Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

Using machine learning algorithms with improved accuracy to analyze and predict employee attrition

Fiyhan Alsubaie[a] and Murtadha Aldoukhi[b]*

[a]National Commission for Academic Accreditation and Assessment, Riyadh, 13318, Saudi Arabia
[b]Industrial Engineering Department, College of Engineering and Architecture, Alyamamah University, Riyadh-11512, Saudi Arabia

CHRONICLE

Article history:
Received: August 12, 2023
Received in revised format: November 1 , 2023
Accepted: December 13, 2023
Available online: December 13, 2023

Keywords:
Machine Learning
Employee Attrition
Improve Model Accuracy
Prediction
Decision Tree
Random Forest
Binary Logistic Regression

ABSTRACT

Human migration is based on pull factors that individuals evaluate when it comes to moving to a different territory. Likewise, employee attrition is a phenomenon that represents the tendency to a reduction in employees within an organization. This research paper aims to develop and evaluate machine learning algorithms, namely Decision Tree, Random Forest, and Binary Logistic Regression, to predict employee attrition using the IBM dataset available on Kaggle. The objective is to provide organizations with a proactive approach to employee retention and human resource management by creating accurate predictive models. Employee attrition has significant implications for an organization's reputation, profitability, and overall structure. By accurately predicting employee attrition, organizations can identify the factors contributing to it and implement data-driven human resources management practices. This study contributes to improving decision-making processes, including hiring and firing decisions, and ultimately enhances an organization's capital. The IBM dataset used in this study consists of anonymized employee records and their employment outcomes. It provides a comprehensive HR data representation for analysis and prediction. Three machine learning algorithms, Decision Tree, Random Forest, and Binary Logistic Regression, were utilized in this research. These algorithms were selected for their potential to improve accuracy in predicting employee attrition. The Logistic Regression model yielded the highest accuracy of 87.44% among the tested algorithms. By leveraging this study's findings, organizations can develop predictive models to identify factors contributing to employee attrition. These insights can inform strategic decisions and optimize human resource management practices.

1. Introduction

Employee attrition refers to the reduction in the number of employees within an organization, whether voluntary or involuntary. Such attrition can occur due to retirement, resignation, termination, or the elimination of positions by the company. In today's business world, employee attrition poses significant challenges and chronic problems for organizations, impacting their competitiveness.

When employees leave an organization, it results in a loss of institutional knowledge, decreased morale, and increased workload for remaining employees. Furthermore, the departure of a significant number of employees may raise concerns for potential applicants. Ultimately, these factors can adversely affect an organization's productivity, profitability, and reputation among customers and stakeholders. In 2021, the Labor Statistical Bureau reported the rate of employee attrition in the United States to be 57.3%.

Business leaders have recognized the profound impact of employee attrition. Human resources are a crucial and valuable asset for any organization. However, many organizations still perceive human resources as an expense rather than a strategic investment. Effective utilization of employees is vital since their efficiency directly impacts the organization's success

* Corresponding author.
E-mail address: m_aldoukhi@yu.edu.sa (M. Aldoukhi)

doi: 10.5267/dsl.2023.12.006

(Branham, 2005). A high attrition rate indicates employee dissatisfaction and a lack of labor force stability, which is detrimental to an organization's competitiveness and long-term growth. It also leads to uncertain costs, disruptions in production and work atmosphere, along with expenses related to recruitment, selection, training, and development. For organizations, understanding the reasons for attrition is essential. While decisions made by employees are often seen as the primary cause, organizations themselves play a crucial role, as their actions and policies influence employee decisions (Dalton & Mesch, 1990). Machine learning algorithms offer a promising solution for predicting employee attrition. However, the accuracy of these algorithms is critical for reliable results that can impact decision-making processes. Accurate prediction of employee attrition benefits organizations by identifying its root causes, fostering an engaged workforce, and improving overall organizational performance.

This study aims to analyze various factors affecting employee attrition in organizations, such as staff training and development, performance appraisal, staff attitude, and delegation of duties. Additionally, we employ Decision Tree, Random Forest, and Binary Logistic Regression machine learning algorithms, which have demonstrated improved accuracy, to predict employee attrition. By identifying the factors contributing to attrition, organizations can make data-driven decisions and take proactive steps to retain employees. For instance, recruitment efforts can be focused on attracting candidates who exhibit characteristics associated with longer tenures.

The remainder of this paper is organized as follows: Section 2 presents the related work, Section 3 describes the methodology and initial results, Section 4 discusses the improved model accuracy and provides analysis, and Section 5 concludes the study.

2. Related Work

This section presents the related works that discuss and analyze employee attrition. In general, machine learning has been actively used in different areas, including but not limited to, malware detection in IoT-based enterprise information systems (Gaurav, Gupta, & Panigrahi, 2023), phishing website detection with semantic features (Almomani et al., 2022) and assisting security and privacy issues in healthcare (Wassan et al., 2022). In addition, one of the important uses of machine learning is to predict employee attrition. Liu (Liu, 2014), presented a case study from the Chilean labor market, in which he studied employee turnover by analyzing 112 responses. Another study of employee attrition was conducted by Nagadevara and Srinivasan (Nagadevara & Srinivasan, 2007), in which they evaluated the impact of demographic attributes and employee absenteeism on attrition. The results they obtained after improving the model accuracy of the Logistic Regression, for example, is 79.58%.

Additionally, Rombaut and Guerry (Rombaut & Guerry, 2018) focused on work-specific factors. The logistic Regression algorithm was utilized by Ponnuru, Merugumala, Padigala, Vanga, and Kantapalli (2020) to predict employee turnover. The authors in (Najafi-Zangeneh, Shams-Gharneh, Arjomandi-Nezhad, & Zolfani, 2021) studied the reason for employee attrition with a focus on feature selections. However, they only used the logistic regression model, without any focus on model accuracy improvement. Although the authors in (Fallucchi, Coladangelo, Giuliano, & De Luca, 2020) and (Qutub, Al-Mehmadi, Al-Hssan, Aljohani, & Alghamdi, 2021) used a variety of models to predict employee attrition, these were basic models and did not address methods for improving accuracy.

Bhartiya, Jannu, Shukla and Chapaneri (2019) followed a structured methodology to perform their study, starting with acquiring the data and finishing with using different machine learning algorithms. Among the machine learning algorithms they used, Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, and Naive Bayes, the Forest classifier provided the highest accuracy at 83.3%. A similar study was conducted by the authors in (Joseph, Udupa, Jangale, Kotkar, & Pawar, 2021) to study emotional evaluation and its impact on employee attrition. They obtained an accuracy score of 86% to predict the attrition rate. The authors in (Alao & Adeyemo, 2013) analyzed the data of 309 employees who were employed in and left the Higher Institutions in Nigeria between 1978 and 2006. They also used the decision tree algorithm to predict employee attrition, achieving a best accuracy rate of 74%. They found that employee attrition was primarily influenced by salary and length of stay. For the same purpose, a deep learning technique was used by the authors in (Al-Darraji et al., 2021) to study employee attritions, and they obtained 94% accuracy in their predicting model. A European-wide survey was conducted by the authors in (Lazzari, Alvarez, & Ruggieri, 2022) to study the reasons behind employee turnover. In their study, they used logistic regression and LightGBM. A collection of different researchers also worked on studying the factors affecting employee attrition and predicting if an employee will stay in the organization or leave in (Subhashini & Gopinath, 2020; Vasa & Masrani, 2019; Zhao, Hryniewicki, Cheng, Fu, & Zhu, 2018).

As shown in Table 1, this study builds on the available literature intending to improve the models' accuracy by using different methodologies where, to the best of our knowledge, they were not used in similar studies. These methodologies included the pruning method for the Decision Tree model, the cross-validation method for the Random Forest model and the stepwise method for Binary Linear Regression. Consequently, this would affect the strategic decisions organizations may take to deal with employee attrition issues.

Table 1
Literature review summary

	Machine Learning Algorithm				Evaluating and Improving Model Performance Techniques		
Authors	Decision Tree	Random Forest	Binary Logistic Regression	Number of factors considered	Pruning	Cross-validation	Stepwise
Liu, 2014	X	X	X	3	X	X	X
Nagadevara and Srinivasan, 2007	X	X	✓	34	X	X	X
Rombaut and Guerry, 2018	✓	X	✓	13	X	X	X
Ponnuru et al., 2020	X	X	✓		X	X	X
Najafi-Zangeneh et al., 2021	X	X	X	32	X	X	X
Fallucchi et al., 2020	✓	✓	✓	35	X	✓	X
Qutub et al., 2021	✓	✓	✓	35	X	✓	X
Bhartiya et al., 2021	✓	✓	X	35	X	X	X
Joseph et al., 2021	✓	✓	✓		X	X	X
Alao & Adeyemo, 2019	✓	X	X		✓	X	X
Al-Darraji et al., 2021	✓	✓	✓	35	X	✓	X
Lazzari et al., 2022	✓	✓	✓		X	✓	X
Vasa and K. Masrani, 2019	✓	✓	✓		X	✓	X
Subhashini and Gopinath, 2020	✓	✓	X	35	X	X	X
Zhao et al., 2018	✓	✓	✓		X	✓	X
Our Study	✓	✓	✓	35	✓	✓	✓

3. Methodology

In this section, we present the methodology we employed to solve the problem, starting with data collection and data exploration. Subsequently, we explain the machine learning algorithms utilized, and the initial results obtained. Table 2 provides a comprehensive list of all the variables and factors utilized in this study, along with their respective descriptions. These factors were identified and selected based on existing research and previous studies, which have frequently associated them with employee attrition.

