Predicting Individual Employee Turnover using Machine Learning Techniques

Research Paper Business Analytics

Rozemarijn Veldhuis

Supervised by
Dr. Sandjai Bhulai

Vrije Universiteit Amsterdam
Faculty of Sciences
Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

April 14, 2017

Abstract

Employee turnover is an extensively investigated issue within organizations. As might be guessed, employee turnover has many negative effects. The prediction of employee turnover serves several purposes. Firstly, it identifies individuals that have a high probability of leaving the organization and secondly, it presents an insight in the top predictors of turnover. Therefore, prediction of employee turnover might offer the opportunity for managers to reduce turnover. The aim of this research is to predict individual employee turnover using different machine learning techniques. This study investigates to what extent employee turnover can be predicted and which machine learning techniques work best. The study is conducted on a small imbalanced dataset. Data mining techniques are used to handle the limitations in the provided dataset. The results indicate a modest improvement in accurately predicting employee turnover.

Content

1 Introduction

Employee turnover is an extensively investigated term within organizations. The term "turnover" is defined by Price (1977) as: "The ratio of the employees of an organization who left in a particular period of time with the average number of employees in that organization during the same period of time". It turns out that employee turnover has many negative effects. Employee turnover within an organization goes together with certain costs, which can be split into direct costs and indirect costs. Examples of direct costs are replacement, recruitment and selection, temporary staff and management time. In the same manner, indirect costs are morale, pressure on remaining staff, costs of learning, product/service quality, organizational memory and the loss of social capital (Dess and Shaw, 2001).
Prediction of employee turnover can help in gaining insight in the top predictors of turnover and the identification of at-risk employees might give managers the chance to anticipate on time to prevent turnover. Voluntary turnover is of interest because in most cases, this represents the highest amount of turnover within an organization.
In machine learning and statistics, classification is the most familiar and effective technique used to classify and predict values. Classification is considered as an instance of supervised learning, which means that a training set of correctly identified observations is available for learning.

1.1 Purpose of this study

The purpose of this study is to predict individual employee turnover using several machine learning techniques. The aim is to predict whether an employee will stay at or leave the organization. We will investigate to what extent employee turnover can be predicted and which technique works best. Furthermore, an overview of the top predictors of turnover will be given. In this paper, multiple classification methods and techniques are used to confirm and verify the results. Every method or technique has its advantages and disadvantages. The performance of each technique can depend on the available data. For this research, a small dataset is available, which includes limited number of attributes and records. In the end, the models will be evaluated by several evaluation metrics.

1.2 Paper overview

The rest of this paper is structured as follows. In Section 2, a review is provided of previous work done on the prediction of employee turnover. Then, we will give

a description of the data in Section 3. In Section 4, we will describe the methodology. In Section 5, the obtained results are discussed. The paper concludes with a summary of the achievements and discussion of future work.

2 Relevant research

A lot of research is done to predict employee turnover based on demographic variables by applying machine learning techniques. In this chapter, we will indicate some interesting researches that will give a theoretical basis for our model choice. Furthermore, the most important explanatory attributes from previous research will be presented.

Models
Hong and Chao (2007) applied Logit and Probit models to predict voluntary employee turnover. The difference between these models is the used link function. They demonstrated that the prediction accuracy of the Logit model is superior to that of the Probit model. Logit is defined as the inverse of the logistic function. Logistic Regression can be interpreted as modeling log odds.

A study of Sikaroudi et al. (2015) show that employee turnover is predictable. They used different models. Random Forest had the best performance with a predicting accuracy of 90.6%. A downside of this model is that it did not utilize any post prune technique in recording elapsed time of processing.

In a recent study, Punnoose and Ajit (2016) applied Extreme Gradient Boosting (XGB) to predict employee turnover besides the more usual techniques like Logistic Regression, Naïve Bayes, Random Forest, SVM, LDA, and KNN. They obtained very good results with the XGB classifier. Punnoose and Ajit conclude that the two tree-based classifiers in Random Forest and XGB perform better than the other classifiers during training. During testing XGB is significantly better than Random Forest. The XGB classifier outperforms the other classifiers in terms of accuracy and memory utilization.

