EPiC Series in Education Science

Volume 5, 2023, Pages 17–30

Proceedings of NEMISA Digital Skills Conference
2023: Scaling Data Skills For Multidisciplinary Impact

EPiC
Education
Science

A Predictive Model to Forecast Employee Churn for HR Analytics

Vengai Musanga[1*] and Colin Chibaya[2*]
[1] Zimbabwe National Defense University, Harare, Zimbabwe
[2] Sol Plaatje University, Kimberley, South Africa
vengaimusanga@gmail.com, colin.chibaya@spu.ac.za

Abstract

The challenge of employee churn is a major issue for most businesses and organizations. Unexpected employee departures can tarnish service delivery, harm customer loyalty, degrade quality of services, drop productivity, and hurt goodwill. The ability to predict employee churn is crucial for retaining valuable employees. This study proposes a predictive model that uses machine learning to forecast employee churn. The predictive model uses feature selection through Pearson correlation methods, information gain, and recursive feature elimination, combined with strong classification methods such as random forest, logistic regression, decision trees, gradient boosting machines, and K-nearest neighbours. The IBM dataset was used for training and testing the proposed predictive model. The accuracy of the different algorithms improved after applying particular feature selection methods. The results yielded showed that the random forest technique outperformed other models in terms of accuracy in the prediction of employee churn.

1 Introduction

The prediction of employee churn remains a major issue for most businesses and organizations globally. In this study, employee churn is defined as the loss of intellectual assets from a company (Pradhan et al.,2017). Employees may decide to leave an organization for various reasons such as dissatisfaction with salary, bureaucracy in the organization, or limited career growth (Khaled et al., 2021). However, the departure of skilled employees is detrimental and can lead to decreased productivity. The need for intelligent systems that can predict employee churn is apparent (Xiaojuan, 2016).

Human resources are considered the most valuable assets within an organization (Fulmer et al., 2013) and employee motivation is crucial in determining their continued stay. Organizations should go an extra mile to retain their manpower. However, it is hard to predict an employee's plans to leave. To address the issue of employee retention, machine learning approaches can be used to analyze past employee data and identify patterns that predict the likelihood of employees leaving the organization (Reyes et al., 2019). In this study, we make a comparison to see which machine learning technique best predicts employee churn.

Past research has explored the use of machine learning techniques for employee churn prediction, but the full potential of feature engineering and selection has not been fully exploited. This study aims to fill that gap by developing an integrated supervised machine learning model (Raza, 2022) that compares multiple feature selection and data balancing methods to improve employee churn predictive accuracy (Omar et al., 2014). The IBM dataset was chosen for this study as it encompasses common factors found in most sectors.

The proposed machine learning model is envisaged to be able to identify the critical factors in retaining valuable employees and assist human resources practitioners in their staff retention efforts. Subsequent sections of the rest of this study cover related literature, a description of the data used, the methods followed in completing this study, the results, and the discussions thereto. We close the work with a conclusion which mainly presents our recommendations, contributions of the work, and the likely direction for future research.

2 Related Work

Organizations predict employee churn to anticipate and understand the reasons for potential loss of employees, and to take proactive measures to retain valuable staff and minimize the impact of turnover on the business. By forecasting employee churn, organizations can improve their retention strategies, reduce hiring and training costs, and maintain a stable and productive workforce. Predicting employee churn offers several advantages for organizations. Firstly, improved retention strategies can be implemented by understanding the reasons for employee churn. This information allows organizations to develop targeted retention programmes that aim to reduce turnover and improve employee satisfaction. Another advantage of predicting employee churn is cost savings. Minimizing employee turnover can help to reduce the costs associated with hiring and training new staff. This is important as these costs can add up quickly, especially for organizations with high levels of turnover. A stable workforce also leads to increased productivity. When employees focus on their work without the disruptions caused by high levels of turnover, they are better able to perform at their best. This, in turn, can lead to improved performance and increased productivity. Effective management of employee churn can also give companies a competitive advantage. Companies that attract and retain top talent are better positioned to succeed in the marketplace, as they have a stable and productive workforce. This can give them a significant advantage over competitors that struggle to manage employee churn. Also, high levels of employee churn can negatively impact customer satisfaction. Customers may need to constantly adapt to new staff, leading to frustration and a decrease in satisfaction levels. By reducing employee churn, organizations can improve customer satisfaction and strengthen customer relationships. This can lead to increased loyalty and repeat business, which is essential for long-term success.

Several studies have investigated the ability of machine learning algorithms to forecast employee turnover (Omar et al., 2014; Umayaparvathi & Iyakutti, 2016; Amin et al., 2016). As a starting point in predicting employee churn, Hebbar et al. (2018) utilized logistic regression on the IBM's employee attrition dataset to determine the likelihood of an individual being part of the churn group. This allowed the researchers to get an initial understanding of the relationships between various employee characteristics and the likelihood of them leaving the company. By applying logistic regression to this dataset, Hebbar et al. (2018) were able to identify the key factors that contribute to employee churn, such as job satisfaction, job involvement, and work-life balance. This information can then be used to develop targeted retention programmes and reduce employee churn, leading to cost savings and increased productivity for the organization. The results of this study demonstrated the potential of logistic regression as a powerful tool in predicting employee churn and provided insights into the factors that organizations should focus on to reduce turnover.