Table 2
Variables and corresponding levels

Variable	Type	Ranges/Factor Levels
Age	Continuous	18-60
Attrition	Nominal	0= No,1=Yes
Business Travel	Ordinal	1=Non-Travel, 2=Travel_Rarely, 3=Travel_Frequently
Daily Rate	Continuous	102-1499
Department	Nominal	1=Human Resources, 2=Research & Development, 3=Sales
Distance from Home	Continuous	1-29
Education	Ordinal	1=Below College, 2= College,3= Bachelor, 4=Master,5 = Doctor
Education Field	Nominal	1=Human Resource,2 = Life Sciences, 3= Marketing, 4=Medical, 5= other, 6= Technical Degree
Environment Satisfaction	Ordinal	1= Low, 2=Medium, 3-High, 4=Very High
Gender	Normal	0 =Female, 1=Male
Hourly Rate	Continuous	30-100
Job Involvement	Ordinal	1= Low, 2=Medium, 3-High, 4=Very High
Job Level	Ordinal	1,2,3,4,5
Job Role	Ordinal	1=Sales Executive ,2= Research Scientist, 3= Laboratory Technician,4= Manufacturing
Job Satisfaction	Ordinal	1= Low, 2=Medium, 3-High, 4=Very High
Marital Status	Nominal	1= Divorced, 2=Married, 3=Single
Monthly Income	Continuous	1009-19999
Monthly Rate	Continuous	2094-26999
Num Companies Worked	Continuous	0-9
Overtime	Nominal	0=No,1=Yes
Percent Salary Hike	Continuous	11-25
Performance Rating	Ordinal	1= Low, 2= Good, 3=Excellent, 4=Outstanding
Relationship Satisfaction	Ordinal	1= Low, 2=Medium, 3-High, 4=Very High
Stock Option Level	Ordinal	0,1,2,3,4
Total Working Years	Continuous	0-40
Training Times Last Year	Continuous	0-6
Work-Life Balance	Ordinal	1= Bad,2=Good,3=Better,4=Best
Years at Company	Continuous	0-40
Years in Current Role	Continuous	0-18
Years Since Last Promotion	Continuous	0-15
Years with Current Manager	Continuous	0-17

3.1. Data Collection

This study utilizes a descriptive study design to comprehensively explore the phenomenon of employee attrition. The data for this study is sourced from the Kaggle website (Subhash Pavan, 2017). The dataset contains information on 1,471 employees and encompasses 35 variables. Additionally, relevant data pertaining to factors influencing attrition has been recorded. This study considers several key factors, including age, as older employees may be more inclined to retire or leave the company. Moreover, the study also takes into account the years of experience at the company, as individuals who have been with the organization for an extended period may be less likely to leave, while those with less tenure may be more prone to attrition. Distance from home is another factor considered, as employees with longer commutes may be more inclined to explore alternative employment opportunities closer to their residence. Work-life balance is also examined, recognizing that employees who perceive a poor work-life balance may be more prone to considering job changes. Lastly, the factor of compensation is analyzed as it is considered a factor relevant to attrition.

The limitations of this study are considered during the data collection process, according to the source we obtained the data from (Kaggle). These limitations can be summarized as follows: Firstly, some of the employees may be uncooperative in filling out the questionnaires that were issued to them. This may be due to personal reasons which may not be prevented but minimized. Secondly, there may be instances where employees fail to return their questionnaires on time. Thirdly, some employee responses may lack the depth of information as per the expectation of the researcher. Finally, another potential limitation is the lack of employee commitment to providing research-related information due to their daily duties in the organization.

3.1. Data Cleaning and Processing

Data processing and cleaning were implemented in the R software. The dataset was found to have no missing values. Variables such as employee count, number of employees, and standard work hours are constants. The variables found to be non-informative about the variability of attritions are removed from the dataset. Conversely, the employee number variable, which is a unique ID for every observation, is removed due to the potential of leading to over-fitting. Nominal variables, loaded as character strings, were converted to factor variables. Ordinal variables, on the other hand, were converted to ordered factors.

3.2. Data Exploration

To gain insights into the underlying features of the dataset, an exploratory analysis was conducted. This study employed graphical representations and descriptive analysis to examine the dataset.

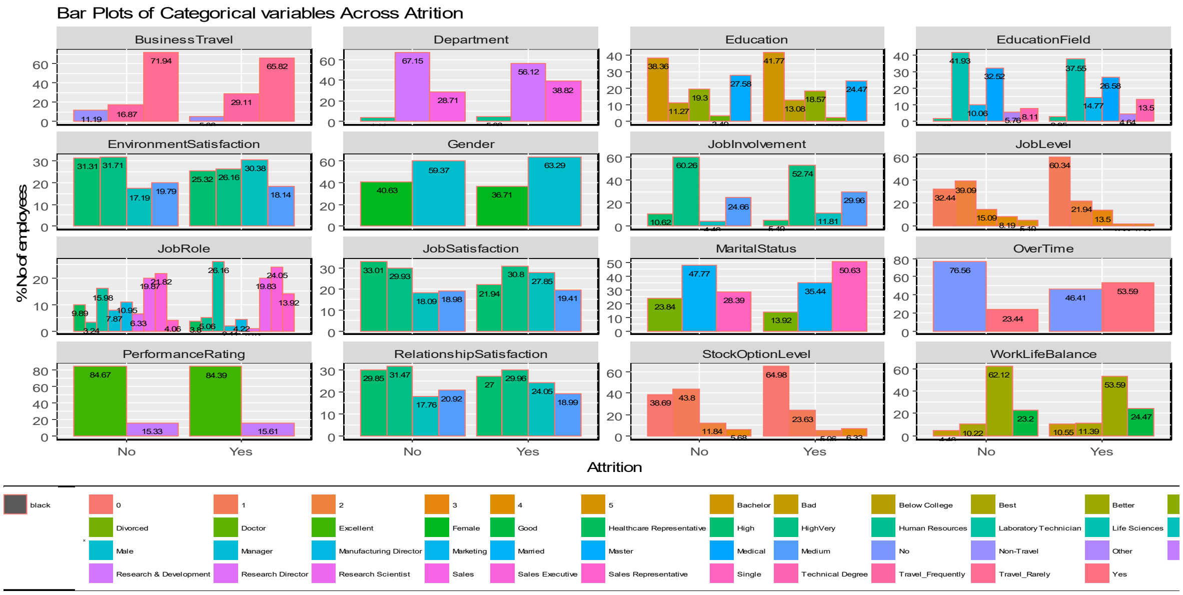


Fig. 1. Bar graphs of categorical variables

Fig. 1 presents bar graphs for all categorical variables (nominal or ordinal) in both the attrition and non-attrition groups. The bars are normalized within each group. Variables that show a strong association with attrition exhibit significant differences in percentage frequency distribution between the two groups. Conversely, variables such as gender, business level, performance rating, work-life balance, education, and education field have similar shapes in both the attrition and non-attrition groups, indicating that the percentage frequencies are relatively consistent across the two groups. As a result, it is expected that these variables have a weak or no significant association with attrition. However, variables like overtime, stock options level, job level, and job role demonstrate significant differences in frequency distribution between the two attrition groups, suggesting a strong relationship with attrition. Table 3 provides descriptive statistics for the continuous variables. For variables such as age and daily rate, where the mean and median values are close, we can conclude that they have an approximately symmetrical distribution.