Sexton et al. (2005) utilized a Neural Network solution to predict voluntary employee turnover, because this method has been found a successful prediction tool for business problems. This study found that a trained Neural Network can sufficiently predict the turnover rate for a small mid-west manufacturing company.

Attributes
Cotton and Tuttle (1986) presented a meta-analytic review of voluntary turnover studies. It was found that the most important attributes for voluntary turnover

were age, length of service, salary, overall job satisfaction, and employee's perceptions of fairness. Moreover, similar studies conclude that personal variables such as age, gender, ethnicity, education and marital status, are important predictors.

Grissom, Nicholson-Crotty and Keiser (2012) found statistical evidence that manager gender matters for satisfaction and turnover in the public sector. They conclude that male teachers with female managers have lower satisfaction and higher turnover rates.

3 Data

3.1 Description of the dataset

The dataset we will use for modelling is the 'Individual Turnover' dataset, which includes information about employees working in a financial services organization across 10 different countries. The Individual Turnover dataset contains 1653 records (according to the number of employees) and 8 attributes (referring to the information about the employee). A description of the available attributes, including the attribute type, is shown in table 1.

ATTRIBUTE	TYPE	DESCRIPTION
ID	Label	Unique number of the employee
BOSSGENDER	Categorical	Gender of the manager
GENDER	Categorical	Gender of the individual
AGE	Numerical	Age of individual in years
LENGTHOFSERVICE	Numerical	Number of completed years at the organization
APPRAISALRATING	Categorical	Performance appraisal rating
COUNTRY	Categorical	Country in which the individual is working
LEAVERSTATUS	Categorical	Whether the individual has left the organization or not

Table 1: Description of the dataset "Individul Turnover".

Table 2 shows an overview of the range values of the attributes. For the categorical attributes, the frequency of the category is included between brackets. Furthermore, observed statistics of the attributes and the number of missing values are shown in respectively column 3 and 4. For the attributes BOSSGENDER and GENDER, '0' indicates that the employee is a female and '1' defines a male employee. APPRAISALRATING is defined on a scale from 1 to 5, with 5 being the highest. For the LEAVERSTATUS attribute, '0' indicates whether an employee

has left the organization (1) or not (0). The distribution between stayed and left employees is 87 percent against 13 percent.

ATTRIBUTE	RANGE/FREQUENCY	LOCATION	# MV
ID	{1, 2, ..., 1653}	-	-
BOSSGENDER	{0 (520), 1 (1133)}	mode = 1	-
GENDER	{0 (832), 1 (819)}	mode = 0	2
AGE	[16, 66]	mean = 38.198; σ = 9.523	-
LENGTHOFSERVICE	[0, 42]	mean = 9.799; σ = 9.703	-
APPRAISALRATING	{1 (18), 2 (106), 3 (114), 4 (1221), 5 (194)}	mode = 4	-
COUNTRY	{1 Belgium (40), 2 Sweden (36), 3 Italy (70), 4 France (37), 5 Poland (40), 6 Mexico (46), 7 Spain (76), 8 United Kingdom (196), 9 United States (1065), 10 Australia (47)}	mode = United States	-
LEAVERSTATUS	{0 (1441), 1 (211)}	mode = 0	1

Table 2: Range values of the attributes.

3.2 Exploration and pre-processing

Data cleansing
The attribute ID is removed from the original dataset, because this attribute indicates the unique number of the employee and has no predictive value. Besides that, the dataset contains three missing values, one is for LEAVERSTATUS and two are for GENDER. Because these missing values are from only three employees out of 1653, we choose to remove these employees from the data.

Data standardization
The numeric variables in the dataset, AGE and LENGTHOFSERVICE are normalized by min-max normalization [1]. This is done because training a neural network is more efficient when numeric independent data are normalized, so that their magnitudes are relatively similar. This leads to a better predictor.

Variances in individual turnover per country
An overview of the distribution of employees that left the organization per country is shown in figure 1. This histogram shows the percentage of leavers from each country. It looks like there are quite some differences in the part of leavers. Sweden has the highest percentage of leavers with 22 percent, which is 9 percent above the leavers in the total dataset. Besides that, the United States have the lowest percent of leavers with 11 percent.

