**International Journal of Information Technology and Management**

ISSN online: 1741-5179 - ISSN print: 1461-4111
https://www.inderscience.com/ijitm

# Applying machine learning algorithms to determine and predict the reasons and models for employee turnover

Shardul Shankar, Ranjana Vyas, Vijayshri Tewari

**DOI:** 10.1504/IJITM.2024.10061685

**Article History:**

| | |
|---|---|
| Received: | 04 March 2019 |
| Last revised: | 03 September 2020 |
| Accepted: | 24 November 2021 |
| Published online: | 22 January 2024 |

# Applying machine learning algorithms to determine and predict the reasons and models for employee turnover

Shardul Shankar*

Department of Management Studies,
Indian Institute of Information Technology,
Allahabad – 211015, India
Email: Shane.shardul@gmail.com
*Corresponding author

Ranjana Vyas

Department of Information Technology,
Indian Institute of Information Technology,
Allahabad – 211015, India
Email: ranjana@iiita.ac.in

Vijayshri Tewari

Department of Management Studies,
Indian Institute of Information Technology,
Allahabad – 211015, India
Email: vijayshri@iiita.ac.in

**Abstract:** In recent years, organisations have struggled with the turnover of employees, which has become one of the biggest issues that not only has inadvertent consequences for an organisation's growth, productivity, and performance but also has negative implications for the intrinsic cost associated with it. To cater to this problem, one such method is the use of machine learning algorithms. But one of the biggest issues in HR information system (HRIS) analysis is the presence of noise in data, leading to inaccurate predictions. This paper tries to examine the efficiency of six such algorithms, to determine the robustness, accuracy in real-time analysis of data, and then use that company's historical data to predict employee turnover for the present year. The dataset was mined from the HRIS database of a global organisation in the USA and Canada in the span of ten years to compare these algorithms to examine voluntary turnover, using Python and RStudio analytical tools.

**Keywords:** employee turnover; machine learning; predictive algorithms; classification; voluntary turnover.

**Reference** to this paper should be made as follows: Shankar, S., Vyas, R. and Tewari, V. (2024) 'Applying machine learning algorithms to determine and predict the reasons and models for employee turnover', *Int. J. Information Technology and Management*, Vol. 23, No. 1, pp.48–63.

**Biographical notes:** Shardul Shankar is a research scholar at the Indian Institute of Information Technology, Allahabad. He is a scholar at IIIT Allahabad. He received his Bachelors of Technology (Electronics and Communication Engineering) from Bhagwant University, Ajmer; PGDM (Marketing and HR) from School of Management Sciences, Varanasi; and MBA from IIIT Allahabad. His areas of interest include emotional intelligence, leadership, human resources, cognitive behaviour, computations, and analytics, etc. He has several papers published in national and international journals.

Ranjana Vyas is an Assistant Professor at the Department of Information Technology, Indian Institute of Information Technology, Allahabad, and the Coordinator-NewGen IEDC, IIIT-A. She is also a DAAD Fellow – Univ. of Kaiserslautern (Germany). Her research areas include data mining, data science, and business intelligence. She has published multiple articles in various reputed journals.

Vijayshri Tewari is a Professor at the Department of Management Studies, Indian Institute of Information Technology, Allahabad. Masters in Psychology and HRM, her career has led her to specialise increasingly in the management of complex academic deliverables and projects. She has indulged in the application of modern leadership methods, motivation, and clarity in communication, Strategic HR, and personal inter-relations in her research endeavours. Her own background and interests in quality service delivery have also made her confident in handling students of different backgrounds and calibre and reconciliation of different approaches to otherwise common problems. Her publication includes many national and international papers in reputed journals and has few books to her accomplishments.

## 1 Introduction

Employee turnover has been one of the biggest issues for any organisation and managers for quite a long time now. Even though, there has been a lot of historical research done over an elongated period of time; but no fixed and standardised model has been able to be created, which can be used as a generic model in all the organisations (Glebbeek and Bax, 2004). Employee turnover has always been looked upon as a negative issue for the HR of the organisation, but works of literature have always suggested otherwise. One thing is for certain though, that employee turnover is usually triggered by negative emotions or actions that occur around an individual, and its negative impacts range from disruptions in business and project continuity to workplace morale to long-term growth and productivity (Punnoose and Ajit, 2016).

Employee turnover, in this dynamic environment, has become more and more crucial to organisations and their managers. This is because when employees leave the organisation, they not only leave their spaces vacant, which is needed to be replaced, they also take away with them the knowledge of the organisation, which these organisations spend a lot of time, money, and energy to build upon (Wright, 2004; Gupta and Singh, 2018). And when these people leave, the organisations have to usually incur more cost, time, and energy to build the knowledge of the replaced employee or recruited employees.

Thus, the organisations try to stay a step ahead of their employees in these situations and try to keep themselves impact-ready. One of the most recent ways that the organisations are increasingly using to deal with this problem is by predicting the future turnover and attrition of these employees using various IT/IS tools (Saradhi and Palshikar, 2011; Punnoose and Ajit, 2016; Hung et al., 2006; Larivière and van den Poel, 2005). For that, these organisations have started predicting their employees leaving with the help of numerous Machine Learning Algorithms, so that the HR department of the organisation has a pro-active implementation plan ready for employee retention.