Table 3
Descriptive statistics

Predictor	n	Min	Q1	median	Mean	Q3	Max
Age	1470	18	30	36	36.92	30	60
Daily Rate	1470	102	465	802	802.49	465	1499
Distance from Home	1470	1	2	7	9.19	2	29
Hourly Rate	1470	30	48	66	65.89	48	100
Monthly Income	1470	1009	2911	4919	6502.93	2911	19999
Monthly Rate	1470	2094	8047	14235.5	14313.1	8047	26999
Num Companies Worked	1470	0	1	2	2.69	1	9
Percent Salary Hike	1470	11	12	14	15.21	12	25
Total Working Years	1470	0	6	10	11.28	6	40
Training Times Last Year	1470	0	2	3	2.8	2	6
Years at Company	1470	0	3	5	7.01	3	40
Years in Current Role	1470	0	2	3	4.23	2	18
Years Since Last Promotion	1470	0	0	1	2.19	0	15
Years with Current Manager	1470	0	2	3	4.12	2	17

In Fig. 2, the boxplot of the continuous variables (age, daily rate, hourly rate, and monthly income) across the attrition groups indicates an approximate symmetry. The mean and standard deviation provide estimates of the central tendency and dispersion of these variables, which are measured on a monetary scale. On the other hand, variables such as years since the last promotion, total years of work, training time last year, years at the company, years in the current role, and years spent with the current manager exhibit negative skewness, with outliers at the upper extreme. This can be attributed to the fact that most individuals spend a short amount of time with their current managers, and only a few have been under the same supervisor for an extended period. All these variables are measured in years. The findings from these variables suggest a low retention level, as most employees in the sample spend little time under the same supervisors and have durations shorter than the median. Additionally, training times are generally short for most employees. In this case, the median and interquartile range quartiles are the most appropriate measures. Furthermore, the variable "distance from home" exhibits negative skewness, indicating that the majority of the sampled individuals do not work far from their homes.

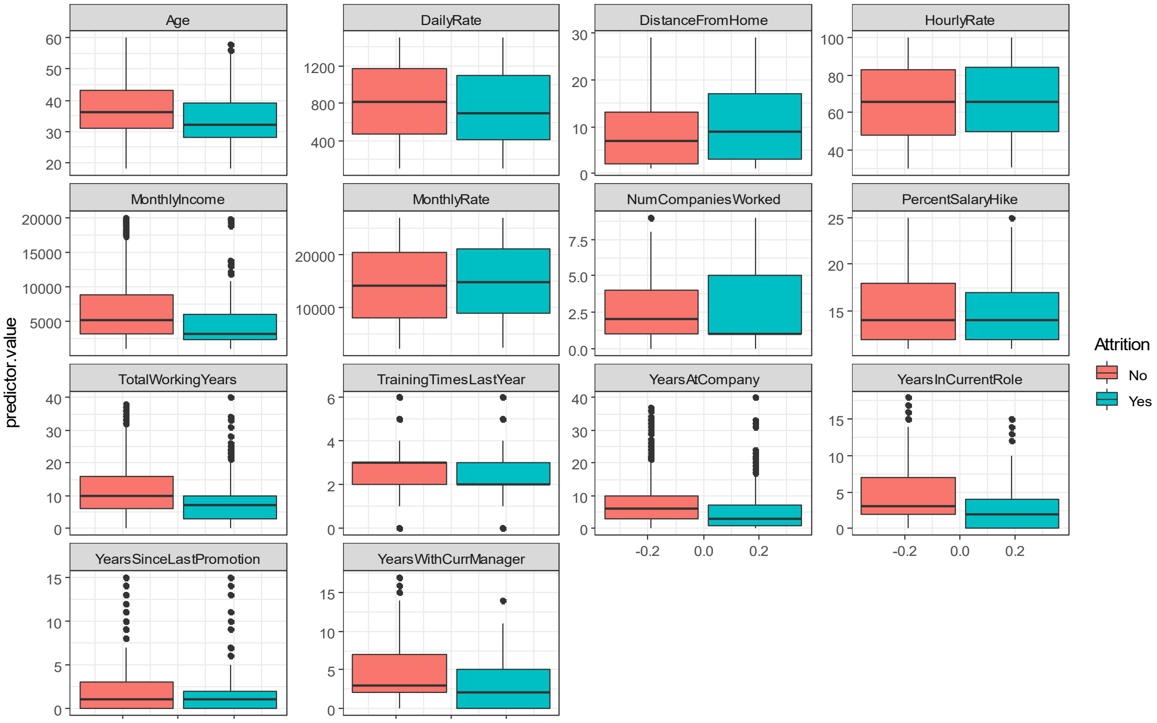


Fig. 2. Box plots of the continuous predictors

3.3. Machine Learning Algorithms

In this section, three models were used to predict employee attrition. The models utilized include the Decision Tree, Random Forest, and Binary Logistic Regression model. Besides the reasons for their interpretability and ease of use, these models were selected according to the nature of the response variable (employee attrition), which is a binary outcome. The Decision Tree and Random Forest algorithms have a common advantage which is they are effective at identifying the most relevant variables for predicting employee attrition. The Binary Logistic Regression algorithm also has the capability of providing a probabilistic interpretation of the results. Further discussion about improving model accuracy is available in the results section. This study made use of R software to run our models. Initially, the models were fitted using the default R setting. Later, to improve the fitted models and increase the accuracy of the models, we used different statistical procedures explained in each algorithm. The data was initially split randomly into two subsets: a training set with 1000 observations, and the remaining 470 observations were reserved as the testing set. The training set was used to fit the models, while the testing set was utilized in testing and calculating the model accuracies. In general, the practicality and accessibility of implementing these models may vary depending on the organization's existing HR systems and processes, as well as the availability and quality of data. To deploy these models for predicting employee attrition in organizations, it is important to consider the availability of sufficient high-quality employee data, the integration with the current system and consideration of ethical and legal implications. Ignoring these elements may limit the adoption and utilization of these models in organizations.

3.3.1. Decision Tree Model

A Decision Tree is a powerful algorithm used for classification and regression purposes. In the R software, we used the RPART library to build the Decision Tree algorithm. The algorithm uses recursive partitioning, which is a statistical method to split the training set so that the outcome in each subgroup (node) remains as homogeneous as possible. The algorithm splits the response variable according to the predictor variables, starting with the predictor having the highest association with the response. The splitting continues until specific thresholds are reached. Homogeneity is measured by the overall degree of node impurity. Two famous measures can be used for node impurity; the Gini index and entropy, with Gini being used on this occasion. For more information about the origins of the Gini index and its usage, we refer the reader to (Ceriani & Verme, 2012). The Gini index formula is expressed as follows:

$$Gini = \sum (P * (1 - P)) \tag{1}$$

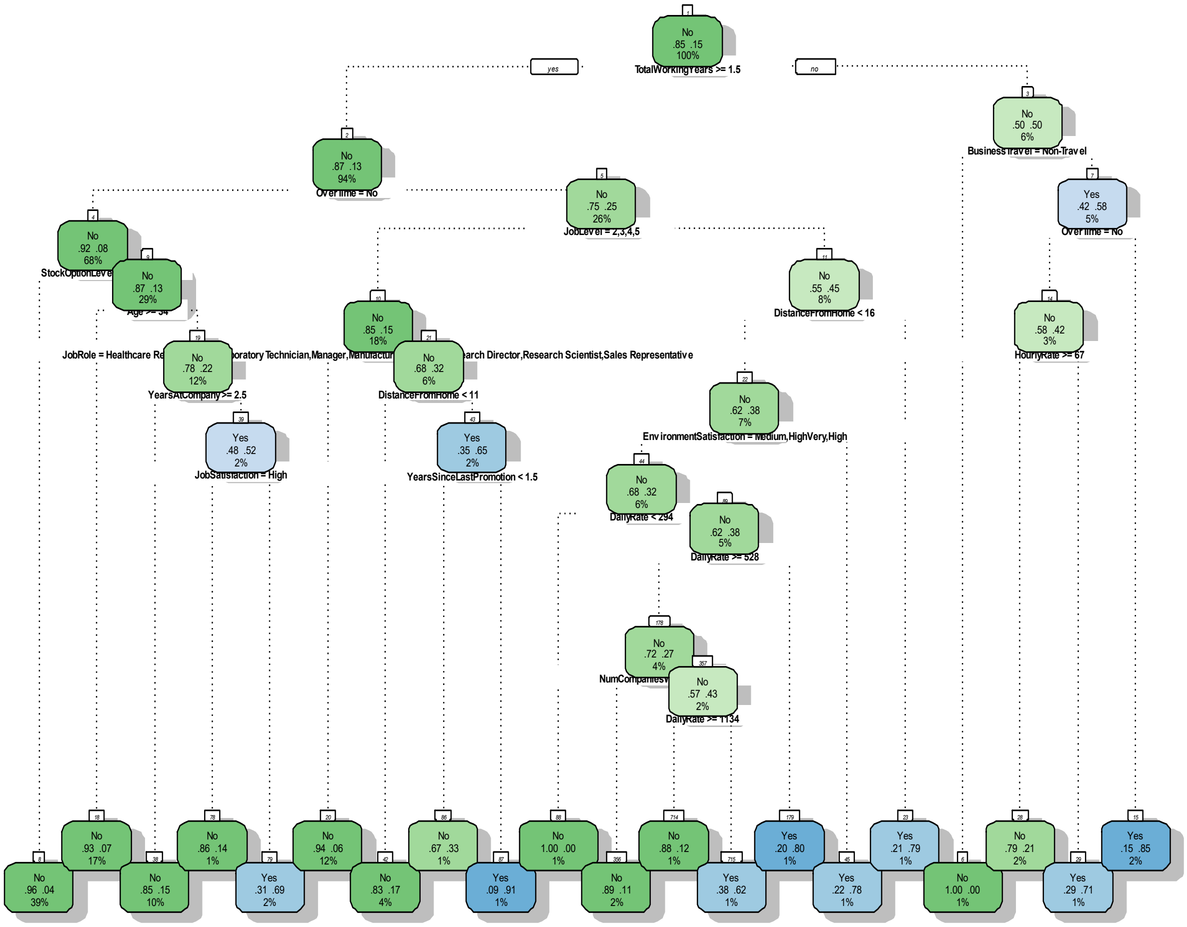
where P is the probability of misclassification (proportion of misclassification on a given node) while the model aims to minimize these measures. It is essential to determine when to stop the splitting process as excessively deep trees can lead to over-fitting of the data. The splitting process concludes when all the leaf nodes become pure, and the number of observations on the leaf hits the pre-specified minimum or the pre-specified minimum number of training observations that can't be allocated to every leaf node with any splitting method. The nsplit parameter is the minimum number of observations in a node for a split to be attempted. In contrast, the minimum bucket is the minimum number of observations in any terminal. By default, R software sets the split value to be 20 and the minimum bucket to be a third of the nsplit, as shown in Fig 3.