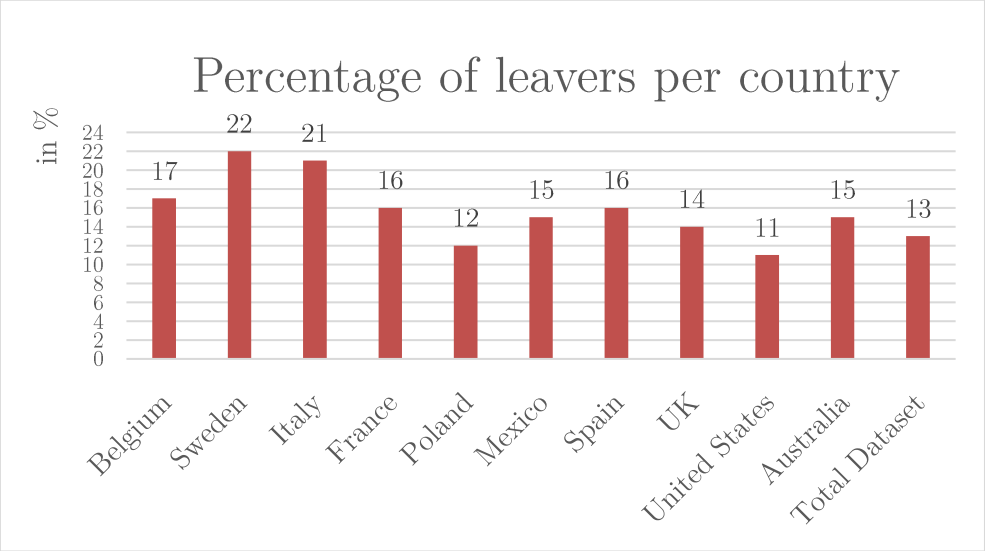

Figure 1: Distribution of employees that left per country.

To verify the independence of individual turnover for each country we apply the Chi-square test. This test is a statistical method to determine if two categorical variables have a significant correlation between them.
The Chi-square test gives a parameter value of 14.509 with 9 degrees of freedom and a P-value of 0.105. Because the P-value is bigger than 0.05, we can conclude that there is no significant difference in individual turnover between the ten different countries. Therefore, we can put all the data from each country in one dataset for modelling.

After the data exploration and pre-processing, we end up with a dataset with 1650 records and 7 attributes.

3.3 Handling imbalanced data

We have to deal with an imbalanced dataset. This means that the classes of the predicted attribute LEAVERSTATUS are not represented equally. There is one class of the predicted attribute that is in minority. The part of employees that stayed at the organization is 87 percent, where the other 13 percent has left the organization. A problem that can occur is that the learning algorithm generates a

trivial classifier that classifies every employee as the majority class, because this is the most common value, and therefore give a high accuracy. To tackle this problem, we can balance the training set by over-sampling the minority class and/or under-sampling the majority class. Therefore, we apply the Synthetic Minority Oversampling Technique (SMOTE) [2]. SMOTE is a statistical technique for increasing the number of cases in a balanced way.
SMOTE works by generating new employees from existing employees that are supplied as input. The new employees are created by taking samples of the feature space for each target class and its nearest neighbours, and generating new examples that combine features of the target case with features of its neighbours. This approach increases the features available to each class and makes the samples more general. The parameters of this technique are k, %under and %over. k equals the number of nearest neighbours, so how many data points are used to create a new one. The parameters %under and %over control the amount of over-sampling of the minority class and under-sampling of the majority classes.

4 Methods

4.1 Procedure

The dataset was implemented into R to predict employee turnover using four learning algorithms, which are Logistic Regression, Random Forest, an Artificial Neural Network and Extreme Gradient Boosting. In section 4.2, a description is given of each learning algorithm. In order to train the models and test how well the models perform on unknown data, we split the dataset into a training and a test set. The training set consist of 80 percent of the data, and the remaining 20 percent is included in the test set. After that, the performance of the models is examined based on the evaluation metrics described in section 4.3

4.2 Predictive models for employee turnover

In this section, the applied models are described. Every method or technique has its advantages and disadvantages.