But, previous literature has revealed that with the advent of HRIS and IT/IS, these algorithms have not been able to substantially predict these outcomes due to ineffectiveness in the removal of noise that creeps in the dataset (Wei and Chiu, 2002). Also, the limited understanding of HR and IT/IS domains by the individuals in these two departments, and non-understanding to bring the connectivity of these two domains into one leads to a minimised exploitation of the outcome's benefits. This is one of the reasons that ROI in HRIS has been a difficult thing to measure (Sahu and Gupta, 1999), which generally attenuates the noise and capability of these algorithms.

Keeping these constraints in mind, this paper tries to address the problem of employee turnover in the organisation with the help of certain key machine learning algorithms. The paper is categorised into three subsections; in the first section, we try to determine and elaborate the previous historical literature researches done in the area of employee turnover by developing a theoretical background. In the second section, we try and explore the applications of various machine learning algorithms in terms of their efficiency and accuracy. This would be done by using the data provided by the HRIS of an organisation that has been operational for over 130 years, and we try and classify the turnover problem using the traditional and modern algorithms using supervised techniques. The third section contains the contrasting comparison of the superior efficiency and accuracy of one machine learning algorithm over the other, and in the process, we try to explain the reason for better performance over the other using appropriate statistical analysis. The conclusion is attained by successfully predicting the current year's expected turnover using the predictive analytics mechanism and predictive model of the most noise-ridden and accurate machine learning algorithm.

## 2 Theoretical background and related literature

Employee turnover is amongst the most concerning and hence the most studied phenomenon in an organisation (Shaw et al., 1998; Archaux et al., 2004; Hung et al., 2006; Rosset et al., 2003; Wei and Chiu, 2002; Morik and Köpcke, 2004; Coussement and van den Poel, 2008; Larivière and van den Poel, 2005). Usually, employee turnover has been defined as the outflow or exodus of employees from an organisation (Stovel and Bontis, 2002). This becomes a cause of concern for an organisation because the loss of employees usually is a loss of intellectual capital of the organisation, as when employees leave, along with them leaves the knowledge of the organisation, which can be general knowledge, technical knowledge, or knowledge specific to the organisation (Wright, 2004).

But the employee turnover is mostly dominated by categorised occasions of the turnover (Morrell et al., 2001b), i.e. involuntary turnover and/or voluntary turnover. In a review of the factors for turnover, research clearly described that both intrinsic and

extrinsic factors are the reasons for employee turnover (Joseph et al., 2007). It divided the turnover intention predictors into personal, demographic, perception, and satisfaction as one of the biggest predictors for the reason for employee turnover (Saradhi and Palshikar, 2011). The research has clearly shown substantial growth in the determinants of employee turnover, and the focus has moved from reasons and has shifted to probable causes for a high turnover (Glebbeek and Bax, 2004).

High turnover has a lot of negative effects on an organisation's performance, as the people leaving not only leads to the loss of intellectual capital but also social capital (Punnoose and Ajit, 2016; Morrell et al., 2001a, 2004). For an organisation, replacing people with a specific and specialised skill-set has always been a problem, but when it is accompanied by a continuous and fast attrition rate, it becomes even more critically important. These places, when are filled, lead to substantial amounts of direct and indirect costs (including recruitment costs to training costs and others). Also, the tasks allotted to these leaving employees are hampered and new people take some time to settle down in a new environment, and thus overall productivity of a project, employees, and thus, the organisation, is reduced.

Due to these reasons, organisations have started to look for various methods to deal with and reduce the turnover rate of their productive employees. One of the more recent and increasingly popular methods that are being continuously and readily used by organisations is machine learning techniques (Punnoose and Ajit, 2016). These Machine learning algorithms are being increasingly used because of many reasons, i.e., they have unbiased data that they analyse and interpret, they are able to efficiently evaluate the various factors affecting these turnover intentions, and more importantly, they are able to pretty accurately predict and forecast the turnover rate, which in turn, helps the organisations to plan accordingly for the future course of action (Saradhi and Palshikar, 2011; Punnoose and Ajit, 2016).

There has been an increase in the study for performance evaluation of machine learning algorithms in recent times. Notably, researchers like Zhao et al. (2019) have used algorithms like neural networks on the reasons for employee turnover. Ameer et al. (2020) and Liu et al. (2020) have compared logistic regression, tree models, and boosting algorithms for predicting turnover.

Some recent researchers have used over ten classifiers and predictors to find which model works the most efficient in predicting the turnover of employees (Huang et al., 2004; Sexton et al., 2005; Liu et al., 2020; McGinty and Lylova, 2020; Dutta and Bandyopadhyay, 2020; Usha and Balaji, 2020; Setiawan et al., 2020) and have tried to find out which of these models are the most influential in predicting turnover, as they help in understanding the features that affect the leaving of an employee (Friedman and Holtom, 2002).

Despite the increased interest in understanding the use of machine learning in predicting employee turnover, there are serious issues in the testing of these models. This can be referred to as multiple attributes, such as excessive confidentiality of HR data (Quinn et al., 2002), and the inconsistency and noise in the available data (Chien and Chen, 2008; Zhao et al., 2019). This leads to questionable accuracy as inconsistent data usually leads to unreliable accuracies, thereby creating misleading outcomes (Sikaroudi et al., 2015; Tzeng et al., 2004; Sexton et al., 2005).