Subsequently, a comparative study was conducted using random forest and support vector machine models to determine the key characteristics of the IBM employee attrition dataset. The study performed exploratory data analysis to understand the relationships between various employee attributes and the likelihood of churn. During this process, different data visualization techniques were used to represent the findings, such as bar charts, histograms, and scatter plots. The aim of this comparative study was to evaluate the performance of random forest and support vector machine models in predicting employee churn and compare their results with the findings of the previous logistic regression study. The results of this study helped to determine the strengths and weaknesses of each model, and to identify the most important factors that contribute to employee churn. This information can be used by organizations to develop more effective retention strategies, reduce employee turnover, and increase productivity. This comparative study demonstrated the potential of machine learning algorithms in predicting employee churn. The study provided a comprehensive evaluation of these algorithms and yielded results that showed that both random forest and support vector machine models had strong performance in predicting employee churn and were able to accurately identify the key factors that contribute to turnover. The study also highlighted the importance of performing exploratory data analysis and visualizing the findings, as this helps to understand the relationships between employee characteristics and the likelihood of churn. However, feature selection methods were not explored, a gap we explore further in this study. Hopefully, accuracy may improve with the adoption of feature selection methods.

In a study conducted by Dam (2021), the author investigated the use of various feature selection techniques for determining the most significant features in predicting employee churn. The author believed that by identifying the most informative features, a more accurate and efficient model for predicting employee churn could be developed. The author's findings provided insights into the importance of feature selection in the prediction of employee churn. In his study, Dam (2021) compared the benefits of three feature selection methods: wrapper, filter, and embedded. He ultimately chose Recursive Feature Elimination, a wrapper method, as his method of choice due to previous research indicating that wrapper methods are effective in identifying the most important features in datasets of medium to large size. In a study conducted by Zhao and colleagues (Zhao et al., 2019), the authors demonstrated the concept of feature importance in an XGBoost model that was trained on a dataset consisting of 1,000 items. The study showed how XGBoost models can be used to determine the relative importance of different features in a dataset, which can provide valuable insights for feature selection and model building. The results of the study demonstrated the practical applications of XGBoost models in evaluating feature importance and the potential benefits of using this approach in real-world data analysis. Other research showed that Tenure was also a good predictor of employee churn (Punnoose & Ajit, 2016).

The effect of employee satisfaction on employee churn was once investigated using regression methods (El-Rayes et al., 2020). The study found that employee satisfaction is a crucial factor in employee turnover, but the lack of advanced machine learning techniques for predicting employee attrition resulted in limited accuracy in the predictions. Similarly, Yigit and Sourabizadeh (2017) investigated employee turnover utilizing multiple techniques with differing levels of complexity. They performed two experiments, one that included feature selection and one without. In both experiments, they found that Support Vector Machine (SVM) was the most effective method for forecasting employee churn. Falluchi et al. (2020) studied a broad spectrum of algorithms and found that Gaussian Naïve Bayes (NB) had the highest recall rate at 0.541. However, using Gaussian NB on employee attrition data poses a challenge as it assumes all predictors are independent, which may not always be the case. This, combined with the algorithm's lower performance, results in a less precise prediction of employee turnover. Among the algorithms studied, decision tree performed second best, followed by logistic regression. Despite their limitations, conventional algorithms hold explanatory power and therefore merit consideration. Punnoose and Ajit (2016), however, discovered that the XGBoost algorithm surpassed Random Forest (RF), Logistic Regression (LR), and Gaussian NB in terms of accuracy. Additionally, XGBoost's inherent regularization helps to avoid overfitting. The authors stated that the data used for predicting employee attrition contains noise, and XGBoost can effectively handle this by interpreting the noise as relevant information rather than ignoring it. This leads to the model learning the noise and resulting in a non-generalizable model (Ying, 2022). In their research, the authors analyzed data from a global retailer with 73115 rows and 33 columns, and compared several machine learning techniques (LR, NB, RF, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), SVM, XGB). The results showed that XGBoost achieved an AUC score of 0.86 on the test set, while SVM and RF scored 0.52 and 0.51, respectively. The authors concluded that XGBoost is more effective for forecasting employee churn (Punnoose & Ajit, 2016). Also, Gabrani and Kwatra (2018) determined that job satisfaction, length of employment, and evaluation are dependable indicators of employee turnover. Most of the research has shown that tree-based models, such as AdaBoost, Gradient Boosting Tree, Random Forest, and Extreme Gradient Boosting classifiers, consistently outperformed other models. The Multilayer Perceptron classification model was also used in some studies, with varying results.

Research on predicting employee churn was conducted using NB, LR, DT, and RF methods (Saradhi & Palshikar, 2011). Another study by Khare et al. (2013) developed an attrition risk equation using LR to forecast employee turnover. Nonetheless, both studies omitted feature engineering and feature selection techniques. Basha et al. (2018) employed a gradient boosting classifier to construct the model. After evaluating the model's recall, precision, and accuracy, the gradient boosting tree outperformed other algorithms with an accuracy of 96%. The results also indicated that employees who are dissatisfied are more likely to leave the organization, while those who have a longer tenure and are engaged in their work are less likely to depart. In their study, Sisodia et al. (2017) assessed the performance of machine learning models for predicting employee turnover. The researchers aimed to compare the predictions of employee churn using various machine learning models such as SVM, NB, and DT. The results revealed that the RF model had a relatively high prediction accuracy. It was noted that a higher level of accuracy could have been achieved if more advanced machine learning techniques were used and the significance of features were validated through multi-criteria models (Gabrani & Kwatra, 2018). A comparison of six machine learning models was performed in the IT sector(Tharani & Raj, 2020). The authors found that among the six evaluated models, Extreme Gradient Boosting was the top performer. This conclusion aligns with earlier research that reached similar results (Zhao et al., 2019; Punnoose & Ajit, 2016). Tharani and Raj (2020) also regarded RF, Multilayer Perceptron, SVM, and KNN as good-performing models. These findings are in partial agreement with prior literature that evaluated RF and Multilayer Perceptron as good-performing models (Punnoose & Ajit, 2016). On the other hand, SVM and KNN were poor performers (Zhao et al., 2019; Punnoose & Ajit, 2016).

