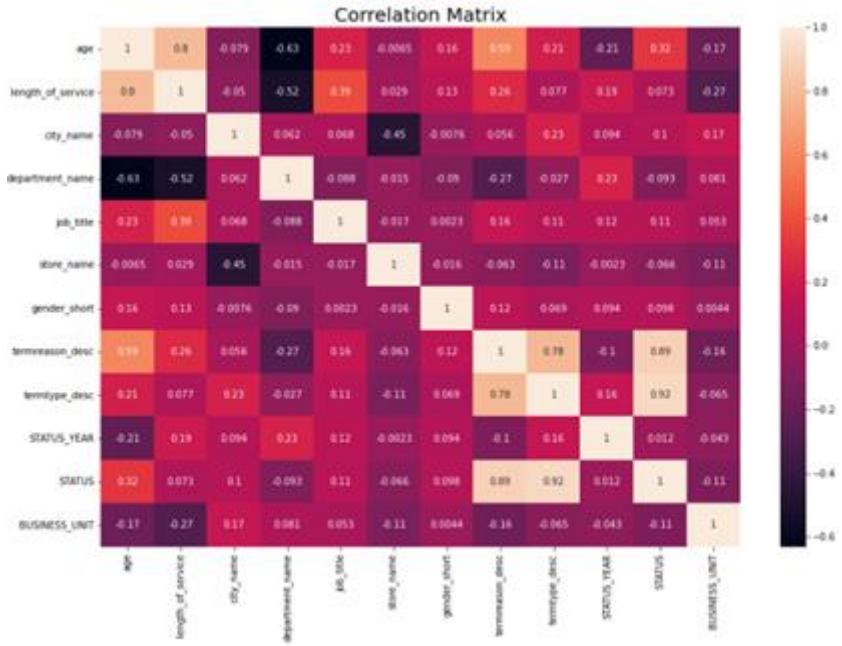

Figure1: Correlation Matrix of Employee Attrition

Experimental Analysis

Employee attributes such as pay, city name, gender, and so on are provided, as was described in the datasets that came before this one. The numbers that are derived from this data are then fed into machine learning algorithms, and churn rate is projected to get insights into whether or not an employee would leave the firm. This section makes use of a total of three datasets; consequently, it is broken up into three distinct chunks of experimental analysis, each of which is accompanied by a summary of the performance analysis.

A. Experimental Analysis on Employee Dataset

This dataset, which has a total of 18 columns and 49653 rows, is used to acquire insights on employee terminations that have taken place over the course of the past 10 years. The data can be accessed to do so. This dataset is used specifically for the purpose of improving one's ability to anticipate the rate of employee turnover that may occur in the future. This dataset is used to train the model, and the machine learning classifiers decision trees and Naive Bayes are used in the

training process. The confusion matrix was used to gather the information and numerical values that are presented in the table that follows.

Table1: Obtained values from Confusion Matrix

Name of Algorithms	Train/Test	Values	Accuracy
Decision Trees	Train Set	[37880,625] [000,38563]	99.18
	Test Set	[9477,186] [000,9605]	**99.03**
Gaussian Naïve Bayes	Train Set	[34941,3564] [22351,16212]	65.84
	Test Set	[8751,912] [5669,3936]	65.84

The confusion matrix that was developed is depicted in the following graph, which also includes numerical values relating to false positive and true negative cases. The deployment of decision trees, on the other hand, resulted in 9477 true positive cases of employee attrition prediction and 186 false positive cases. This is something that can be noticed. In a manner analogous, GNB came up with 8751 cases of actual positivity and 912 cases of fake positivity for the same. As a result of this, it is clear to see that the testing precision of decision trees was demonstrated to be higher than that of GNB.

