Analysing which factors are of influence in predicting the employee turnover

Research Paper Business Analytics

HMN Yousaf

Supervised by

Dr. Sandjai Bhulai

Nov 23, 2016

Abstract

Employee turnover is a serious issue for organizations in current global economy. Human resources are the key assets for businesses to sustain their competitive advantage. Organizations want to understand the key issues behind employee turnover phenomena. Prediction models are highly related to human resource management to understand the employee turnover patterns from historical data. This research analyzes the factors which have influence in predicting the employee turnover. The study is conducted on a dataset provided by focus orange and different predictive models are tested on this dataset. The results of this research indicate that several factors like age, location, currency and business level etc. have an influence on employee turnover. The limitations involved in the provided dataset are handled by data mining techniques. This study is useful for both industry and research perspective.

Contents

1 Introduction

'Employee turnover' as a term is widely used in knowledge based organizations. Productivity of such organizations is highly dependent on their employees. When employees leave an organization, they carry with them invaluable knowledge which is often the source of competitive advantage for the business. In a highly competitive market, employee turnover poses risk and challenges for organizations. The impact of turnover has received considerable attention by senior management and human resources professionals. It has proven to be one of the most costly and seemingly intractable human resource challenges confronting by several organizations. Understanding the reasons of employee turnover and its impact on a business is essential in all organizations. Turnover can be considered as a subgroup of human resource management (HRM). HRM function is to motivate employees and enhance workforce effectiveness. Integrating information technologies and HRM will provide smarter work. Globally competitive organizations will depend on the uniqueness of their human resources and the systems for managing human resources effectively to gain competitive advantage (Pfeffer 1994, Bartlett & Ghoshal 1997, Barney & Wright 1998)

1.1 Purpose of this Study

The purpose of this research study is to analyze the factors which influences the employee turnover in an organization using data mining techniques. Different predictive models are used to understand the turnover phenomenon. The result of this analysis may provide the recommendations which can enhance the efficiency and effectiveness of human resource planning processes that are used to focus on the employee turnover problem.

This study is conducted on real data which was provided by Focus Orange Technology, which is an Amsterdam based company. They help other companies to optimize the effective investment in their human capital and they have a Crunchr data platform which collects and validates all employee data and converts this into meaningful insights. For this study, they provided one of their client's data which was anonymized to fulfil the privacy rules. The necessary information which is used to answer our research question was present in the data.

1.2 Paper Overview

This research paper is organized as follows: the background information about employee turnover and related research is given in Section 2. The data which is used in this research is described in Section 3 with all data preprocessing and exploration details. Section 4 contains the methods and techniques which are conducted to answers our research question. The obtained results from different models are discussed and compared in Section 5. Lastly the conclusion and discussion about this research are given in Section 6.

2 Background & Related Research

The term turnover is defined by Price (1977) as the ratio of the number of organizational members that have left during the period under consideration divided by the average number of people in the organization during that period. Employee turnover has been one of the most studied subjects in organizational behavior literature (Schwab, 1991), yet continues to elude any concrete conclusions. The reason so much attention has been paid to the issue of turnover is because turnover has some significant effects on organizations (Denvir and Mcmahon, 1992). Many researchers argue that high turnover rates might have negative effects on the profitability of organizations if not managed properly (Wasmuth and Davis, 1993).

Controlling employee turnover can constitutes a complex and challenging task for both the workplace and managers. Managers may have difficulty understanding and accepting employee turnover within their organization, due to a myopic perspective of the situation. However, identifying the primary causes, quantifying the problem, and finding possible solutions to high employee turnover can prove to be valuable information for managers who wish to make a difference (Mobley, 1982). Although there is no standard framework for understanding the employees turnover process as whole, a wide range of factors have been found useful in interpreting employee turnover (Kevin et al. 2004).