3 Data

The data used in the study was sampled from the IBM dataset which contains information about the employees such as personal details (name, employee ID, address, date of birth, and gender); contact details such as email and phone number; employment information such as hire date, job title, and department; compensation and benefits such as salary, bonuses, and insurance; performance evaluations such as reviews, ratings, and feedback; attendance and time off information such as sick leave, and vacation days; as well as education and training information, including degrees, certifications, and the courses taken. Initially the dataset had 35 features. However, 5 were removed due to redundancy, remaining with 30 features. Table 1 shows the 30 key features that were considered, along with the related data types.

Number	Feature	Data type
1	Age	Numeric
2	Business Travel	Categorical
3	Daily Rate	Numeric
4	Department	Categorical
5	Distance from Home	Numeric
6	Education	Categorical
7	Education Field	Categorical
8	Gender	Categorical
9	Environment satisfaction	Categorical
10	Hourly Rate	Numeric
11	Job Involvement	Categorical
12	Job Level	Categorical
13	Job Role	Categorical
14	Job Satisfaction	Categorical
15	Marital Status	Categorical
16	Monthly Income	Numeric
17	Monthly Rate	Numeric
18	Number of Companies Worked	Numeric
19	Overtime	Categorical
20	Percent Salary Hike	Numeric
21	Performance Rating	Categorical
22	Relationship Satisfaction	Categorical
23	Stock Option Level	Categorical
24	Total Working Years	Numeric
25	Training Times last Year	Numeric
26	Work Life Balance	Categorical
27	Years at Company	Numeric
28	Years in Current Role	Numeric
29	Years since Last Promotion	Numeric
30	Years with Current Manager	Numeric

Table 1: HR dataset features

4 Methods

The steps involved in predicting employee churn included importing the IBM dataset into a Jupyter Notebook as the first step. Subsequently, exploratory data analysis was conducted to gain a deeper insight into the data. Oversampling was done to address the problem of imbalanced data. To proceed, 70% of the data was sampled into training data, while the remaining 30% was designated as testing data. This ratio of 70/30 split was adopted based on research findings indicating optimal results (Gholamy et al., 2018).

Methods of feature selection, including Pearson correlation, Information gain, and Recursive feature elimination, were applied to identify the most crucial factors for prediction. Machine learning algorithms, including logistic regression, random forest, gradient boosting. decision trees, and K-nearest neighbour, were applied to each result of the feature selection methods. The accuracy, precision, recall, and F-Score of the algorithms were evaluated using test data for comparison.

The machine learning algorithms were rerun on the data without the use of feature selection methods. The classification results before and after feature selection were compared and the feature selection and classification approach with the highest accuracy and precision was selected. Finally, the key factors affecting employee turnover were analyzed and strategies for retaining employees were evaluated.

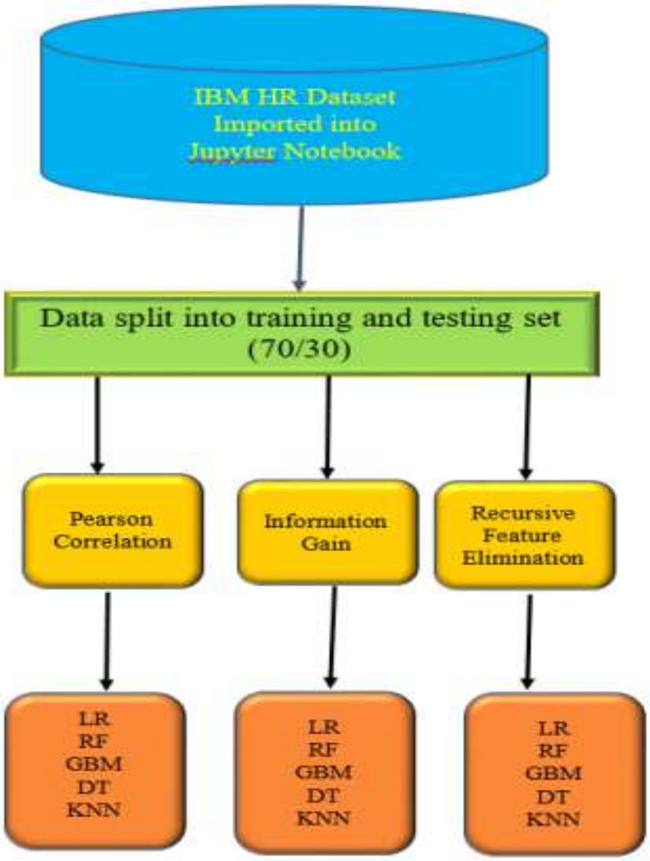

Fig 1.