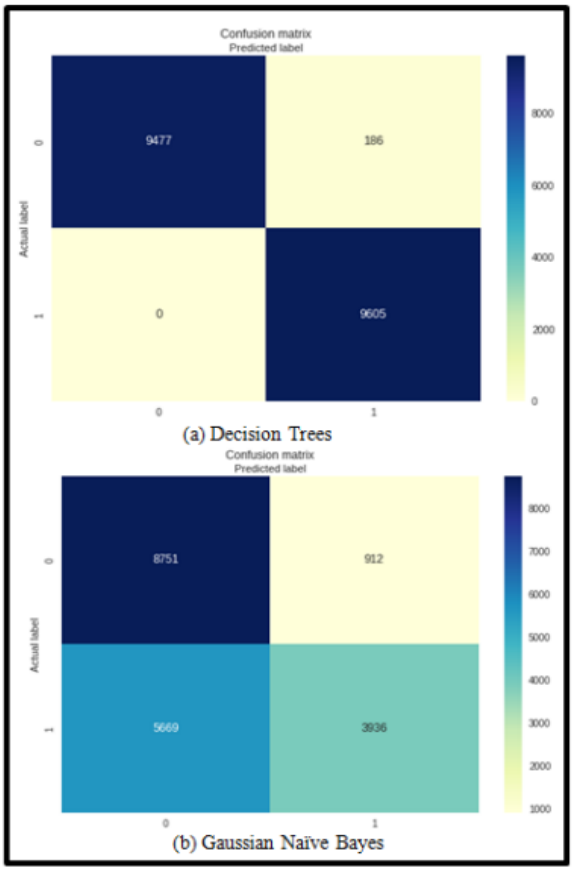

Figure2: Output of confusion matrix

Table2: Obtained values from classification report

Name of Algorithm	Precision		Recall		FI-Score	
Decision Trees	Positive Cases	Negative Cases	Positive Cases	Negative Cases	Positive Cases	Negative Cases
	1.00	0.98	0.98	1.00	0.99	0.99
Accuracy	**0.99**					
Gaussian Naïve Bayes	Positive Cases	Negative Cases	Positive Cases	Negative Cases	Positive Cases	Negative Cases
	0.61	0.81	0.91	0.41	0.73	0.54
Accuracy	**0.66**					

The numerical data that was collected from table 2 provides an accurate and precise overview of the values that were obtained through the categorization report. As a result, the total results are given together with their various degrees of accuracy. On the other hand, it has been found that the model produces the most accurate results when it is tested using decision trees, which results in an accuracy of 99%. This was discovered when it was found that this method created the highest level of accuracy.

B. Experimental Analysis on Telco Churn Dataset

This customer retention data collection, which has a total of 7043 rows and 21 columns, was designed with the goal of predicting the behavior of customers regarding their individual retention rates. Machine learning classifiers AdaBoost and XGBoost are used during the training process of the model, which is carried out on the dataset. The confusion matrix was used to gather the information and numerical values that are presented in the table that follows.

Table3: Obtained values from Confusion Matrix

Name of Algorithms	Train/Test	Values	Accuracy
AdaBoost	Train Set	[3775,377] [689,784]	81.04
	Test Set	[908,103] [191,205]	79.10
XGBoost	Train Set	[3836,316] [674,799]	82.39
	Test Set	[924,87] [187,209]	**80.52**

The confusion matrix that was developed is depicted in the following graph, which also includes numerical values relating to false positive and true negative cases. The deployment of ADaBoost, on the other hand, resulted in 908 incidents of genuine positive employee attrition prediction and 103 cases of false positive employee attrition prediction. In a manner analogous, XGBoost generated 924 genuine positive results and 87 erroneous positive results for the same. The conclusion that can be drawn from this is that the testing precision of XGBoost was demonstrated to be superior than that of ADaBoost.