Kramer et al. (1995); Saks. (1996); Srinivasan, V. & Valk, R. (2008) among others have attempted to answer the question of what determines people's intention to quit by investigating possible antecedents of employees intentions to quit. There are several reasons why people leave organizations which could be used to predict intentions to quit and actual turnover. The range of factors may include lack of commitment, job dissatisfaction, unclear expectations of peers and supervisors, ambiguity of performance evaluation methods, job content, tenure, salary and demographics.

Booth and Hamer (2007) found that labor turnover is related to a variety of environmental factors and organizational factors such as company culture and values, supervisory style, fair pay, corporate value, giving support to each other, trust and respect between employees, manageable workload, development and career building satisfaction and degree of job satisfaction.

Previous research findings also indicated that some causes of employee turnover are job-related factors that are somewhat within the direct control of the employer. Examples of such factors would be dissatisfaction with working conditions, supervising conflicts, scheduling conflicts or salary discrepancies. Understanding the causes of job-related turnover is crucial in being able to identify problems within an organization that might be controlled by the employer. Corrective steps taken in this area include training programs for supervisors, clarification of the employee's purpose or role and identifying scheduling solutions (Ulschak & Snowantle, 1992).

3 Data Description, Exploration and Preprocessing

3.1 Data Description

The original data was provided in single csv file and it consists of 39 attributes and 10,616 instances. There are 10 attributes which have 100% missing values so they are removed before exploring anything in data. The rest of the 29 attributes can be divided into two categories. Their description and type are given in following tables:

Table 1 Attributes containing only employee ID's

Attributes	Type	Description
Job ID	categorical	Unique identifier of job
Position ID	categorical	Unique identifier of position
Employee ID	categorical	Unique identifier of employee
Functional manager ID	categorical	Unique identifier of functional manager

Table 2 Attributes containing employee information

Attributes	Type	Description
First name	categorical	First name of employee
Last name	categorical	Family name of employee
Gender	boolean	Gender of employee
Date of birth	date	Date of birth of employee
Date in service	date	Date when employee joined company
Date in position	date	Date when employee started current job
FTE	numerical	Part/Full-time percentage (100 mean full time)
Position title	categorical	Title of employee
Contract type	categorical	Contract type of employee
Employee status	categorical	Status of the job of employee
Position Grade	categorical	Grade of the job
Talent status	categorical	Talent status of employee
Performance Score	categorical	Performance status of employee
Potential Score	categorical	Potential of employee
Mobility	categorical	Mobility of employee (none, local, regional, global)
Retention risk	categorical	Retention risk of employee (none, low, medium, high)
Retention risk reason	categorical	Retention risk reason of employee (none, compensation, career)
Business level	categorical	Business Unit Hierarchy
Functional Area	categorical	Functional Area Hierarchy
Location level 1-3	categorical	Geographical Hierarchy (country, city, address)
Employee grade	categorical	Employee grade of employee (can be local and/or global grade)
Base salary	numerical	Compensation information in local currency
Currency	categorical	The local currency used for compensation
Hire type	categorical	Indication of an additional employee
Leave type	categorical	Indication of an employee removed

3.2 Data Exploration and Preprocessing

The attributes which have only ID information are removed from the data set because they do not show any useful property that could help in model building. Similarly, the attributes first name and last name are removed because they are specific to every instance and do not generalize any model. The attribute 'Employee status' is also removed because 99% of the instances have the same value for this attribute and the model cannot learn any useful information from this attribute. The 'Title' attribute was anonymized in the given data but it appears like an id value attribute. We cannot consider it in our analysis because it has about 2K factor levels within less then 9K instances and models do not work with such large number of factor levels. Another reason to remove this attribute is its business definition because it cannot be considered in numerical terms.

Before considering the missing values in all attributes, we did some data conversion which seems logical from business and modeling point of view. The proportion of "Leave type" values are checked because this attribute will be our response variable and we do not want to miss any valuable information. We assumed that all the instances that have no value in the Leave type attribute are set to "Stayed" which means employees are still working in a company. The proportion of all the other values in this attribute was very low so we merged them and set them to "Left" which means these employees have left the company. So, now the attribute "Leave type" is binary and it will be easy to model the binary classification problem.