4.1 Data imbalances

The target variable data refers to the variable in a dataset that is being predicted or modeled. In this case, the target variable data is imbalanced, meaning that the distribution of the data between the different classes is not equal (Dam, 2021). This can have negative effects on the performance of a machine learning model, as it can lead to decreased accuracy and bias towards the majority class. To mitigate these effects, two techniques were used to balance the data: Over-sampling and SMOTE. Over-sampling involves duplicating samples from the minority class to increase its representation in the dataset, while SMOTE (Synthetic Minority Over-sampling Technique) creates new synthetic samples for the minority class based on the existing samples that are closest to it in the feature domain. These techniques aim to balance the data, so that the machine learning model can be trained on more representative data and produce better results (Dam, 2021).

4.2 Feature Selection

To improve the prediction of employee turnover, the study used various feature selection methods to identify the most important features for the task (Pranto, 2022). This preprocessing step is beneficial for several reasons, including improvement in performance, reduction of overfitting, and reduced computational costs (Dam, 2021). The study evaluated three different feature selection methods: Pearson Correlation, Information Gain, and Recursive Feature Elimination. Pearson Correlation (PC) selects features based on statistical measures and is known for its velocity and efficiency (Dam, 2021). Features with a correlation greater than 0.8 were removed, leaving only columns with a correlation below 0.8. After this, classification algorithms were applied to the feature subset.

Information Gain (IG) evaluates the reduction of randomness in the data after transforming the dataset (Pranto, 2022). It analyzes each feature's contribution to the target variable by calculating its information gain. The features were ranked by descending information gain and those with a threshold of 0.005 or higher were included in the feature subset (Gupta, 2022). Again, classification algorithms were applied to this subset. Recursive Feature Elimination (RFE) is a feature selection method that iteratively measures feature importance and eliminates the least significant features (Sharma & Yadav, 2021). The features were ranked by importance using the RF method and the least important ones were removed until the desired number of features was reached. The RF-RFE algorithm was used to create the feature subset, followed by the implementation of classification algorithms.

4.3 Classification Algorithms

The study utilized a diverse range of five classification algorithms to analyze the data. The first algorithm is LR, which is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The second algorithm used is RF, which is an ensemble learning method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The third algorithm is Gradient Boosting Machine (GBM), which is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The fourth algorithm used is DT, which is a simple representation for classifying examples by recursively partitioning the feature space. The final algorithm applied is KNN, which is a non-parametric method used for classification and regression. The algorithm assumes that similar things exist in close proximity.

The LR algorithm is a statistical technique used for binary classification, where the goal is to predict the probability of a categorical event happening (e.g., yes or no, true or false). The LR algorithm works by fitting a logistic function, also known as a sigmoid curve, to the data. The logistic function maps any real-valued number to a value between 0 and 1, which can then be interpreted as a probability value. In the context of LR, the logistic function represents the relationship between the input features and the binary outcome. By estimating the parameters of the logistic function using maximum likelihood estimation, the LR algorithm can produce a model that predicts the probability of the event of interest for any new observation. Here, *x* represents the value of the output.

$$S(\mathrm{x})= \frac{1}{1+e^{-x}}$$

An analysis of the relationship between the dependent variable and independent variables is carried out using this algorithm. To accomplish this, a line must be fitted and the error between the line and the data points must be minimized. Weighting factors define the importance of predictors (Dam, 2021). The ease with which Logistic Regression can be implemented and trained made it a suitable baseline model for the study.

A RF is a predictive modeling technique that makes use of DTs as its foundation. The advantage of this approach lies in its ability to mitigate overfitting, a common problem in machine learning, by aggregating the predictions from multiple trees rather than relying on the result from a single DT. This aggregation helps to reduce the variance in the predictions, leading to improved performance compared to a single DT, while maintaining the same level of bias. In the model being described, all features were transformed into numerical values, after which the output variable was predicted using the RF technique. The use of DTs as the basis for prediction in RF offers several benefits, including its ability to handle complex non-linear relationships between the input variables and the output variable, as well as its ability to handle large amounts of data and to handle missing data. In summary, RF is a powerful technique for predictive modeling that can significantly improve the performance of a single DT by combining the results from multiple trees and reducing variance while maintaining bias. The conversion of all features to numeric values is an important step in this process, as it enables the model to handle a wide range of data and make accurate predictions.

GBM is a powerful machine learning technique that combines multiple weak learners to form a strong ensemble model. In GBM, the weak learners are combined iteratively in a way that improves the overall accuracy of the model. The process of combining weak learners in GBM can be seen as an optimization problem, where the goal is to minimize the error rate of the ensemble model. This optimization is achieved through a gradient descent procedure, where the error rate is progressively and repeatedly reduced. In the study being described, all features were converted to numerical values, allowing the GBM algorithm to process the data and make predictions. The target variable in this study was "Attrition", which was predicted using the GBM technique. The ability of GBM to handle a wide range of data types and its ability to handle complex non-linear relationships between the input variables and the output variable make it a popular choice for predictive modeling. In summary, GBM is a powerful technique for predictive modeling that involves combining multiple weak learners to form a strong ensemble model. The optimization process used in GBM reduces the error rate of the model and improves its overall accuracy. The conversion of all features to numeric values is an important step in this process, as it enables the GBM algorithm to process the data and make accurate predictions.