The root node is the group that contains all the training data, with 15% of the cases being attrition cases, while 85% are no attrition cases. This tells us that, before considering the extra predictor information, we know that the probability of attrition is 15%. The predictor variables are additional case information and are expected to reduce uncertainty about the outcome class. Attempts to include the predictors are therefore expected to increase prediction accuracy.

Total working years is the predictor that has the highest association with attrition. The training set (root node) was split based on total working years to create two groups; a group for employees who had total working years equal to 1.5 or more (node 2) and the other group with working years less than 1.5 (node3). The first group had 938 (94%) of the total employees on the training set. Among these 938 employees, 13% were attrition cases, while the remaining 87% had no attrition. Of the 118 (6%) employees in node 3, 50% were attrition cases, while the remaining 50% were not. At node 23 (employee with working years <1.5), business travel is the predictor that has the highest association with attrition. This group was then subdivided into two smaller groups; these are nodes 6 and 7 respectively. At node six, there are only 9 (1% of the training set), and all are non-attrition cases. In node 7 there are 22 cases, of which 42% are attrition cases, while 58% are not. Node 6 cannot split further because the nine observations are less than 20, which is the default nsplit. Also, the node is 100% pure.

Conversely, the variable with the highest association with attrition at node 7 is overtime. Therefore, the node was split further into two nodes: node 14, comprising employees from node 7 who do not get overpayment, and node 15, comprising employees from node 7 who receive an overpayment. In node 14, 58% are attrition cases, while 42% are not attrition cases. Conversely, in node 15, 15% are attrition cases, while 85% are not attrition cases. Node 15 has 20 observations which is the split; hence cannot be split further.

As shown in Table 4, predicting the testing set with the full model, the model's specificity is 94.8%, while the sensitivity is 34.1%. The overall model accuracy is 83.4%.



Rattle 2019-Dec-03 21:59:28 User

Fig. 3. The fitted tree

Table 4
Metric summary of specificity and sensitivity of the Decision Tree model

		Actual		
		No	Yes	Total
Predicted	No	362	58	420
	Yes	20	30	50
	Total	382	88	470

3.3.2. Random forest Model

A random forest classifier is a decision tree model built by several decision trees from bootstrap samples. During prediction, each decision tree is allowed to make a prediction, and subsequently, the most predicted (most voted) category becomes the model prediction. To achieve high accuracy, the bootstrap sampling needs to be controlled by some basic parameters. One is the number of trees, and the other is the minimum number of variables to be considered before splitting a node. The first parameter is called ntree in R software and it controls the number of trees used. The second parameter, mtry, is the number of variables tried at each split.

The initial run is the default R model, where it tries a random sample of 5 variables at each split and builds 500 trees. The out-of-bag error rate is 13.1%, representing the error percentage. For more information about how to estimate the out-of-bag, we refer the reader to Breiman (Breiman, 1996). Table 5 presents the specification of the initial model.

Table 5
Summary of the initial model

Model Specification	
Number of Trees	500
Mtry	5
Out of Bag Error Rate	13.1%

Model importance was calculated based on mean Gini decrease; Gini importance is the probability that a new data point will be wrongly classified on a given node. Gini importance is leveraged to calculate mean Gini importance which is a measure of variable importance. It measures how having a given variable on the model decreases the chance of the wrong prediction for any new data point. The variable with the highest mean Gini decrease is monthly income, the variable can be assumed to be among the drivers of attrition, and the model weakens with high magnitude by not having the variable on the model. The second in importance is age while the least important is performance rating. A full breakdown is available in Fig 4.

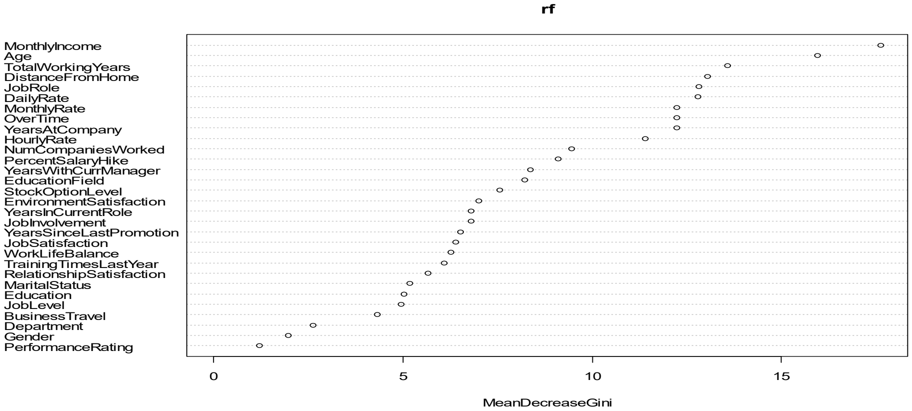


Fig. 4. Graph of variable importance

As shown in Table 6, by predicting employee attrition using the test dataset, 99.21% (379 of 382 no attritions) were correctly predicted, and 13.64% (12 of 88 attritions) were also correctly predicted. The model has, therefore, a higher specificity but a low sensitivity. The overall correct prediction is 83.4%.

Table 6
Metric summary of specificity and sensitivity of Random Forest model

		Actual		
		No	Yes	Total
Predicted	No	379	76	455
	Yes	3	12	15
	Total	382	88	470

3.3.3. Binary Logistic Regression Model

A binary logistic regression model analyzes datasets with one or more independent variables and a binary dependent variable. The model assumes a linear relationship between the log of odds and the independent variables. The odds here refer to the ratio between the probability of an event happening to the probability of it not happening. The general equation for the model as follows:

$$\text{Log}\ (\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + e \qquad (2)$$

where for this study, p = probability of attrition (i.e., attrition =Yes), β_0 ... β_n are the regression coefficients, while X_1...X_n are the independent variables and e is the error. From this relationship, the predictions of the probability of attrition are achieved by making up the subject of the following formula:

$$p = \frac{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}{1 + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n} + e \qquad (3)$$

The model assumes a linear relationship between the log of odds. This assumption makes sense for only the numerical variables. A fair linear relationship is evident from the scatter plots below in Fig. 5. The relationship is stronger for age, monthly income, and total working years.

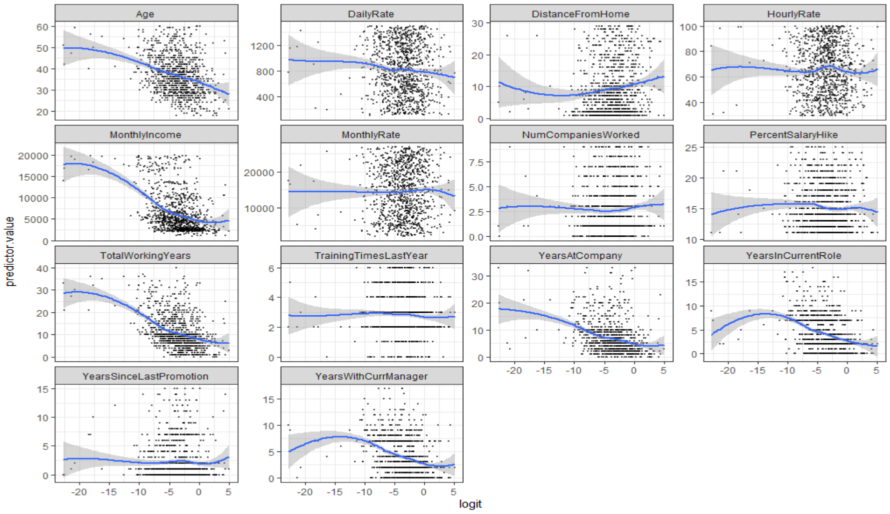


Fig. 5. Scatter plots of predicted values versus logit (log of odds)

Data used for this study is a random sample, where each observation is an employee. Therefore, it is reasonable to assume that the observations are independent.