In order to get a comprehensive assessment of these models, reliable feature importance, data visualisation, and correct statistical analysis must be used in order to get

an accurate and reliable analysis of HR data for predicting the turnover of the employees, so that precarious interpretation can be avoided to better gauge the factors influencing the employee turnover.

## 3 Methods

### *3.1 Theoretical background of predictive algorithms*

Organisations are now using predictive models in various business applications (Pathak et al., 2018; Saradhi and Palshikar, 2011; Shankar and Tewari, 2021a) in order to predict a lot of business decisions and applications. In the paper, we try and classify the data with the help of supervised learning (Michie et al., 1994). The problems of supervised learning are constituted of data classification with the help of classes or clusters in the given data training examples. This is done by dividing the labelled objects into training data, but the data to be tested are not used as training data, but are used as test data. With the help of this, a learned function is created which is usually known as the predictive model (Saradhi, 2008). This subsection presents the theoretical understanding of the classification and predictive models used on the data to determine employee turnover using employee history data.

#### *3.1.1 Boosted model (XGboost)*

Boosting in machine learning refers to a process of producing accurate prediction rules with the combination of moderately inaccurate rules-of-thumb (Freund and Schapire, 1995, 1996a, 1966b, 1997; Freund et al., 1999). Weak training sets are superimposed on the modified data, and with the help of weighted sum, predictions are combined and produced. XGBoost is a tree-based method and is also referred to as extreme gradient boosting method (Chen and Guestrin, 2016). This method is used over other boosting techniques because they have been designed to optimise the scalability, reliability, and model performance (Zhao et al., 2019; Punnoose and Ajit, 2016).

#### *3.1.2 Decision trees*

In a given dataset, decision tree constructs a tree-like model where we split the population into multiple sets of homogeneous data with the help of splitters in input variables, and where each node represents an attribute and the branches represent their corresponding values (Duda et al., 2001; Saradhi and Palshikar, 2011). Some of the biggest advantages of a decision tree over other algorithms are that it is easy to create, understand and interpret, and being a non-parametric method, substantially lesser data cleaning is required.

But one of the biggest problem with it is that even the smallest changes in the training data leads to a huge amount of variations in its model and thereby the classifier. But despite their obvious instability, they are still one of the most used predictive models in practice (Hung et al., 2006).

#### *3.1.3 Random forests*

Random forests is a method proposed by Breiman (1996, 2001), which is used to create multiple decision trees on the training data but only using a sampled set of attributes. This

process is commonly known as bagging. In the bagging process, new training data is created using the sampling method on the original dataset. This allows some important sets to be chosen multiple times and some not at all. This process of bagging reduces the variance of classification errors, thereby increasing their accuracy (Chan and Paelinckx, 2008). This, in turn, makes the random forest more robust against the problem of over-fitting.

### *3.1.4 Neural networks*

Neural Networks began for the study of biological analogies (Wasserman, 1989), but have since evolved into statistical algorithms, learning, and pattern recognition. The multiple layers of NN, i.e., input layer, multiple hidden layers, and the output layer have predetermined 'neurons' that define its architecture (Somers, 1999). Modifications of these networks have been shown to outperform gradient algorithms and even real-life problems (Gupta et al., 2000; Sexton et al., 1998; Baum and Haussler, 1989; Zhao et al., 2019; Ameer et al., 2020). The structure of a neural network allows it to model its functions into a universally accurate function, especially when given multiple hidden layers (Murphy, 2012). These models can be further extended into deeper developments, giving us the fundamental concepts of deep learning algorithms. This is why, neural networks are the most heavily researched topics in the area of artificial intelligence and machine learning (Zhao et al., 2019; Murphy, 2012).

### *3.1.5 Logistic regression*

One of the basic techniques of classification assumptions is Logistic Regression. It obtains the posterior possibilities, and then it assumes a model and then predicts the attributes of the model. This method is usually comparatively used against the SVMs, Decision Tree, and Random Forests. It predicts for categorical variables, and the parameters are found to be close to that of random forests' test data (Coussement and van den Poel, 2008). The model is estimated by the maximum likelihood technique (King and Zeng, 2001).

### *3.1.6 K-nearest neighbours*

This model is usually used for classification and regression, as a non-parametric algorithmic statistical test. They are used to identify a number of data points (here: K), and they measure the nearest neighbour of the new data point to classify it into the dataset. This new distance is measured by finding the average of the K nearest neighbours (Friedman and Holtom, 2002; Murphy, 2012). K-Nearest Neighbours have a great distinct advantage of being fairly accurate with smaller datasets, but an increase in feature dimensions drastically reduces the accuracy of the model (Zhao et al., 2019).

## 4 Data

Our data is the origination of a global organisation that has been in operations for over 130 years and has had its offices spread across Canada and the USA. The data is extensive data for a period of 10 years, and in the period of data reference, only those

business units were our units of observation which were active during the whole time of reference. The data window was created on a yearly basis, hence, when the employee left the organisation, the records did not appear in the dataset the following year. Each instance stood as the recorded attribute of the employee of the incremental year. The data was gathered from the HRIS database of the organisation. Thus, the data contains the annual snapshot of all the active employees and the occurrence of terminations.