Table 3 Conversion of attribute "Leave type" values

Leave Type	% of instances
Voluntary	0.42%
Terminated	0.26%
Left	0.14%
Involuntary	1.38%
Stayed	97.80%

Leave Type	% of instances
Left	2.20%
Stayed	97.80%

Similarly, the attribute "Hire type" contains values Hire, Rehire and a lot of missing data. We assumed that the missing values mean that the employee are not new hires but the old employees. This way the missing values for the "Hire type" attribute reached to 0%. There is no missing value in the FTE attribute but its description shows that it can be converted to a binary attribute. So, if the FTE value is equal to 100% then its mean the employee is working full-time and if less than 100% then the employee is working part-time. Moreover, the two out of six categories of attribute "Contract type" show very low proportion of data which is less than 0.01%. We decided to merge these two categories with "Temporary" which is also makes sense by a business point of view.

Table 4 Updated proportion of attributes values after data processing

Contract Type	% of instances		Hire Type	% of instances		FTE	% of instances
Probation	1.27%		Rehire	0.18%		Part-time	3.55%
Temporary	3.46%		New hire	1.71%		Full-time	96.45%
Fixed Term	4.15%		Old	98.11%			
Permanent	41.55%						
Missing	49.57%						

Afterwards, the attributes are checked for missing values. The following figure shows the missing data proportion of every attribute in red.
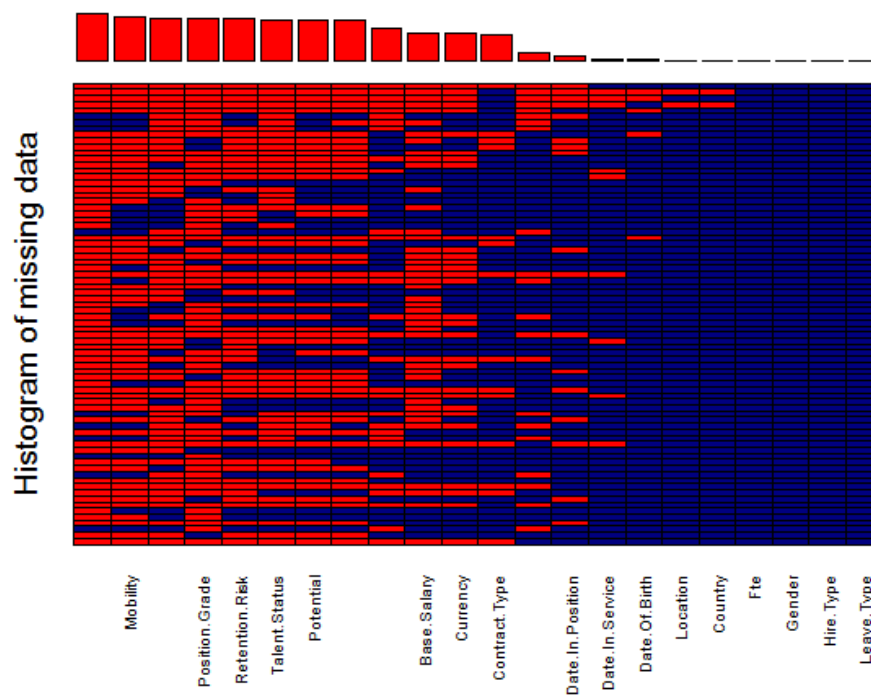

Figure 1 –Missing data proportion of attributes

As can be seen from the above figure that about 8 of the attributes have more than 70% of missing data. These attributes are excluded from our analysis because no valuable information can be learned from these.