Generalized variance inflation factors (GVIF) is the applicable measure of multicollinearity when factor variables are involved. The variable department was above the recommended 5 by the rule of thumb. No other variable had a $GVIF^{\left(\frac{1}{2*Df}\right)}$ greater than 5**.** The monthly rate and job role have a high $GVIF^{\left(\frac{1}{2*Df}\right)}$, although they don't exceed the threshold of 5. Attempts to add an interaction complicate the model further, and the interaction terms' coefficients couldn't be computed. Therefore, the job role is expected to be related to the department (the roles assigned to an employee depend on the department they are in). The data support the idea because dropping either the department or the job role impacts decisions. The other variable $GVIF^{\left(\frac{1}{2*Df}\right)}$ remain below 5. The variable department was therefore dropped from the model. Table 7 contains all the details about the generalized variance inflation factors.

Table 7
Generalized variance inflation factors

Predictor	GVIF	Df	$\mathrm{GVIF}^{\left(\frac{1}{2*\mathrm{Df}}\right)}$
Age	2.10	1	1.45
Business Travel	1.44	2	1.09
Daily Rate	1.13	1	1.06
Department	**55260610.00**	**2**	**86.22**
Distance from Home	1.25	1	1.12
Education	1.98	4	1.09
Education Field	5.12	5	1.18
Environment Satisfaction	1.74	3	1.10
Gender	1.16	1	1.08
Hourly Rate	1.18	1	1.09
Job Involvement	1.61	3	1.08
Job Level	111.58	4	1.80
Job Role	1741493000.00	8	3.78
Job Satisfaction	1.56	3	1.08
Marital Status	3.38	2	1.36
Monthly Income	15.58	1	3.95
Monthly Rate	1.17	1	1.08
Num Companies Worked	1.58	1	1.26
Overtime	1.51	1	1.23
Percent Salary Hike	2.46	1	1.57
Performance Rating	2.42	1	1.56
Relationship Satisfaction	1.70	3	1.09
Stock Option Level	4.10	3	1.27
Total Working Years	5.80	1	2.41
Training Times Last Year	1.11	1	1.06
Work-Life Balance	1.53	3	1.07
Years at Company	7.29	1	2.70
Years in Current Role	3.49	1	1.87
Years Since Last Promotion	3.00	1	1.73
Years with Currency Manager	3.54	1	1.88

As shown in Fig. 6, a graph of cook distance shows some extreme values. Outliers are therefore present in this dataset. It is determined that standardized error values exceeding ±3 were flagged as points of influence to identify influential outliers. There are 4 influential outliers, although their values don't seem far from the other variables, as shown in Fig. 7.

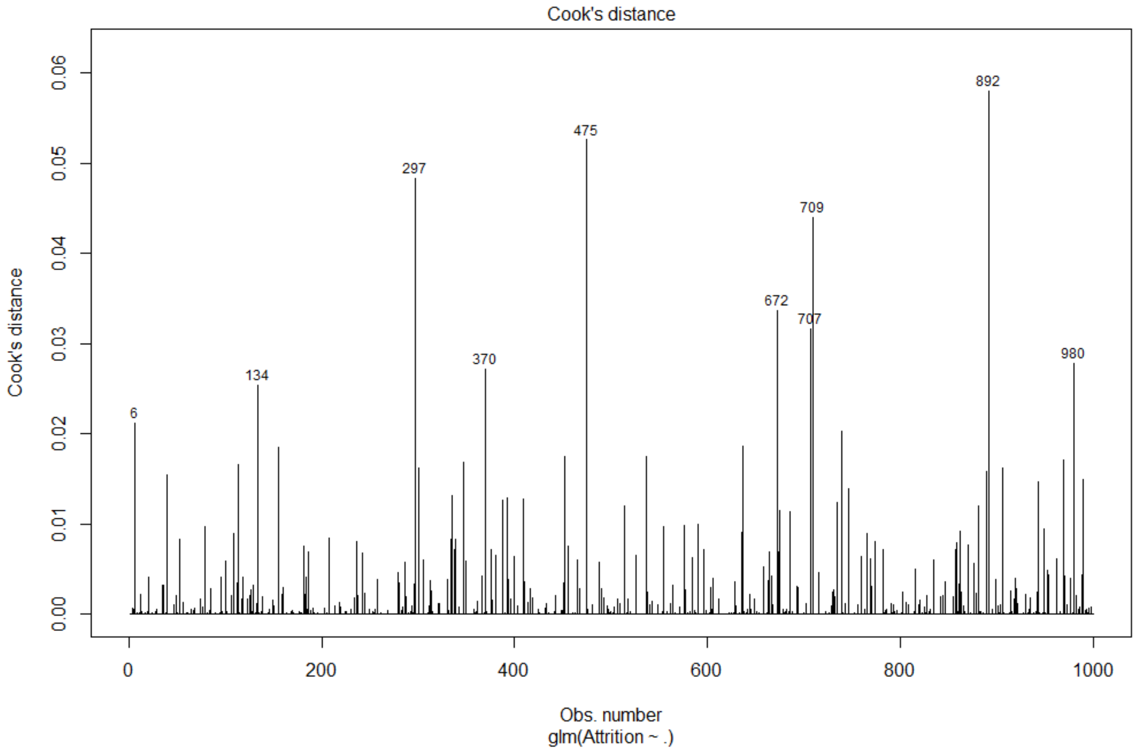


Fig. 6. A graph of cook distances

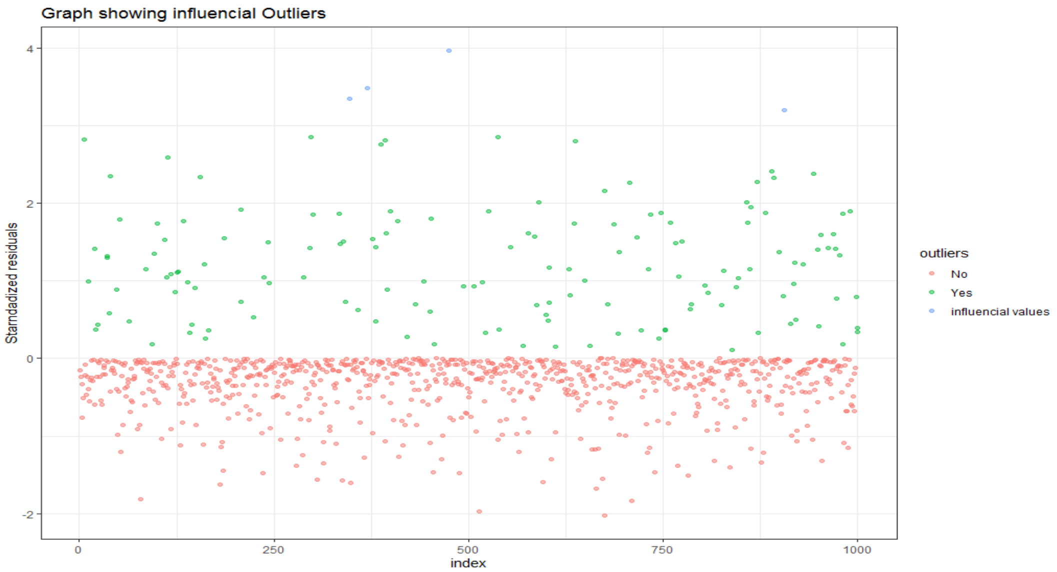


Fig. 7. Influential outliers

The significance of the model explains the importance of the independent variables to the model. It serves as a test to determine whether adding our predicted values improves the prediction. The test compares a model with all the predictors against a model with no predictor variable (Null/intercept only model). The two models are compared based on their deviance from a saturated model. It is assumed that the only available information is about the response variable (attrition), and no information is known about the predictors. As shown in Table 8, there were only $\frac{149}{1000}$ cases of attrition for the training dataset and $\frac{851}{1000}$ cases without attrition.

Table 8
Frequency distribution of attrition

Attrition	n	Prob
No	851	0.851
Yes	149	0.149

The odds of attrition to no attrition are therefore $\frac{0.149}{0.851} = 0.1750881$. Since this value is less than 1, the chance of an employee leaving the company is less than staying. The log of odds (Log $(\frac{p}{1-p})$) is -1.74247. This is the intercept coefficient, and it estimates the expected log (odds), assuming that the values of all the independent variables are 0. At the 5% level, the p-value is significant ($p < 2 * 10^{-16}$), which means that there is a significant difference between the chance of attrition and no attrition. The information is cast in Table 9 below.