The dataset contains 49,686 instances, and 17 attributes:

1 employee identification number
2 recording date
3 employee birth date
4 hiring date
5 termination date
6 employee age
7 employee gender
8 employee city location
9 employee designation
10 employee department
11 length of service
12 employee store number
13 business units
14 employment status
15 status year
16 turnover reason
17 turnover type.

The most recent year of the data is taken as the test data, and the remaining nine years of data is taken as the training data for the analysis. The data is decreed from the firm's records but does not guarantee the reliability of the data (Glebbeek and Bax, 2004); as noise is inherent to machine-generated data.

### *4.1 Data preprocessing*

As discussed in the introduction section, after getting the data, the first thing done was preprocessing it in order to remove missing entries and noise in the data. The following methods were adopted in the process of data preprocessing:

### *4.2 Missing value*

In order to get a more meaningful result, missing values were dealt with, even though some of the algorithms are well-versed in dealing with missing values (such as XGBoost). Whenever a missing value was encountered, the row of the data was dropped.

### *4.3 Data type conversion*

Since some of the algorithms used in this study have trouble understanding the categorical data, they were thus converted into numerical data. This is usually done by one-hot encoding (Quinn et al., 2002; Punnoose and Ajit, 2016), but here it was done by label encoding by using the Scikit-learn Package in Python (Zhao et al., 2019; Pedregosa et al., 2011). The traditional method increases the feature dimensions but at the caveat of multiple distinct values. The method used in the study is a feature selection method as it drastically increases the prediction accuracy of the classification algorithms, and automatically uses dimensionality reduction methods if needed.

### *4.4 Feature scaling*

The feature scaling method allows the machine learning algorithm to reduce feature ranges by adjusting the feature scales (Zhao et al., 2019) because scale gaps reduce the performances of these classifiers at the optimisation stage. In this study, the logarithmic feature scaling method was done because it adjusts the distribution, and consequently, the linear performance of the feature is substantially increased. Both standardisation and normalisation were done on the given data and were plotted on the feature importance diagram.

### *4.5 Experimental methodology*

The data given to us has various columns. For data analysis, the preprocessing of the data and feature selection are important things. There were some features that were useless in terms of context, and some were useful. We can use various techniques to choose the important features. The various libraries like Scikit library were used for importing the features.

When we feed the data to our Machine Learning model, we have to prepare the data which includes the categorical classification. One more important point that has to be kept in mind is to choose the right evaluation metrics.

Here are the different types of metrics that have been used in this study:

1 precision

2 recall

3 F1 score

For decision tree, linear regression, and support vector machine the following procedure was adopted:

1 the data was encoded using the python package named as LabelEncoder

2 then the data was split into training and testing data in a ratio of 60:40

3 the features for the data were 8 and the label for the data was STATUS

4 then using the package present for the decision tree, linear regression, and support vector machine, the accuracy, feature selection, cross-validation, and the F1 score of the data were found.

For neural network the following procedure was adopted:

1 The model was built using the Sequential model technique which is provided by the Keras.

2 Now we added the layers to the model. For each layer, we specified the number of neurons and also the activation function for them. It should be kept in mind that only the first layer expects the input dimension. And the numbers of units in the final layers are the target variables.

3 While compiling the model, there are various things that have to be kept in mind. We need to specify the loss function. We have used a binary cross-entropy function for the loss function of the neural network.

4 Then we have to evaluate the model performance. Here, we used accuracy, feature selection, and cross-validation.

5 Lastly we need to fit our model onto the training data just as we do for traditional Machine Learning algorithms.

## 5 Statistical analysis

Here are the metrics that have been used to conduct the comparison of the model's performance:

- precision = (true positives)/(true positives + false positives)
- Recall = (true positives)/(true positives + false negatives)
- F1 score = the harmonic mean of 'precision' and 'recall'.

The precision allows you to estimate the true positives out of all the positives. This means that out of all the predicted positives, how many are truly positive. The recall actually measures the true positives out of the total labelled actual positives.

The F1 score test is done to conduct the weighted average through the harmonic mean of precision and recall. This allows for balancing the accuracy of both factors. It is especially useful when we have a large number of True Negatives which are usually excluded by business organisations as they have a very low effect on the tangible costs, but a very high effect on the non-tangible costs.

## 6 Results

The experimental analysis of the data and its results are given in this section. The first step in this model development was data preprocessing and reaching the desired predictive class labels. In data cleaning, the rows of missing values were dropped. The categorical data were label encoded using SkiLit Package by Python.

Then the data was split 70:30 into training and test data. After that, the present attributes were used to create the predictive models (Shankar and Tewari, 2021b). We found that not all attributes would be helping us in model development, so attributes were

pruned, and then each algorithm was used to predict on the 30% of the test data sample. All the predictive models are run on RStudio and Python.

Table 1 presents the summary of the dataset.