4.2.1 Logistic Regression

Logistic Regression is a statistical method that is appropriate for examining the relationship between a categorical response variable and one or more categorical or continuous explanatory variables. Logistic regression is used for analysis of data when the predicted outcome is either a 1 or 0. This is the case for predicting

turnover, where people leave an organization (1) or they stay (0). We aim to predict turnover, a categorical response variable, using demographical variables, which can be categorical or continuous. A downside to Logistic Regression is that the output can be a little difficult to interpret.
The model is generally presented in the following format, where β refers to the parameters and x represents the independent variables:

$$\log\ odds\ = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

The log $odds$ or log-odds ratio, is defined by:

$$\ln[\tfrac{p}{1-p}]$$

The log(odds) expresses the natural logarithm of the ratio between the probability that an event will occur, p(Y = 1), to the probability that it will not occur. We are usually concerned with the predicted probability of an event occurring and that is defined by:

$$p = \frac{1}{1+e^{-z}} \quad where\ z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

4.2.2 Random Forest

Random Forest belongs to the decision tree models. A decision tree model starts with all the employees and then sorts them into smaller and smaller groups based on their likelihood of turnover. This creates a tree-like structure with a central node and many leaves for each path. The Random Forest technique takes the decision tree concept further by producing a large number of decision trees. This approach technique takes random selections of data from the dataset, and groups them into their own decision trees. It then takes the average of all the trees that it creates to make a prediction. An advantage of Random Forests is that it can be used to rank the importance of variables. Other advantages are that Random Forests are simple to understand and interpret and are robust to noisy data. A disadvantage of this technique is that it does not deal with data including categorical variables with different number of levels.

4.2.3 Artificial Neural Network

Artificial Neural Networks (ANNs) are statistical learning models, based on the neural structure of the brain that are used in machine learning. The brain basically learns from experience. The goal of an Artificial Neural Network is to solve a problem in the same way that the human brain would.
The ANN can be divided into three main parts: the input layer, the hidden layer and the output layer.

The input layer consists of all the predictor attributes. Each value from the input layer is sent to all the hidden nodes in the hidden layer. The values entering a hidden node are multiplied by weights.
To get the final value, the activation function is applied to the hidden layer sums. The sigmoid function has been used as the activation function of artificial neurons. A sigmoid function is a mathematical function having an "S" shaped curve (sigmoid curve) and is defined by the formula:

$$S\ t\ = \frac{1}{1+e^{-t}}.$$

The purpose of the sigmoid activation function is to transform the input signal into an output signal between 0 and 1.
An Artificial Neural Network can have any number of layers, and any number of nodes per layer. In this study, we will build a model with one hidden layer. The number of hidden nodes is chosen after training the model.

4.2.4 Extreme Gradient Boosting

Extreme Gradient Boosting (XGB) belongs also to the decision tree models. In that, XGB is similar to Random Forests but it uses a different approach to train the model. It uses a more regularized-model formalization to control over-fitting. XGB follows the principle of gradient boosting. Boosting is an ensemble technique in which learners are learned sequentially. With early learners fitting simple models to the data, which are any better than models fully based on chance agreement, and then analysing the data for errors. Later models focus primarily on those examples. The idea is that many weak learners are combined into a complex strong learner. A theoretical background of the model construction of XGB is presented by Chen and Guestrin (2016).

4.3 Evaluation metrics

To evaluate the performance of the model, we use 10-fold cross-validation. To examine how well the models perform, several evaluation metrics are used. In this section, we give a description of the used evaluation metrics.

Classification accuracy
The classification accuracy is the ratio of the number of correct predictions out of all predictions made. This measure is presented as a percentage, where 100% indicates that all the predictions are well made.

Cohen's kappa

A more robust measure is Cohen's kappa. This is a measure between 0 and 1, where 1 indicates and takes into account the possibility that a good prediction occurred by chance. Because the data has unbalanced classes, this metric is very important. In table 3 we can see how well the model performs for the intervals of kappa.

The definition of κ is:

$$\kappa \equiv \frac{O - E}{1 - E}$$

where O is the observed number of correct classifications, identical to the accuracy. E is the expected accuracy under chance agreement.

KAPPA	PERFORMANCE
<0	Poor
0 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
> 0.81	Almost perfect

Table 3: Models performance for Cohen's Kappa.