Table 9
Regression coefficient of the null model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.74247	0.08881	-19.62	$<2 * 10^{-16}$***

Comparing the above null model to a model with all predictors is done based on deviance. The difference between the deviance of the two models is assumed to follow a chi-square distribution. At a 5% level of significance, there is a significant difference between the two models (χ2(60) = 380.95, $p < 2.2 * 10^{-16}$), as shown in Table 10. The independent variables, therefore, significantly improve the response variable's prediction.

Table 10
Test of model significance

	Df	Deviance
Null model	999	841.94
Full model	939	460.99
Diff	60	380.95
sig		$< 2.2 * 10^{-16}$***

As shown in Table 11, by predicting employee attrition using the test dataset, 94.5% (361 of 382 no attritions) were correctly predicted, and 52.27% (46 of 88 attritions) correctly predicted the model in which it has a higher specificity but higher low sensitivity. The correct overall prediction is 86.60%.

Table 11
Metric summary of specificity and sensitivity of Binary Linear Regression model

		Actual		
		No	Yes	Total
Predicted	No	361	42	403
	Yes	21	46	67
	Total	382	88	470

4. Results and analysis

In this section, the obtained results of the improved accuracy models are presented. In addition, an analysis of the results and effects on the variables is also presented.

4.1. Improving the accuracy of the Decision Tree model

Fitting a full-grown tree is not advisable because the tree tends to over-fit and overemphasize small splits that do not significantly minimize Gini. In order to mitigate the over-fitting in this model, we used the pruning methodology. Pruning is a method of removing such nodes to achieve a smaller (less complex) tree whose performance is close to that of a full-grown tree. This way, we improve interpretation at the cost of a slight bias. Pruning in the RPART library is controlled by the complexity parameters (CP). A small CP leads to a deep-grown tree with a risk of over-fit. In contrast, a significant CP leads to a small tree whose performance is small because we are likely to remove virtual nodes. The CP was decided by trying several CPs and then choosing the complexity parameter that leads to minimum cross-validation error. For this data, using 10-fold cross-validation means that the training set is randomly split into 10 subgroups. These subgroups will contain around 100 observations each. Each time we omit one subgroup and fit a decision tree on the 900 observations. For each value of CP to be tested, the prediction of the remaining 100 observations is made, and the correct classification for each CP value is calculated. The subgroups are rotated until all 10 subgroups are exploited, resulting in 10 sets of accuracy values. The corresponding accuracy values are averaged for each CP value to get the cross-validation error. The CP value with minimum cross-validation is the most desirable. From the tried CP's, the optimal is 0.030201342, which corresponds to a cross-validation accuracy equal to 0.8440370. A graph of complexity parameters by cross-validation is shown in Fig. 8.

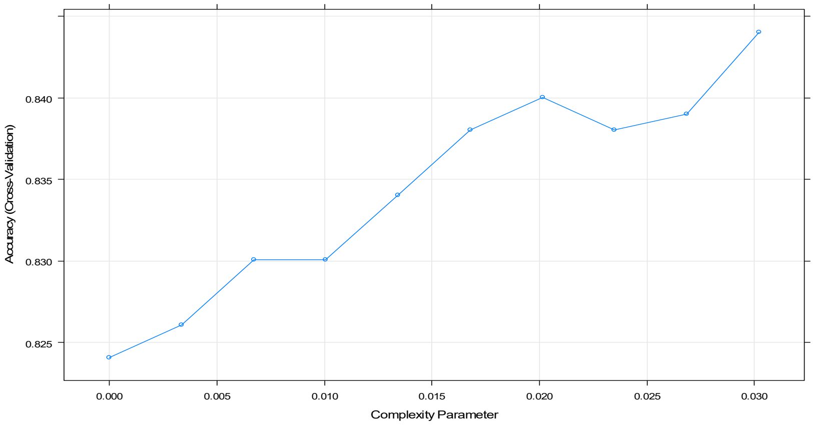


Fig. 8. A graph of complexity parameters by cross-validation

However, this pertains only to the root node, indicating that the baseline model is the best tool. The baseline model is a model which doesn't consider the predictor variables. The model makes predictions by classifying each incoming observation as a no attrition. This is why we expect 85% accuracy. The 84% cross-validation accuracy is within this neighborhood. Based on the data shown in Table 12, the model's specificity is 100% when predicting the testing set, but the sensitivity is 0%. Therefore, we traded all the sensitivity for 100% specificity. The overall accuracy falls to 81.3%, which indicates a loss of 2.1% in overall accuracy in exchange for a simpler model.

Table 12
Confusion matrix for the pruned model

		Actual		
		No	Yes	Total
	No	382	88	470
Predicted	Yes	0	0	0
	Total	382	88	470

4.2. Improving the accuracy of the Random Forest model

By optimizing the model mtry parameter using 10-fold cross-validation, the model with mtry =4 and ntree =101 has the highest performance, and the cross-validation error is 13.5%. It is found that the model with mtry =4 was iterated over 1:34 while ntree was iterated over 101:1004. The heatmap in Figure 9 shows the cross-validation error for each combination of try and tree. On the bottom left, the darkest tile represents the lowest cross-validation error for this case (mtry=4,ntree=101).

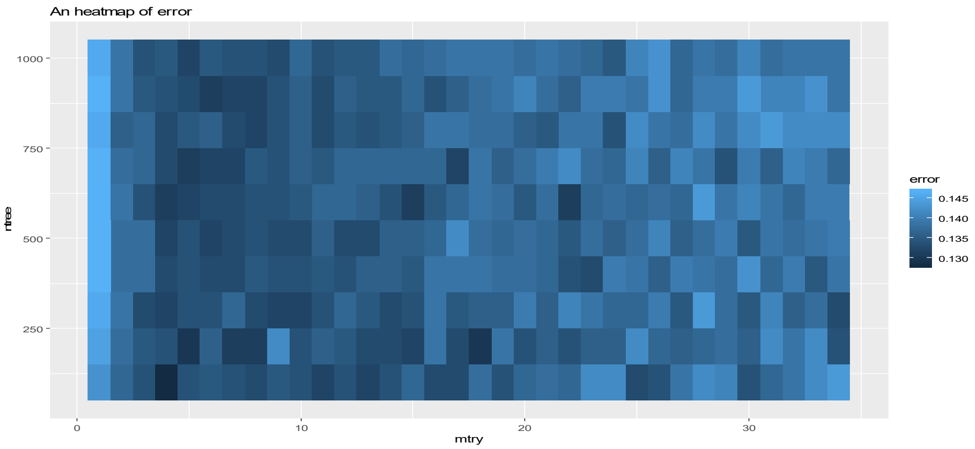


Fig. 9. A heatmap of error

The tuned model increased specificity to 99.74%, and sensitivity increased to 15.91%. With the tuning, as the model sensitivity, the model specificity improves too. The overall accuracy also increased to 84.04%. Table 13 summarizes the results concluded by the Random Forest model.

Table 13
Confusion matrix of the best of the tuned model

		Actual		
		No	Yes	Total
Predicted	No	381	74	455
	Yes	1	14	15
	Total	382	88	470

4.3. Improving the accuracy of the Binary Logistic Regression model

The model with 32 predictors happens to be complex. In this part, an attempt to improve both the model's accuracy and complexity was conducted using the stepwise selection technique.

According to Desboulets in (Desboulets, 2018), stepwise regression performs a search in logistic regression for the best model by starting with an empty model and then adding the next model, improving the model's fit. More candidate variables are then added until the model fit can't improve significantly anymore (forward selection). The backward selection method also starts with the full model, removes the variable, and then removes the variables one by one until the best fit is achieved. The model fit is measured by the Akaike information criterion (AIC), whose formula is as follows:

$$AIC = 2K - 2\ln(\widehat{L}) \qquad (4)$$

It can also be expressed as 2k-deviance, where k is the number of the estimated parameters in the model. The best model is selected according to the value of AIC, in which the selected model is associated with the lowest value of the AIC. In this analysis, both forward and backward selection methods are used. Stepwise regression, therefore, penalizes models' excessive parameters and poor fit. Therefore, the following 11 predictors in Table 14 were removed.

Table 14
Removed variables

index	Variable
1	Age
2	Daily Rate
3	Education
4	Hourly Rate
5	Marital Status
6	Monthly Income
7	Monthly Rate
8	Percent Salary Hike
9	Performance Rating
10	Years at Company
11	Years with Currency Manager

Comparing the stepwise and complete models, it is observed that there is no significant difference between the two models, as shown in Table 15. However, the 11 is a significant improvement in model complexity.