**Table 1** Dataset summary

| *Status year* | *ACTIVE* | *TERMINATED* | *TOTAL* | *Percent terminated* |
|---|---|---|---|---|
| 2016 | 4,799 | 162 | 4,961 | 3.27 |
| 2015 | 4,962 | 253 | 5,215 | 4.85 |
| 2014 | 5,215 | 105 | 5,320 | 1.97 |
| 2013 | 5,101 | 130 | 5,231 | 2.49 |
| 2012 | 4,972 | 110 | 5,082 | 2.17 |
| 2011 | 4,840 | 123 | 4,963 | 2.48 |
| 2010 | 4,710 | 142 | 4,852 | 2.93 |
| 2009 | 4,603 | 164 | 4,767 | 3.44 |
| 2008 | 4,521 | 162 | 4,683 | 3.46 |
| 2007 | 4,445 | 134 | 4,579 | 2.93 |

## 7 Performance analysis of the algorithms

After the initial data cleaning and preprocessing, the data are trained on the various models with the help of training data and then evaluated using the test data. The termination according to various constraints are checked (i.e., by employee age, length of service, city name, department, job title, business units, store name, status year, termination type and termination reason) and was measured on employment status. Eleven attributes are ignored and the models are tested on the remaining eight subset attributes.

After that, the data are evaluated with the help of accuracy metrics. This procedure is used because it allows for more accurate model validation as it measures the probability of classifier rank being randomly chosen. The model results are shown in Table 2.

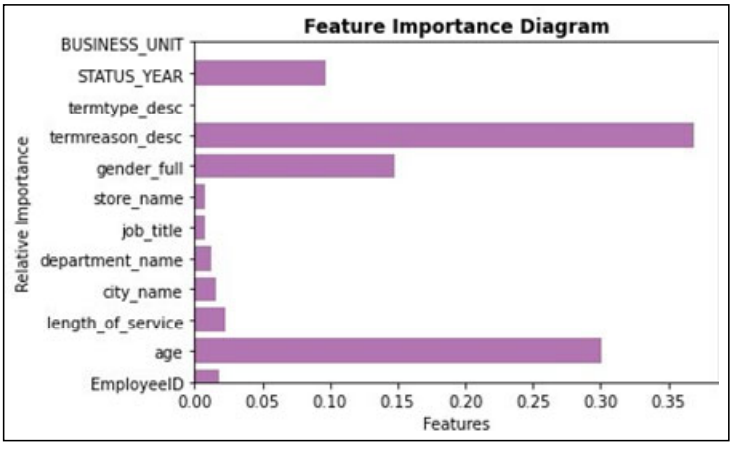
**Table 2** Accuracy metrics of the algorithms

| *Models* | *Accuracy* |
|---|---|
| Neural networks | 0.994 |
| Logistic regression | 0.977 |
| XGBoost | 0.995 |
| Decision tree | 0.991 |
| Random forest | 0.995 |
| K-nearest neighbour | 0.980 |

With the help of these employees' turnover prediction model, now, the employees who have resigned/released can certainly be calculated. It can be easily seen that the XGboost model was able to present the highest Accuracy, and hence could be used to predict the attrition and turnover most accurately. They were closely followed by random forests and neural networks with a 1,000th fractional difference, which is very low for a data this big.

The next was the feature scaling process, which led to identifying and presenting the categorisation of the feature importance. After selecting the features, the features were ranked according to their feature importance. All the models were able to produce their own feature rankings, but since XGBoost had the best accuracy of them all, it was fitting to use this model's feature selection with the help of the feature importance diagram. Figure 1 presents the feature importance diagram using the XGBoost model.

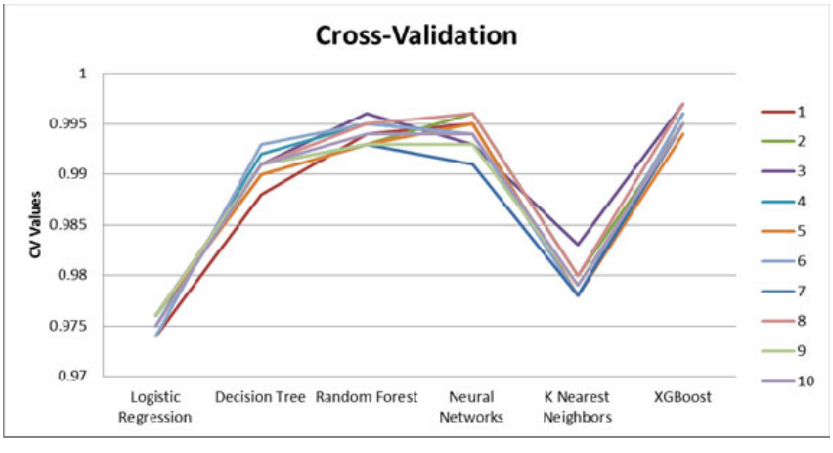
**Figure 1** Feature importance diagram using XGBoost model (see online version for colours)




It can be clearly seen in Figure 1 that the feature that most affects our models are the reason for termination, which includes layoffs, resignations, and retirements; followed by age of the employee. These allowed us to look into the critical factors that can act as features for the study.

This was followed by measuring the performance of the data. This was done in three major steps, the first was cross-validating the dataset, then creating the confusion matrix of the models, and then doing the statistical analysis on the precision and recall values of the models. The cross-validation that was used was simple cross-validation and k-fold cross-validation for neural networks. The cross-validation table is given in Figure 2.