Area under the ROC-curve

The area under the ROC-curve (AUC) measures the trade-off between true positive rates and false positive rates. This measure is obtained from creating the ROC-curve and reflects the specificity and sensitivity of the model. The performance of the test depends on how well the test separates the stayers and leavers. An area of 1 represents a perfect model, where an area of .5 represents a worthless model. Table 4 shows the performance for the intervals of the area under the curve.

AUC	PERFORMANCE
0.51 - 0.60	Fail
0.61 - 0.70	Poor
0.71 - 0.80	Fair
0.81 - 0.90	Good
0.91 - 1.00	Excellent

Table 4: Models performance for AUC.

5 Results

This section describes the results of the predictive models according the evaluation metrics. First, the models were trained without applying SMOTE. However, the resulting classification matrices for learning algorithm show that only one class is predicted. This is the class of the employees that stayed at the organization which is in 87 percent of the cases true and the highest accuracy is accomplished. Therefore, we trained the models after applying SMOTE for several values for the parameters k, %over and %under. For each learning algorithm, the SMOTE parameters are selected based on the highest value for Cohen's kappa.

5.1 Balanced datasets

SMOTE is applied on the training set to create a more balanced dataset in order to better train the models. The values for k are: 0, 1, 3, 5, 7, 9. Furthermore, %over: 300, 500 and at last a combination of under and oversampling is used: %over is 100 and %under is 200. The parameter settings, for which the model reaches the highest Kappa, are shown in table 5. Also, the distribution between the employees that stayed and the employees is left is presented. Each machine learning technique has different optimal parameters for SMOTE. For example, the settings for Logistic Regression indicate for each left employee in the training set, five extra employees are generated using three nearest neighbours. Moreover, it can be seen that for the Neural Network the class of left employees are only duplicated twice and no new unique employees are generated.

MODEL	SMOTE SETTINGS	STAYED : LEFT
LOGISTIC REGRESSION	k = 3, over% = 500	62.5% : 37.5%
RANDOM FOREST	k = 7, over% = 100, under% = 200	50% : 50%
NEURAL NETWORK	k = 0, over% = 300	60% : 40%
XGB	k = 3, over% = 300	60% : 40%

Table 5: SMOTE settings for each model including distribution stayed and left.

5.2 Evaluation of models' performance

Table 6 shows the results of the evaluation metrics for the models. XGB has the highest classification accuracy, in 83 percent of the cases it predicts the right outcome. Furthermore, we see that the performance based on Cohen's kappa is comparable for all the models, with 0.14, and 0.13 for XGB. This value for kappa indicates a slight performance as can be seen in 3. At last, the area under the ROC

curve is presented. These values are also comparable, and show a poor performance (table 4).

MODEL	CLASSIFICATION ACCURACY	COHEN'S KAPPA	AUC
LOGISTIC REGRESSION	77.2%	0.14	0.59
RANDOM FOREST	65.6%	0.14	0.63
NEURAL NETWORK	81.5%	0.14	0.57
XGB	83.0%	0.13	0.56

Table 6: Performance of the models

Improving the models
Based on Cohen's kappa and the AUC values, it can be concluded that there is not really a well performing model. We tried to the improve results by two changes. The first was by using repeated cross-validation instead of cross-validation. The second possible improvement was by imputing the two missing values from the GENDER attribute with the mode of GENDER from the country the employee works. However, these two changes did not improve the model performance and therefore the models are not changed.

5.3 Model parameters

After training, the optimal parameters for the multiple classifiers are found. For the Random Forest, the optimal number of trees is 17. The optimal number of hidden units in the hidden layer of the Neural Network is 5. And for Extreme Gradient Boosting, the final values used for the model are: $nrounds = 150$, $max\_depth = 3$, $\eta = 0.4$, $\gamma = 0$ and $colsample\_bytree = 0.6$.
Where *nrounds* defines the number of trees, *max_depth* equals the maximum depth of a tree, η is the learning rate, γ specifies the minimum loss reduction required to make a split, and *colsample_bytree* defines the fraction of columns to be randomly samples for each tree.