DTs are a widely used machine learning technique for data prediction and classification. They have a tree-like structure that can be easily visualized and understood, making them a popular choice for both researchers and practitioners.. In a DT, every internal node represents a test on an attribute of the data, and every branch represents the outcome of that test. The leaf nodes of the tree represent the class labels of the data. This flowchart-like layout makes it easy to understand how the model arrived at its predictions and can help in interpreting the results. DTs are effective for both regression and classification problems, and they can handle both continuous and categorical data. They are particularly useful for dealing with complex non-linear relationships between the input variables and the output variable. DTs can also handle large amounts of data and can handle missing data effectively. In summary, DTs are a powerful technique for data prediction and classification that can handle complex relationships between the input variables and the output variable. The tree-like structure of DTs makes them easy to understand and interpret, and they can handle a wide range of data types and missing data.

KNN is a popular machine learning algorithm used for binary classification tasks. In KNN, data points are classified based on the class of their nearest neighbors. The number of neighbors considered is specified by the user and is represented by the parameter "k". To determine the nearest neighbors, a distance measure is used. The most common distance measure used in KNN is the Euclidean distance, which calculates the straight-line distance between two points in a multidimensional space. This distance is calculated for each instance in the dataset, and the k instances with the smallest distances are identified as the nearest neighbors. Once the nearest neighbors have been identified, the unlabeled instances are classified based on the majority class of their k-nearest neighbors. If k is set to 3 and 2 out of the 3 nearest neighbors belong to class A and 1 belongs to class B, the unlabeled instance will be classified as belonging to class A. This process is repeated for all instances in the dataset, and the resulting classifications form the output of the KNN algorithm. In conclusion, KNN is a simple and effective method for binary classification tasks, as it makes use of the class information of nearby instances to make predictions. By using the Euclidean distance and a specified value of k, KNN can effectively classify new, unlabeled instances.

5 Results

To evaluate the various strategies, the feature selection techniques and corresponding algorithms were executed, and the results of their accuracy were summarized in Table 2. The purpose of comparing the different approaches was to determine which method performed the best in terms of accuracy. The feature selection methods used were applied to identify the most important features in the data set, which then served as inputs for the relevant algorithms. The accuracy scores of the algorithms were recorded and presented in Table 2 for easy comparison and analysis. This helped to determine the most effective combination of feature selection and algorithm for the specific problem at hand.

Feature selection method	Accuracy %				
	LR	RF	GBM	DT	KNN
PC	91.62	91.76	91.49	82.16	88.51
IG	88.78	92.57	90.68	82.57	87.43
RFE	87.43	92.30	90.14	79.59	87.70

Table 2: Classification accuracy with feature selection

In addition to evaluating the performance of the algorithms with the use of feature selection methods, the algorithms were also tested without these methods to compare the results. The accuracy scores of the algorithms in this scenario were recorded and presented in Table 3. The purpose of this comparison was to determine if feature selection improved the accuracy of the algorithms, or if the algorithms could produce acceptable results without it. By presenting the accuracy scores without feature selection in Table 3, a clearer understanding of the impact of feature selection on the performance of the algorithms could be obtained. This information would then inform the decision of whether to use feature selection in a given scenario, or not.

	Accuracy %				
	LR	RF	GBM	DT	KNN
Original, untreated data	84.35	86.17	87.53	76.64	84.13

Table 3: Classification accuracy without feature selection

When analyzing Table 2 and Table 3, it becomes evident that the accuracy of the results continually improves with the use of feature selection methods. One method in particular, Pearson Correlation, stood out as being particularly effective, resulting in a significant increase in accuracy when compared to other methods. Out of all the classification techniques used, the RF Classifier demonstrated the highest accuracy, showing a 6.84% improvement when Pearson Correlation was employed. The accuracy of the GBM also saw an improvement of 4.52%, while the DT method showed a 7.2% increase. The KNN method saw a 5.21% increase in accuracy, and the LR method had the largest improvement of all, with an 8.61% increase in accuracy.

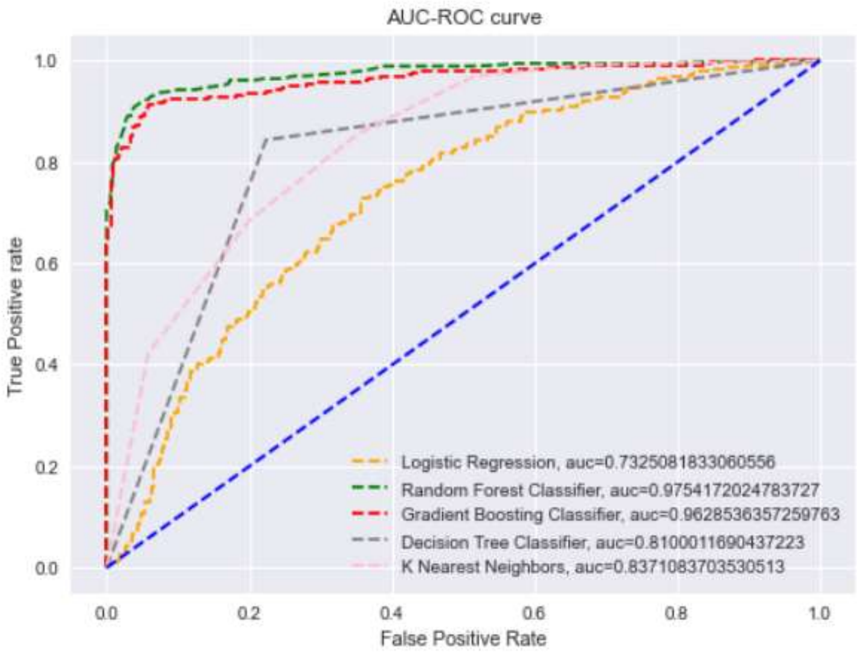
AUC-ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression, auc=0.7325081833060556
Random Forest Classifier, auc=0.9754172024783727
Gradient Boosting Classifier, auc=0.9628536357259763
Decision Tree Classifier, auc=0.8100011690437223
K Nearest Neighbors, auc=0.8371083703530513