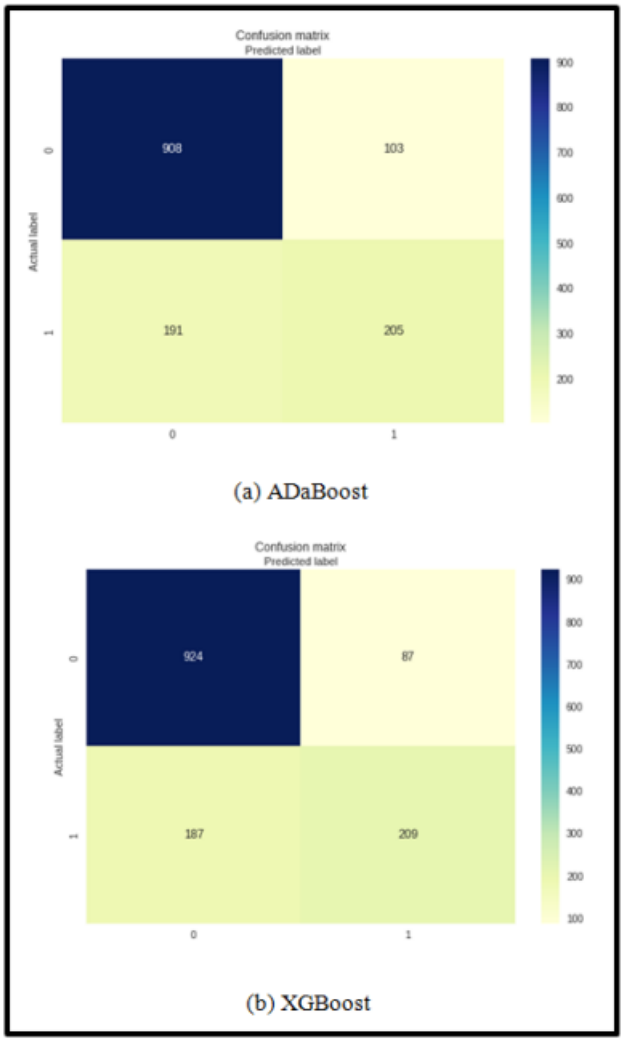

Figure3: Output of confusion matrix

Conclusions

The implementation of machine learning techniques and the prediction of the corresponding staff turnover rate are the key goals of the report that has been provided. Because the churn rate and turnover of an organization have a direct impact on the profits of that organization, the decision to hire and fire employees and individuals becomes a very critical one to make in order to maintain the proper working balance in the environment. As a result, I have proposed that we forecast the same situation by utilizing three distinct datasets in conjunction with a variety of ML-based approaches. It has been determined through empirical research that KNN, XGBoost, and decision trees, in that order, provided optimum degrees of accuracy that were 94%, 81%, and 99%, respectively.

References

[1] Jarrahi, M.H. Artificial intelligence and the future of work: HumanAI symbiosis in organizational decision making. Bus. Horiz. 2018, 61, 577–586
[2] Duan, Y.; Edwards, J.S.; Dwivedi, Y.K. Artificial intelligence for decision making in the era of Big Data—Evolution, challenges and research agenda. Int. J. Inf. Manag. 2019, 48, 63–71
[3] Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. Computers 2020, 9, 86
[4] Sarah. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Vaesens,"New insights into a churn prediction in the telecommunication sector. An profit driven datamining approach," European journal of operational research, vol. 218, no. 1, pp. 211- 229, 2012
[5] K.Coussement and D. VandenPoel, "Integrating the voice of customers through call centre emails into a decision support system for attrition prediction," Information & Management, vol. 45, no. 3, pp. 164–174, 2008
[6] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to attrition prediction: a data mining approach," Expert systems with applications, vol. 23, no. 2, pp. 103–112, 2002
[7] Singh, M., Varshney, K. R., Wang, J., Mojsilovic, A., Gill, A. R., Faur, P. I. and Ezry, R. (2012). An analytics approach for proactively combating voluntary attrition of employees, Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, IEEE, pp. 317–323
[8] Zhou, N., Gifford, W. M., Yan, J. and Li, H. (2016). End-to-end solution with clustering method for attrition analysis, Services Computing (SCC), 2016 IEEE International Conference on, IEEE, pp. 363–370
[9] Mohbey, K.K. Employee's Attrition Prediction Using Machine Learning Approaches. In Machine Learning and Deep Learning in Real-Time Applications; IGI Global: Hershey, PA, USA, 2020; pp. 121–128
[10] Ponnuru, S.; Merugumala, G.; Padigala, S.; Vanga, R.; Kantapalli, B. Employee Attrition Prediction using Logistic Regression. Int. J. Res. Appl. Sci. Eng. Technol. 2020, 8, 2871–2875
[11] Yang, S.; Ravikumar, P.; Shi, T. IBM Employee Attrition Analysis. arXiv 2020, arXiv:2012.01286
[12] Usha, P.M.; Balaji, N. Analysing employee attrition using machine learning. Karpagm J. Comput. Sci. 2019, 13, 277–282
[13] Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. Computers 2020, 9, 86
[14] Document validation and verification system S Shivadekar, SR Abraham, S Khalid Int. J. Adv. Res. Comput. Eng. Technol.(IJARCET) 5 (3)
[15] Design of a secure online academic document verification system: A case study of Nkumba University M Sarah Nkumba University