5.4 Important attributes

After the models are constructed, we can identify which attributes have the strongest predictive power. Figures 2-5 show the relative importance of the top 5

most influential predictor attributes for the four models. In figure 2, the most influential predictor attributes for the Logistic Regression model are shown. It can be seen that the strongest prediction is made when the value of APPRAISALRATING equals 4, the employees' boss is a female and the employee works in the United States. Besides that, an employee working in Spain and LENGTHOFSERVICE are in the top 5 of most important attributes. In figure 4 the most important attributes are shown for the Neural Network, which are GENDER, AGE, BOSSGENDER, APPRAISALRATING. The numeric attributes AGE and LENGTHOFSERVICE have the most predictability for the decision tree models Random Forest and Extreme Gradient Boosting as can be seen in the figures 3 and 5.

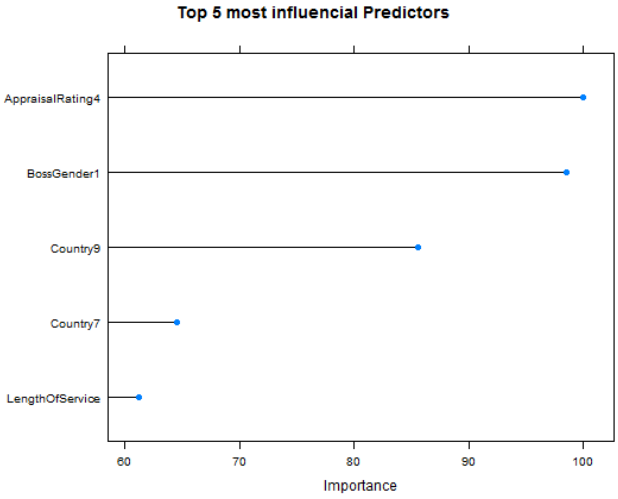

Figure 2: Top 5 most influential predictors for Logistic Regression.

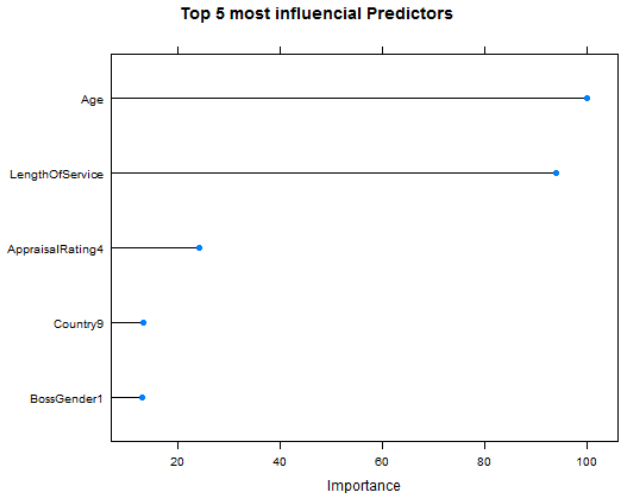

Figure 3: Top 5 most influential predictors for Random Forest.

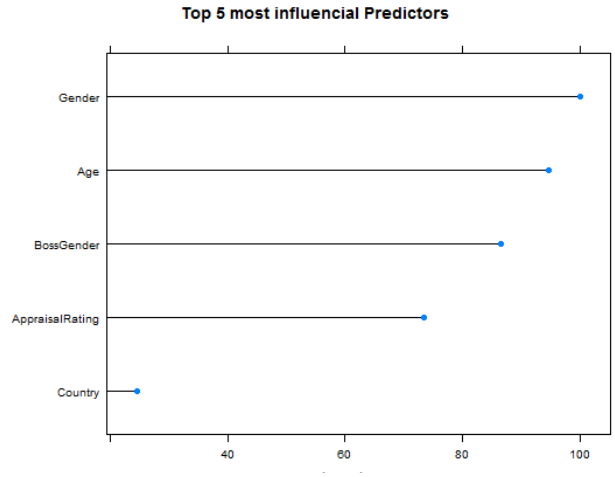

Figure 4: Top 5 most influential predictors for Neural Network.