Fig. 2

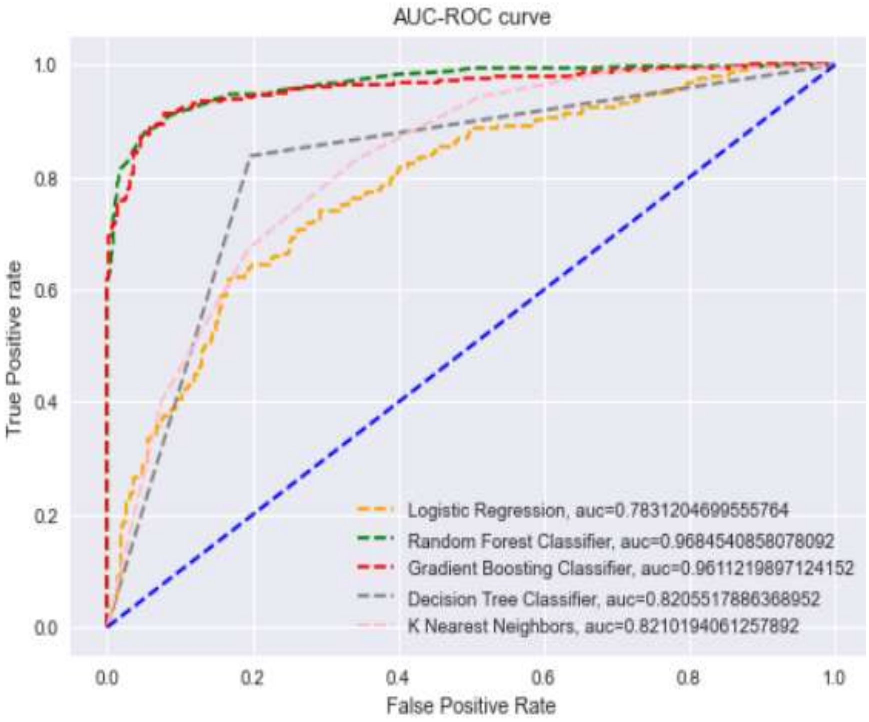
AUC-ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression, auc=0.7831204699555764
Random Forest Classifier, auc=0.9684540858078092
Gradient Boosting Classifier, auc=0.9611219897124152
Decision Tree Classifier, auc=0.8205517886368952
K Nearest Neighbors, auc=0.8210194061257892

Fig 3.

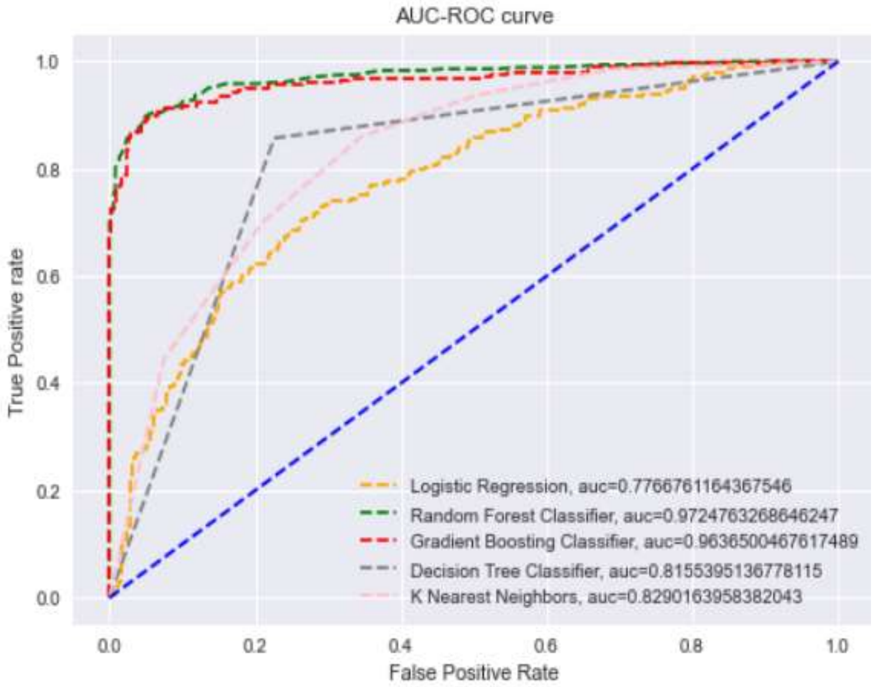
AUC-ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression, auc=0.7766761164367546
Random Forest Classifier, auc=0.9724763268646247
Gradient Boosting Classifier, auc=0.9636500467617489
Decision Tree Classifier, auc=0.8155395136778115
K Nearest Neighbors, auc=0.8290163958382043

Fig 4.