Table 15
Comparing between Stepwise and the entire model

	Df	Deviance
Stepwise model	954	472.65
Full model	939	460.99
Diff	60	11.658
Sig		0.7047

By predicting employee attrition using the test dataset, 97.6% (373 of 382 no attritions) were correctly predicted, and 43.18% (38 of 88 attritions) were correctly predicted. Therefore, the model has a higher specificity but a low sensitivity. Nevertheless, the correct overall prediction is 87.44% which is a slight improvement from the original model, as shown in Table 16.

Table 16
Model accuracy

		Actual		
		No	Yes	Total
	No	373	50	423
Predicted	Yes	9	38	47
	Total	382	88	470

Comparing the ROC curves, the stepwise model has AUC =0.9867, which is higher than a full model, as shown in Fig. 10.

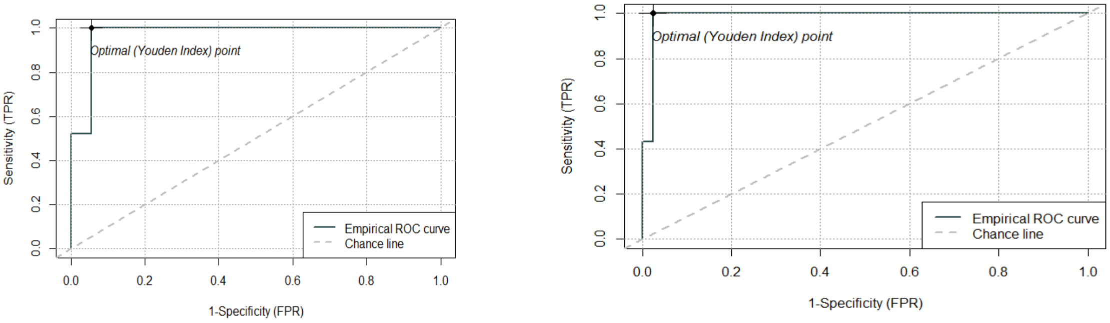
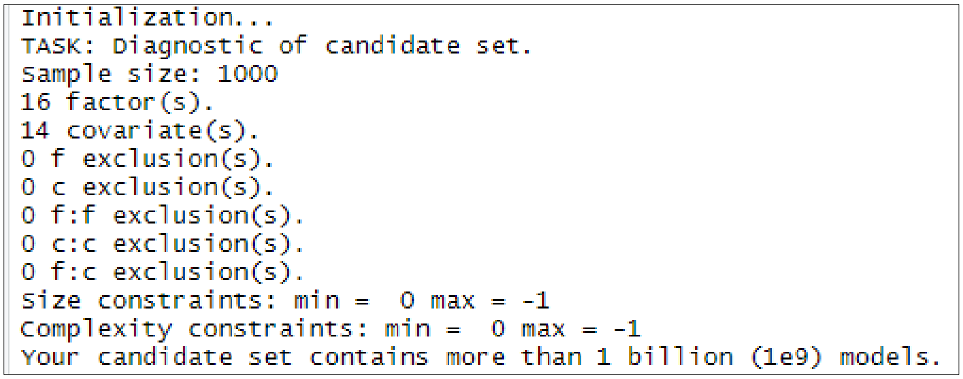


Fig. 10. ROC curve

The best subsets Logistic regression model considers all the possible subsets of the independent variables. The model with the least AIC is therefore considered the best model. With over 1 billion possible subsets, as shown in Fig. 11, both the genetic and the exhaustive method will rarely work properly with this number of models and a low-level machine.



```
Initialization...
TASK: Diagnostic of candidate set.
Sample size: 1000
16 factor(s).
14 covariate(s).
0 f exclusion(s).
0 c exclusion(s).
0 f:f exclusion(s).
0 c:c exclusion(s).
0 f:c exclusion(s).
Size constraints: min =  0 max = -1
Complexity constraints: min =  0 max = -1
Your candidate set contains more than 1 billion (1e9) models.
```

Fig. 11. Selection of the best Logistic

Comparing the accuracies of the 2 logistic regression models, the stepwise regression model was found to be the best with the highest prediction accuracy (87.44%). In interpreting the model coefficients, it is worth noticing that there are 3 independent variables: continuous, ordinal, and nominal. R software handles each type differently. The continuous predictor variables, including distance from home, number of companies worked for, years since last promotion, and years in the current role, were significant predictors of attrition at a 5% significance. Distance from home, number of companies worked for and years since last promotion have a positive coefficient. This means that as the variables increase, the log of odds increases with these variables. However, the variables also have an odds ratio greater than 1, indicating that they increase the chance of attrition (they are protective risk factors). For instance, the odds ratio for distance from home is 1.063, which means that the odds of attrition increase by 0.063 for each unit increase in distance from home. Conversely, years in the current role have a negative coefficient, which means that the log of odds ratio decreases as the variable increases (it is a protective factor). The odds ratio for the variable is 0.8895, which means that the odds of attrition decrease by 0.79991 times for every extra year in the current role. For nominal variables, we have no order of categories. R software, therefore, compares the log of odds ratio across the categories of nominal variables, using one category as the reference point. Variables of job role and overtime are significant predictors at a 5% significance level. For job roles, healthcare representatives were made the reference category. Sales executives have a significant positive coefficient, indicating a higher log of odds ratio compared to healthcare representatives. Consequently, this category experiences a higher rate of employee attrition than healthcare representatives.

For ordinal variables, R software recognizes that in addition to the categorical variables, the categories have some order. The default method is to fit a series of orthogonal functions to the levels of the variable. There are k-1 contrasts and n number of categories. The first part is the linear contrast, the second part is quadratic, and the third part is cubic. The trend continues until the last contrast, which is to power k-1. The generalization of interpretation aims to explain which contrasts significantly account for any difference between our levels. For business travels, the linear contrasts coefficient is significant and positive (B=1.554,37p =0.0001). This suggests a positive linear trend in log odds across business travel categories. Log odds increase as we go up the categories (the odds of attrition are high as business travels increase). It also found that the quadratic coefficient is insignificant. Environment satisfaction, job involvement, relationship satisfaction, and work-life balance have significant linear contrast coefficients. This signifies that log odds decline as the levels of these factors increase (people satisfied with the job, job environment, etc., will rarely leave the company*)*. Table 17 represents the rest of the regression coefficient and the value of interests.

Table 17
Regression coefficients

		Estimate	Std. Error	z value	EX(p)	Pr(>|z|)
(Intercept)		-1.544	1.354	-1.14	0.213489	0.2541
Business Travel						
	BusinessTravel.L	1.554	0.391	3.978	4.73253	0.0001***
	BusinessTravel.Q	0.091	0.250	0.364	1.095094	0.7158
Distance from Home		0.061	0.015	4.057	1.062644	<0.0001***
EducationField						
	EducationFieldLife Sciences	-1.768	1.026	-1.724	0.170618	0.0847.
	EducationFieldMarketing	-1.718	1.093	-1.572	0.179398	0.1158
	EducationFieldMedical	-1.926	1.034	-1.864	0.145699	0.0624.
	EducationFieldOther	-1.926	1.122	-1.716	0.145753	0.0861.
	EducationFieldTechnical Degree	-0.308	1.053	-0.292	0.735195	0.7702
	EnvironmentSatisfaction.L	-1.217	0.263	-4.625	0.296182	<0.0001***
	EnvironmentSatisfaction.Q	0.345	0.249	1.389	1.412075	0.165
	EnvironmentSatisfaction.C	-0.318	0.254	-1.249	0.727792	0.2118
Gender						
	Male	0.412	0.255	1.613	1.50911	0.1067
Job Involvement	JobInvolvement.L	-1.747	0.425	-4.115	0.174312	<0.0001***
	JobInvolvement.Q	0.242	0.338	0.715	1.273934	0.4743
	JobInvolvement.C	-0.110	0.235	-0.469	0.895521	0.639
Job Level						
	JobLevel.L	0.484	1.249	0.387	1.621789	0.6987
	JobLevel.Q	1.163	0.725	1.602	3.198046	0.1091
	JobLevel.C	-0.791	0.620	-1.275	0.453523	0.2024
	JobLevel 4	0.691	0.489	1.413	1.996149	0.1577
Job Role						
	Job Role Human Resources	0.849	1.076	0.789	2.336187	0.4303
	Job Role Laboratory Technician	0.987	0.853	1.158	2.683736	0.2471
	Job Role Manager	-0.428	1.158	-0.37	0.651811	0.7117
	Job Role Manufacturing Director	0.584	0.791	0.739	1.793448	0.46
	Job Role Research Director	-2.425	1.414	-1.716	0.088445	0.0862.
	Job Role Research Scientist	-0.242	0.878	-0.276	0.784789	0.7825
	Job Role Sales Executive	1.865	0.664	2.81	6.457421	0.005**
	Job Role Sales Representative	1.477	0.944	1.564	4.381846	0.1177
	JobSatisfaction.L	-0.914	0.249	-3.673	0.401034	0.0002***
	JobSatisfaction.Q	-0.109	0.251	-0.433	0.896973	0.6648
	JobSatisfaction.C	-0.358	0.259	-1.378	0.699388	0.1681
Num Companies Worked		0.206	0.055	3.778	1.228815	0.0002***
Overtime						
	Yes	2.531	0.286	8.849	12.56644	< 2e-16
Relationship Satisfaction						
	RelationshipSatisfaction.L	-0.832	0.256	-3.249	0.435339	0.0012**
	RelationshipSatisfaction.Q	0.340	0.257	1.325	1.405355	0.1851
	RelationshipSatisfaction.C	-0.374	0.260	-1.44	0.687647	0.1499
Stock Option Level						
	StockOptionLevel.L	-0.565	0.354	-1.597	0.568343	0.1103
	StockOptionLevel.Q	1.285	0.358	3.595	3.61644	0.0003***
	StockOptionLevel.C	-0.590	0.366	-1.61	0.554444	0.1073
Total Working Years		-0.106	0.037	-2.852	0.899766	0.0043**
Training Times Last Year		-0.144	0.096	-1.498	0.865585	0.1341
Work Life Balance						
	WorkLifeBalance.L	-0.923	0.429	-2.151	0.397238	0.0315*
	WorkLifeBalance.Q	0.481	0.342	1.403	1.617093	0.1605
	WorkLifeBalance.C	0.171	0.236	0.722	1.186253	0.47
Years in Current Role		-0.223	0.060	-3.727	0.799915	0.0002***
Years Since Last Promotion		0.289	0.060	4.79	1.334825	<0.0001***