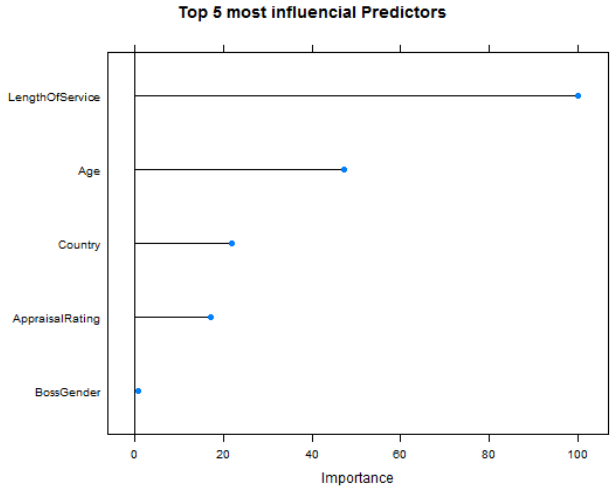

Figure 5: Top 5 most influential predictors for XGB.

6 Conclusion and discussion

This study investigated to what extent employee turnover is predictable using multiple machine learning techniques and which technique works best. Furthermore, this research gives an insight in which attributes have the strongest predictive power for the constructed models.
Although previous research about the prediction of employee turnover got very good results, this study concludes that employee turnover has limited predictability. In many previous studies, the classification accuracy was used as the performance measure. In this research, the models are especially evaluated on Cohen's kappa. The result for this evaluation metric was poor for all the models, and this indicates that the models do slightly perform better than when a prediction is made on chance agreement. The performance of the models is comparable to each other.
Furthermore, an insight in the most important attributes was given. The numeric attributes, gender of the employee and the number of years that an employee works at the organization, have the strongest predictive power for the decision tree based models Random Forest and Extreme Gradient Boosting. For the Logistic Regression model, the most important predictors for leaving the organization are when the performance appraisal rating equals 4, the employees' boss is a female and the employee works in the United States. Finally, the most important attributes for the Neural Network are gender of the employee, age in years, gender of the boss and performance appraisal rating.
The study was conducted on a small imbalanced dataset. For future work, it would be recommended to collect more data. More employees and especially more information about each employee might give a better prediction. Furthermore, this research uses all the available attributes. It might be better to investigate which attributes are significant, and remove the insignificant attributes from the data.

References

[1] J. Han and M. Kamber, Data mining-concepts and techniques, San Francisco: Morgan Kaufmann Publishers, 2006.

[2] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, „SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[3] J. Price, „The study of turnover," IA, Iowa state university press, 1977, pp. 10-25.

[4] S. Kotsiantis, „Supervised Machine Learning: A Review of Classification Techniques," *Informatica,* nr. 31, pp. 249-268, 2007.

[5] J. Grissom, J. Nicholson-Crotty and L. Keiser, „Does my boss's gender matter? Explaning job satisfaction and employee turnover in the bureaucracy," in *Midwest Poltical Science Association*, Chicago, IL, 2011.

[6] R. Sexton, S. McMurtrey, J. Michalopoulos and A. Smith, „Employee turnover: A neural network solution," *Computers & Operations Research,* vol. 10, nr. 32, pp. 2635-2651, 2005.

[7] G. Dess and J. Shaw, „Voluntary turnover, social capital, and organizational performance.," *Academy of Management Review,* nr. 26, pp. 46-456, 2001.

[8] W.-C. Hong and R.-M. Chao, „A comparative test of two employee turnover prediction models," *International Journal of Management,* vol. 24, nr. 2, pp. 216-229, 2007.

[9] E. Sikaroudi, R. Ghousi and A. Sikaroudi, „A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)," *Journal of Industrial and Systems Engineering,* vol. 8, nr. 4, pp. 106-121, 2015.

[10] R. Punnoose and P. Ajit, „Prediction of employee turnover in organizations using machine learning algorithms," *International Journal of Advanced Research in Artifical Intelligence,* vol. 5, nr. 9, 2016.

[11] J. L. Cotton and J. M. Tuttle, „Employee turnover: A meta-analysis nd review with implications for research," *Academy of Management Review,* vol. 11, nr. 1, pp. 55-70, 1986.

[12] T. Chen and C. Guestrin, „XGBoost: A scalable tree boosting system," *arXiv:1603.02754v1*, 2016.