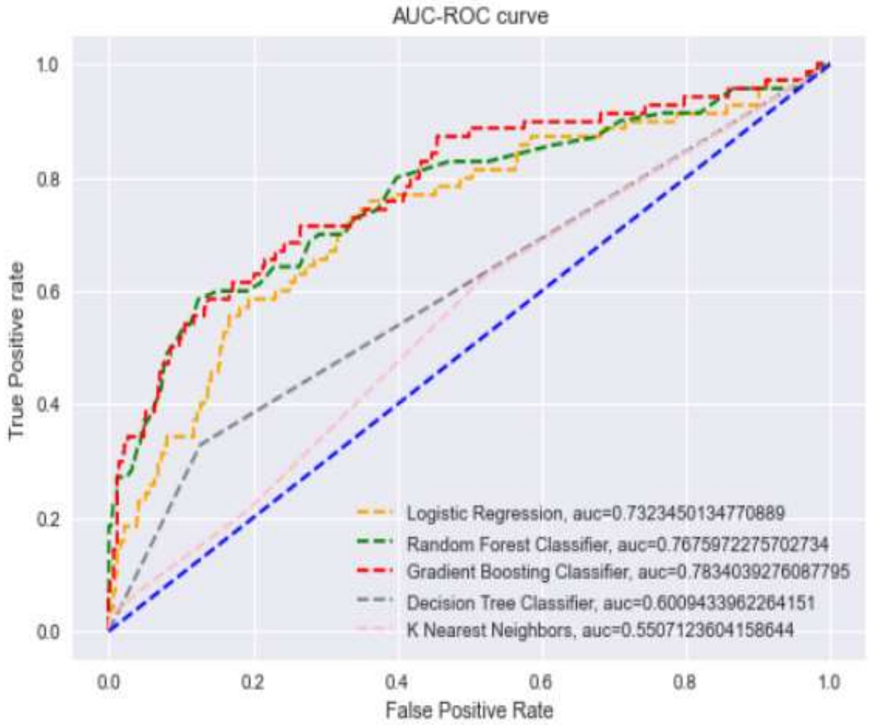
AUC-ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression, auc=0.7323450134770889
Random Forest Classifier, auc=0.7675972275702734
Gradient Boosting Classifier, auc=0.7834039276087795
Decision Tree Classifier, auc=0.6009433962264151
K Nearest Neighbors, auc=0.5507123604158644

Fig. 5

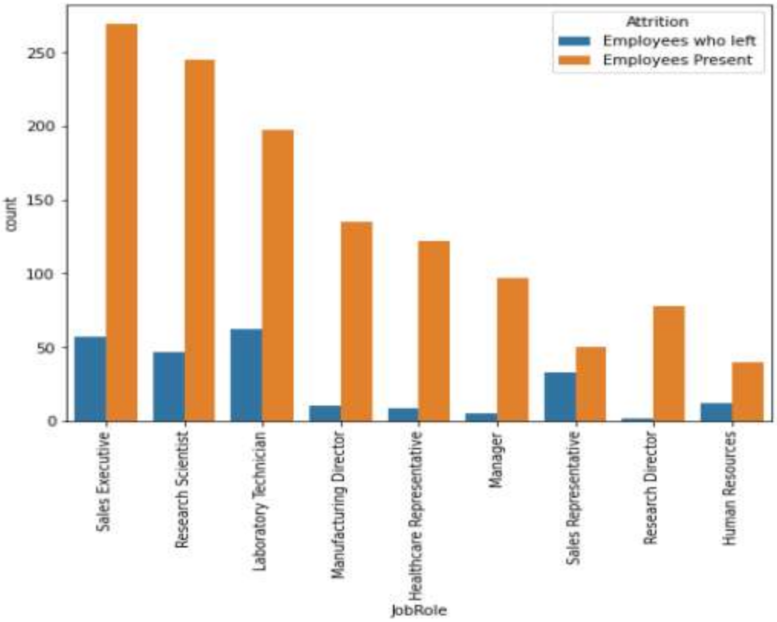

Fig 6.

To evaluate the performance of different classification thresholds, Receiver Operating Characteristic (ROC) curves are utilized, and the Area Under the ROC Curve (AUC-ROC) is calculated both with and without feature selection. The AUC-ROC summarizes the results of each threshold's confusion matrix, and a high AUC (close to 1) indicates that the model is highly separable. The closer the AUC is to 1, the better the model's performance. The results showed that the AUC-ROC of the GBM and RF algorithms was higher than the other algorithms in all cases, implying that the two ensemble algorithms are superior models. The AUC-ROC graph clearly indicates that GBM and RF are strong models, with AUC values close to 1, particularly when Pearson Correlation is used for feature selection. The ROC curves for the different algorithms are presented below in Figures 2 to 5 above.

6 Conclusion

The purpose of the study was to build a supervised machine learning model for employee attrition prediction. To do this, the study compared the performance of five different algorithms: LR, RF, GBM, DT, and KNN. These algorithms were evaluated both with and without feature selection to determine the impact of feature selection on each algorithm performance. The results of the study showed that the RF model had the highest accuracy and AUC (Area Under the Curve) when compared to the other algorithms. On the other hand, the GBM model performed the best on untreated data. The study also found that ensemble algorithms (algorithms that make predictions by combining the outputs of multiple models) showed greater predictive power. In terms of feature selection, the results showed that it improved the performance of the algorithms. The study found that the Pearson Correlation method was the most effective feature selection technique. The study concluded that the employee attrition prediction model built using machine learning can assist management in developing effective retention strategies. The results of this study suggest that future studies should explore the use of unsupervised machine learning and deep learning techniques for employee attrition prediction.