**Figure 2** Cross-validation for the predictive algorithms (see online version for colours)



In order to further validate the models, the performance of the test data is checked by confusion matrix. It creates four combinations of predicted and actual values. The four combinations are true positive, true negative, false negative, and false positives. They allow us to measure the precision, recall, sensitivity, and accuracy of the model. The table of confusion matrix is given in Tables 3–7.

**Table 3** Confusion matrix of k-nearest neighbours

| *n = 49,653* | *Predicted: NO* | *Predicted: YES* |
|---|---|---|
| Actual: NO | 12,021 | 30 |
| Actual: YES | 216 | 147 |

**Table 4** Confusion matrix of decision tree

| *n = 49,653* | *Predicted: NO* | *Predicted: YES* |
|---|---|---|
| Actual: NO | 12,027 | 24 |
| Actual: YES | 85 | 278 |

**Table 5** Confusion matrix of XGBoost

| *n = 49,653* | *Predicted: NO* | *Predicted: YES* |
|---|---|---|
| Actual: NO | 12,037 | 14 |
| Actual: YES | 43 | 320 |

**Table 6** Confusion matrix of random forests

| *n = 49,653* | *Predicted: NO* | *Predicted: YES* |
|---|---|---|
| Actual: NO | 12,039 | 12 |
| Actual: YES | 50 | 313 |

**Table 7** Confusion matrix of logistic regression

| *n = 49,653* | *Predicted: NO* | *Predicted: YES* |
|---|---|---|
| Actual: NO | 12,049 | 2 |
| Actual: YES | 287 | 76 |

## 8 Statistical analysis

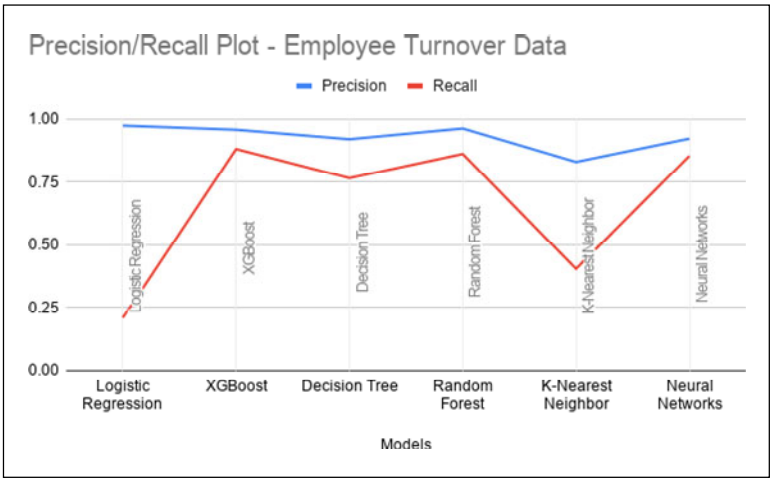
After the creation of the confusion matrix, the F1 score test was done with the classifiers to understand the significant differences with the algorithms. This could've been done by first finding out the precision and recall values from the confusion matrix. They allowed for a better statistical understanding of the performances of the models of interest by giving us the balance in the accuracy of true and false positives. The table of this analysis is given in Table 8.

**Table 8** Classifier performance and scaling effects on the predictive models

| *Models* | *Precision* | *Recall* | *F1 score* |
|---|---|---|---|
| Logistic regression | 0.974 | 0.209 | 0.345 |
| XGBoost | 0.958 | 0.882 | 0.918 |
| Decision tree | 0.921 | 0.766 | 0.836 |
| Random forest | 0.963 | 0.862 | 0.910 |
| K-nearest neighbour | 0.831 | 0.405 | 0.544 |
| Neural networks | 0.923 | 0.855 | 0.887 |

To better understand the scaling effects of the F1 score test, the results were put on the scatter plot; i.e., precision/recall plot. These provide the change and the consistency of the predictive models against the test sample. Figure 3 shows the precision/recall plot.

**Figure 3** Precision/recall plot of the dataset (see online version for colours)




Here, we can clearly see that XGBoost provides the best prediction to the model, which can be corroborated with Table 8 too. With the help of this model, it was easily predicted that 186 out of 49,686 employees would leave in 2017.

## 9 Conclusions and future work

Employee turnover has been one of the biggest problems in an organisation as it not only affects the values but also the intellectual and social capital of the organisation. Thus, for a business analyst, it becomes a matter of supreme importance that an accurate, consistent, and reliable predictive model is built for future action plans. Data was taken from the HRIS of a global organisation from Canada. This paper was created as an outcome of two reasons. First, the employee turnover problem was evaluated with the help of machine learning algorithms, and they were compared with each other by the statistical and mining models and with this, an appropriate employee turnover model was built. Second, from the results of the first models, the most accurate and reliable model

was selected and prediction for the present year was done by identifying and validating the precision and sensitivity of these models.

The important outcome of this paper is that these supervised learning models are a better set of algorithms in terms of high accuracy, low run-time, and high precision and sensitivity; and they can be used to accurately and reliably predict employee turnover over their counterpart models. But it can be seen that there is a substantial gap in these prediction models. In the future, the models can be used to not only determine the quantitative but also qualitative elements, which would not only discuss the present visible value, but also the future intrinsic value of each employee. But, the practical applicability of these models would always be a factor of consideration. Additionally, a study of sentiment analysis techniques and deep learning models would be a good way to predict employee turnover.