5. Conclusion

The research findings indicate that different machine learning algorithms were applied to predict employee attrition. The Decision Tree model, after attempts to improve its accuracy, achieved a score of 81.3%, which represented a decrease of 2.5% compared to the initial results. The decrease in accuracy was observed due to the simplification of the model. Conversely, the Random Forest model, after parameter tuning, demonstrated an accuracy of 84.04%, showcasing a 0.8% increase. Additionally, the Logistic Regression model's accuracy improved to 87.44% through the use of stepwise regression, resulting in a 1% increase. Notably, the Logistic Regression model exhibited higher accuracy in training set prediction compared to the other models. The analysis revealed several significant findings. The positive linear contrast coefficients for business travel indicated that increased business travel heightened the likelihood of employees leaving the company. This finding was reasonable, as business travel exposes individuals to other companies and potential job opportunities, making them susceptible to recruitment efforts. Furthermore, the significance of distance indicated that employees preferred working close to home, leading to increased attrition rates with greater distance. To mitigate attrition, companies can prioritize local hiring or encourage employees to settle nearby when hiring from a distance. Notably, factors such as job satisfaction, job involvement, relationship satisfaction, and work-life balance were found to have a negative relationship with attrition rates. Therefore, companies should strive to ensure employee satisfaction, encourage positive relationships among staff, and maintain a healthy work-life balance to minimize attrition rates. The analysis identified the sales executive role as the most affected by attrition rates. This could be attributed to their frequent interactions with the public, which exposes them to potential job openings in other companies. Additionally, employees with a history of changing employers were more likely to leave, indicating a propensity for job hopping. Surprisingly, overtime was found to be insignificant in influencing attrition, suggesting that employees do not leave solely based on overtime pay. Performance rating was also found to have no impact on attrition.

To effectively leverage machine learning algorithms for analyzing and predicting employee attrition, organizations are advised to follow certain steps. These include identifying relevant data sources, selecting appropriate machine learning algorithms, choosing relevant features and variables, and continuously monitoring and updating the models. It is also crucial to understand the capabilities and limitations of these predictive models and consider potential risks and limitations, including bias, privacy, interpretability, and validation. Future research can explore the use of different machine learning algorithms to predict employee attrition, considering additional factors and features that may impact attrition. Improving data quality through enhanced data cleansing techniques can lead to more accurate predictions. Addressing bias and fairness concerns during model training and evaluation can help develop more equitable and fair models. Furthermore, it may be valuable to predict employee performance and optimize workforce planning by analyzing factors such as job satisfaction, work engagement, and training history. Examining retirement rates and demographic trends can also provide insights into future workforce challenges. However, ethical considerations, such as employee data privacy, must be carefully addressed when designing and implementing these models.

References

Al-Darraji, S., Honi, D. G., Fallucchi, F., Abdulsada, A. I., Giuliano, R., & Abdulmalik, H. A. (2021). Employee attrition prediction using deep neural networks. *Computers*, *10*(11), 1–11. https://doi.org/10.3390/computers10110141

Alao, D. A. B. A., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. *Information Systems & Development Informatics*, *4*(1), 17–28.

Almomani, A., Alauthman, M., Shatnawi, M. T., Alweshah, M., Alrosan, A., Alomoush, W., & Gupta, B. B. (2022). Phishing Website Detection With Semantic Features Based on Machine Learning Classifiers: A Comparative Study. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–24. https://doi.org/10.4018/IJSWIS.297032

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 2019*. https://doi.org/10.1109/I2CT45611.2019.9033784

Branham, L. (2005). Planning to become an employer of choice. *Journal of Organizational Excellence*, *24*(3), 57–68. https://doi.org/10.1002/joe.20060

Breiman, L. (1996). Out-of-Bag Estimation. In *Statistics Department: University of California Berkeley*.

Ceriani, L., & Verme, P. (2012). The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality*, *10*(3), 421–443. https://doi.org/10.1007/s10888-011-9188-x

Dalton, D. R., & Mesch, D. J. (1990). The Impact of Flexible Scheduling on Employee Attendance and Turnover. *Administrative Science Quarterly*, 370–387. Retrieved from http://www.jstor.org/stable/pdf/3150242.pdf?_=1467266017307

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*. https://doi.org/10.3390/econometrics6040045

Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4). https://doi.org/10.3390/computers9040086

Gaurav, A., Gupta, B. B., & Panigrahi, P. K. (2023). A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information system. *Enterprise Information Systems*, *17*(3).

https://doi.org/10.1080/17517575.2021.2023764

Here's what your turnover and retention rates should look like. (n.d.). Retrieved October 14, 2022, from https://www.ceridian.com/blog/turnover- and-retention-rates-benchmark (accessed

Joseph, R., Udupa, S., Jangale, S., Kotkar, K., & Pawar, P. (2021). Employee attrition using machine learning and depression analysis. *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*. https://doi.org/10.1109/ICICCS51141.2021.9432259

Lazzari, M., Alvarez, J. M., & Ruggieri, S. (2022). Predicting and explaining employee turnover intention. *International Journal of Data Science and Analytics*, *14*(3), 279–292. https://doi.org/10.1007/s41060-022-00329-w

Liu, J. L. (2014). *Main causes of voluntary employee turnover: A study of factors and their relationship with expectations and preferences*. University Of Chile.

Nagadevara, V., & Srinivasan, V. (2007). Early Prediction of Employee Attrition in Software Companies-Application of Data Mining Techniques. *The 10th International Conference of the Society of Global Business and Economic Development*.

Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Zolfani, S. H. (2021). An improved machine learning-based employees attrition prediction framework with emphasis on feature selection. *Mathematics*, *9*(11). https://doi.org/10.3390/math9111226

Ponnuru, S., Merugumala, G., Padigala, S., Vanga, R., & Kantapalli, B. (2020). Employee Attrition Prediction using Logistic Regression. *International Journal for Research in Applied Science and Engineering Technology*, *8*(5), 2871–2875. https://doi.org/10.22214/ijraset.2020.5481

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. *International Journal of Machine Learning and Computing*, *11*(2), 110–114. https://doi.org/10.18178/ijmlc.2021.11.2.1022

Rombaut, E., & Guerry, M. A. (2018). Predicting voluntary turnover through human resources database analysis. *Management Research Review*, *41*(1), 96–112. https://doi.org/10.1108/MRR-04-2017-0098

Subhash Pavan. (2017). IBM HR Analytics Employee Attrition & Performance. Retrieved from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Subhashini, M., & Gopinath, R. (2020). Employee Attrition Prediction in Industry Using Machine Learning Techniques. *International Journal of Advanced Research in Engineering and Technology*, *11*(12), 3329–3341. Retrieved from https://doi.org/10.34218/IJARET.11.12.2020.313

Vasa, J., & Masrani, K. (2019). Foreseeing employee attritions using diverse data mining strategies. *International Journal of Recent Technology and Engineering*, *8*(3), 620–626. https://doi.org/10.35940/ijrte.B2406.098319

Wassan, S., Suhail, B., Mubeen, R., Raj, B., Agarwal, U., Khatri, E., … Dhiman, G. (2022). Gradient Boosting for Health IoT Federated Learning. *Sustainability*, *14*(24). https://doi.org/10.3390/su142416842

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. In *Advances in Intelligent Systems and Computing* (Vol. 2). https://doi.org/10.1007/978-3-030-01057-7_56