Acknowledgements

We acknowledge both the moral and technical support given by colleagues at the Sol Plaatje University, Department of Computer Science, Data Science, and Information Technology. This research was also partly funded through the CAIR (Centre for Artificial Intelligence Research) fund, grant agreement number: CSIR/BEI/HNP/CAIR/2020/10, supported by the Government of the Republic of South Africa, through its Department of Science and Innovation (DSI).

References

Pradhan, R.K., Jena, L.K. & Pattnaik, R. (2017). *Employee Retention Strategies in Service Industries: Opportunities and Challenges*. Employees and Employers in Service Organizations, Apple Academic Press, pp 53-70.

Khaled, A., Safeya, Z., Abdullah, A. & Usmani, T.M. (2021). *Employee Retention Prediction in Corporate Organizations using Machine Learning Methods*. Academy of Entrepreneurship Journal, Vol. 27, pp 3-4.

Xiaojuan, Z. (2016). *Forecasting Employee Turnover in Large Organizations*. PhD diss., University of Tennessee.

Fulmer, I.u & Ployhart, R. (2013). *Our Most Important Asset*. Journal of Management, Vol. 40, pp 2-3

Reyes, A., Aquino, C. & Bueno, D.C. (2019). *Why Employees Leave: Factors that Stimulate Resignation Resulting in Creative Retention Ideas*. Researchgate Publication, pp7-8.

Raza, A. (2022). *Predicting Employee Attrition Using Machine Learning Approaches*. Applied Sci. Journal, vol. 12.

Omar, A., Hossam, F., Khalid, J., Osama, H. & Ghatasheh, N. (2014). *Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis*. Life Science Journal. Vol. 11(3), pp 23-24.

Umayaparvathi, V. & Iyakutti, K. (2016). *A survey on customer churn prediction in telecom industry: datasets, methods and metric*. IRJET, Vol. 3, pp 12-16.

Amin , A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., & Hussain, A. (2016). *Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study*. IEEE Access. Vol. 4. pp.7940-7957.

Hebbar, A., Sanath, P., Rajeshwari, S., & Saqquaf, S. (2018). *Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees*. 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology. pp 934-938.

Dam, R.V. (2021). *Predicting Employee Attrition*. Masters Thesis, University of Tilburg.

Zhao, Y., Hryniewicki, M., Cheng, F., Fu, B. & Zhu, X. (2019). *Employee Turnover Prediction with Machine Learning: A Reliable Approach*. PhD thesis, University of Toronto.

Punnoose, R. & Ajit, P. (2016). *Prediction of employee turnover in organizations using machine learning algorithms*. International Journal of Advanced Research in Artificial Intelligence, Vol. 5 Issue 9, pp 32-33.

El-Rayes, N., Fang, M., Smith, M. & Taylor, S. (2020). *Predicting employee attrition using tree based models*. International Journal of Organizational Analysis, Vol 28, pp 33-36.

Yigit, I. & Sourabizadeh, H. (2017). *An Approach for Predicting Employee Churn by Using Data Mining*, IEEE.

Falluchi, F., Codalangelo, M., Giuliano, R. & William, L. (2020). *Predicting Employee Attrition Using Machine Learning Techniques*. Computers Journal, Vol. 9.

Ying, X. (2022). *An Overview of Overfitting and its Solutions*. IOP Publishing, Vol. 1168, No.2.

Gabrani, G. & Kwatra, A. (2018). *Machine Learning Based Predictive Model for Risk Assessment of Employee Attrition*". Computational Science and Its Applications, Springer International Publishing.

Saradhi, V. & Palshikar, G.K. (2011). *Employee churn prediction*. Expert Sys. with Applications, Vol. 38, no. 3.

Khare, R., Kaloya, D., Choudhary, C.K & Gupta, G. (2013). *Employee attrition risk assessment using logistic regression analysis*. Computational Science and Its Applications. Vol. 34.

Basha, M., Rajput, D. S., & Vandhan, V. (2018). *Impact of Gradient Ascent and Boosting Algorithm in Classification.* International Journal of Intelligent Engineering Systems, Vol. 11.

Sisodia, S., Vishwakarma, S. & Pujahari, A. (2017). *Evaluation of machine learning models for employee churn prediction.* International Conference on Inventive Computing and Informatics (ICICI) (pp. 1016-1020), IEEE.

Tharani, M. & Raj, V. (2020). *Predicting employee turnover intention in itites industry using machine learning algorithms.* Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud).

Gholamy, A., Kreinovich, V. & Kesheleva, O. (2018). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. Technical Report: UTEP-CS-18-09, University of Texas.

Pranto, B. (2022). *Entroppy Calculation, Information Gain & Decision Tree Learning*. Available at https://medium.com/analyticsvidhya/entropy-calculation-information-gain-decision-tree-learning-771325d16f..

Gupta, A. (2022). *Feature Selection Techniques in Machine Learning*. Available at: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning..

Sharma, N.V. & Yadav, N.S. (2021). *An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers*. Microprocessors and Microsystems Journal, Vol. 85