## References

Ameer, M., Rahul, S.P. and Manne, S. (2020) 'Human resource analytics using power bi visualization tool', *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May, IEEE, pp.1184–1189.

Archaux, C., Martin, A. and Khenchaf, A. (2004) 'An SVM based churn detector in prepaid mobile telephony', *Proceedings. 2004 International Conference on Information and Communication Technologies: From Theory to Applications*, April, pp.459–460, IEEE.

Baum, E.B. and Haussler, D. (1989) 'What size net gives valid generalization?', *Advances in Neural Information Processing Systems*, Vol. 1, No. 1, pp.81–90.

Breiman, L. (1996) 'Bagging predictors', *Machine Learning*, Vol. 24, No. 2, pp.123–140.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Chan, J.C.W. and Paelinckx, D. (2008) 'Evaluation of random forest and AdaBoost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery', *Remote Sensing of Environment*, Vol. 112, No. 6, pp.2999–3011.

Chen, T. and Guestrin, C. (2016) 'Xgboost: A scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August, pp.785–794.

Chien, C.F. and Chen, L.F. (2008) 'Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry', *Expert Systems with Applications*, Vol. 34, No. 1, pp.280–290.

Coussement, K. and van den Poel, D. (2008) 'Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques', *Expert Systems with Applications*, Vol. 34, No. 1, pp.313–327.

Duda, H., Hart, P.E. and David, G. (2001) *Stork, Pattern Classification*, pp.20–25, John Wiley and Sons Inc.

Dutta, S. and Bandyopadhyay, S.K. (2020) 'Employee attrition prediction using neural network cross validation method', *International Journal of Commerce and Management Research*, Vol. 6, No. 3, pp.80–85.

Freund, Y. and Schapire, R.E. (1995) 'A decision-theoretic generalization of on-line learning and an application to boosting', in Vitányi, P. (Ed.): *Computational Learning Theory. EuroCOLT 1995. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, Vol. 904.

Freund, Y. and Schapire, R.E. (1996a) 'Experiments with a new boosting algorithm', *ICML*, July, Vol. 96, pp.148–156.

Freund, Y. and Schapire, R.E. (1996b) 'Game theory, on-line prediction and boosting', *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, January, pp.325–332, ACM.

Freund, Y. and Schapire, R.E. (1997) 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp.119–139.

Freund, Y., Schapire, R. and Abe, N. (1999) 'A short introduction to boosting', *Journal-Japanese Society For Artificial Intelligence*, Vol. 14, Nos. 771–780, p.1612.

Friedman, R.A. and Holtom, B. (2002) 'The effects of network groups on minority employee turnover intentions', *Human Resource Management*, Vol. 41, No. 4, pp.405–421, Published in cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management.

Glebbeek, A.C. and Bax, E.H. (2004) 'Is high employee turnover really harmful? An empirical test using company records', *Academy of Management Journal*, Vol. 47, No. 2, pp.277–286.

Gupta, A. and Singh, V. (2018) 'Enhancing intention to stay among software professionals', *Academia Revista Latinoamericana de Administración*, Vol. 31, No. 3,

Gupta, J.N., Sexton, R.S. and Tunc, E.A. (2000) 'Selecting scheduling heuristics using neural networks', *INFORMS Journal on Computing*, Vol. 12, No. 2, pp.150–162.

Huang, L.C., Huang, K.S., Huang, H.P. and Jaw, B.S. (2004) 'Applying fuzzy neural network in human resource selection system', *IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS'04*, June, Vol. 1, pp.169–174, IEEE.

Hung, S.Y., Yen, D.C. and Wang, H.Y. (2006) 'Applying data mining to telecom churn management', *Expert Systems with Applications*, Vol. 31, No. 3, pp.515–524.

Joseph, D., Ng, K.Y., Koh, C. and Ang, S. (2007) 'Turnover of information technology professionals: a narrative review, meta-analytic structural equation modeling, and model development', *MIS Quarterly*, Vol. 31, No. 3, pp.547–577.

King, G. and Zeng, L. (2001) 'Logistic regression in rare events data', *Political Analysis*, Vol. 9, No. 2, pp.137–163.

Larivière, B. and van den Poel, D. (2005) 'Predicting customer retention and profitability by using random forests and regression forests techniques', *Expert systems with Applications*, Vol. 29, No. 2, pp.472–484.

Liu, L., Akkineni, S., Story, P. and Davis, C. (2020) 'Using HR analytics to support managerial decisions: a case study', *Proceedings of the 2020 ACM Southeast Conference*, April, pp.168–175.

McGinty, N.A. and Lylova, E.V. (2020) 'Transformation of the HR management in modern organizations', *1st International Conference on Emerging Trends and Challenges in the Management Theory and Practice (ETCMTP 2019)*, February, Atlantis Press, pp.18–21.

Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994) 'Machine learning, neural and statistical classification', *Ellis Horwood Series in Artificial Intelligence*, New York, N.Y.

Morik, K. and Köpcke, H. (2004) 'Analysing customer churn in insurance data – a case study', *European Conference on Principles of Data Mining and Knowledge Discovery*, September, pp.325–336.

Morrell, K., Loan-Clarke, J. and Wilkinson, A. (2001a) 'Unweaving leaving: the use of models in the management of employee turnover', *International Journal of Management Reviews*, Vol. 3, No. 3, pp.219–244.

Morrell, K., Loan-Clarke, J. and Wilkinson, A. (2001b) *Lee and Mitchell's Unfolding Model of Employee Turnover – A Theoretical Assessment*, Loughborough Univ. Business School, UK.

Morrell, K.M., Loan-Clarke, J. and Wilkinson, A.J. (2004) 'Organisational change and employee turnover', *Personnel Review*, Vol. 33, No. 2, pp.161–173.

Murphy, K.P. (2012) '*Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA.

Pathak, A., Tewari, V. and Shankar, S. (2018) 'Impact of emotional intelligence on employability of IT professionals', *Management Insight*, Vol. 14, No. 1, pp.14–21.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … and Vanderplas, J. (2011) 'Scikit-learn: machine learning in Python', *The Journal of Machine Learning Research*, Vol. 12, No. 10, pp.2825–2830.

Punnoose, R. and Ajit, P. (2016) 'Prediction of employee turnover in organizations using machine learning algorithms', *Algorithms*, Vol. 4, No. 5, p.C5.

Quinn, A., Rycraft, J.R. and Schoech, D. (2002) 'Building a model to predict caseworker and supervisor turnover using a neural network and logistic regression', *Journal of Technology in Human Services*, Vol. 19, No. 4, pp.65–85.

Rosset, S., Neumann, E., Eick, U. and Vatnik, N. (2003) 'Customer lifetime value models for decision support', *Data Mining and Knowledge Discovery*, Vol. 7, No. 3, pp.321–339.

Sahu, A. and Gupta, M. (1999) 'An empirical analysis of employee turnover in a software organization', *Indian Journal of Industrial Relations*, Vol. 35, No. 1, pp.55–73.

Saradhi, V.V. (2008) *Kernel Methods: Some Improvements and Some Analysis*, PhD, Indian Institute of Technology Kanpur.

Saradhi, V.V. and Palshikar, G.K. (2011) 'Employee churn prediction', *Expert Systems with Applications*, Vol. 38, No. 3, pp.1999–2006.

Setiawan, I., Suprihanto, S., Nugraha, A.C. and Hutahaean, J. (2020) 'HR analytics: employee attrition analysis using logistic regression', *IOP Conf. Series: Materials Science and Engineering (830)*, IOP Publishing.

Sexton, R.S., Dorsey, R.E. and Johnson, J.D. (1998) 'Toward global optimization of neural networks: a comparison of the genetic algorithm and backpropagation', *Decision Support Systems*, Vol. 22, No. 2, pp.171–185.

Sexton, R.S., McMurtrey, S., Michalopoulos, J.O. and Smith, A.M. (2005) 'Employee turnover: a neural network solution', *Computers & Operations Research*, Vol. 32, No. 10, pp.2635–2651.

Shankar, S. and Tewari, V. (2021a) 'Impact of collective intelligence and collective emotional intelligence on the psychological safety of the organizations', *Vision*, p.09722629211012256.

Shankar, S. and Tewari, V. (2021b) 'Understanding the emotional intelligence discourse on social media: insights from the analysis of Twitter', *Journal of Intelligence*, Vol. 9, No. 4, p.56.

Shaw, J.D., Delery, J.E., Jenkins Jr., G.D. and Gupta, N. (1998) 'An organization-level analysis of voluntary and involuntary turnover', *Academy of Management Journal*, Vol. 41, No. 5, pp.511–525.

Sikaroudi, A.M.E., Ghousi, R. and Sikaroudi, A. (2015) 'A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)', *Journal of Industrial and Systems Engineering*, Vol. 8, No. 4, pp.106–121.

Somers, M.J. (1999) 'Application of two neural network paradigms to the study of voluntary employee turnover', *Journal of Applied Psychology*, Vol. 84, No. 2, p.177.

Stovel, M. and Bontis, N. (2002) 'Voluntary turnover: knowledge management – friend or foe?', *Journal of Intellectual Capital*, Vol. 3, No. 3, pp.303–322.

Tzeng, H.M., Hsieh, J.G. and Lin, Y.L. (2004) 'Predicting nurses' intention to quit with a support vector machine: a new approach to set up an early warning mechanism in human resource management', *CIN: Computers, Informatics, Nursing*, Vol. 22, No. 4, pp.232–242.

Usha, PM. and Balaji, N.V. (2020) 'An analysis of the use of machine learning for employee attrition prediction – a literature', *Journal of Information and Computational Science*, Vol. 10, No. 3, pp.1429–1438.

Wasserman, P.D. (1989) *Neural Computing: Theory and Practice*, Van Nostrand Reinhold Co., Cupertino, CA.

Wei, C.P. and Chiu, I.T. (2002) 'Turning telecommunications call details to churn prediction: a data mining approach', *Expert Systems with Applications*, Vol. 23, No. 2, pp.103–112.

Wright, D. (2004) *The Law of Project Manager*, 1st ed., p.166, Grower Publishing Ltd. – Business & Economics, USA.

Zhao, Y., Yu, Y., Li, Y., Han, G. and Du, X. (2019) 'Machine learning based privacy-preserving fair data trading in big data market', *Information Sciences*, Vol. 478, pp.449–460.