Table 5 Excluded data attributes based on missing values

Attributes	% Missing	Attributes	% Missing
Retention Risk	80%	Mobility	86%
Performance Status	78%	Retention risk reason	90%
Potential	79%	Position Grade	81%
Employee Grade	82%	Talent Status	79%

Further investigation showed that some of the attributes have missing data about 50-60% but the major portion of this missing data comes from instances which are specific to United States. We did not exclude these attributes because these can be useful for Europe and we assumed that the USA offices do not measure these attributes so it is better to split data based on location. One part contains all the instances from Europe and the second part contains all the instances from USA.

Table 6 Europe data attributes

Attributes	% Missing
Gender	0.0%
Date of birth	0.1%
Date in service	0.2%
Date in position	17.0%
FTE	0%
Business level	24%
Functional Area	24%
Location level	0%
Contract type	0.1%
Base pay	9%
Currency	9%
Hire type	0%
Leave type	0%

Table 7 USA data attributes

Attributes	% Missing
Gender	0.0%
Date of birth	0.1%
Date in service	0.1%
Date in position	0.1%
Business level	4%
Hire type	0%
Leave type	0%

The missing values in the Europe dataset are handled step by step for each attribute. First, the instances which have missing data for most of the attributes are excluded from the dataset. Then each attribute is checked again for missing data. The missing values for 'Date in position' attribute are filled by taking values from 'Date in service' attribute because we assume that date in position for these employees should at least equal to date in service.

Since every country has its own currency so the missing values for 'Currency' attribute are taken by considering the 'Location' attribute. To get the missing data for 'Base Pay' we build a simple regression tree model on all available attributes in the Europe data. The predicted values from this model are then inserted in place of missing values in base pay attribute.

There are few attributes and lower proportion of missing values in the USA dataset so missing values are handled by the same techniques as used in the Europe dataset. More-over, the 'FTE' attribute has the same value 'Full-time' for all the instances in this dataset so we decided to remove this attribute from the USA dataset because it does not give any additional information to enhance the model learning.

After dealing with missing data some new attributes are created. Instead of using date of birth directly, the age of each employee is calculated based on the last date of the month when data was provided. Similarly, the duration of employee service and the duration in current position are calculated. The age and these newly calculated durations are then converted into bins so we can use them as a factor instead of dealing them as numerical attributes.

In addition to two data parts namely USA and Europe, we also decided to build the aggregated one. The common attributes of these two datasets are merged together. The number of instances in each data set and the proportion of our response variable values are given in below table.

Table 8 Final datasets with no. of predictors and response percentage

Dataset	No. of instances	No. of predictors	Leave Type	
			Stayed	Left
USA	5069	6	96.88%	3.12%
Europe	4019	12	98.16%	1.84%
Total	9086	8	97.45%	2.55%

4 Modeling Methods

4.1 Procedure

We have three datasets which will be used for our analysis. As we want to analyze the factors which cause the employee to leave the company, we will use the 'Leave type' attribute as our response variable and we will predict either the employee will stay or leave the company based on the predictors in each dataset. To achieve this goal, the standard data mining approach will be applied. The best practice while applying data mining techniques is to split the dataset randomly into training and testing. The main objective of this separation of dataset into training and testing is to make sure that the model is generalized and applicable to any new data instance instead of being specific to a given dataset. This splitting of data is only possible if we have abundant instances so the model could learn enough to generalize the results. Since our data instances are very limited, we will use the same dataset for training and testing purpose.

4.2 Models

Three different prediction models will be used for every dataset. These models will be trained and tested using the same dataset and accuracy will be determined using the cross-validation technique. The selected models are logistic regression, artificial neural networks and random forest.

4.2.1 Logistic Regression

Logistic regression is a specialized form of regression used to predict and explain a categorical dependent variable. It works best when the dependent variable is a binary categorical variable. One special advantage of logistic regression is that it is not restricted by the normality assumption which is a basic assumption in the regression analysis. This technique can also accommodate non-metric variables such as nominal or categorical variables by coding them into dummy variables. Another advantage of logistic regression is that it directly predicts the probability of an event occurring. To make sure that the dependent variable, which is the probability, is bounded between zero and one, the logistic regression defines a relationship between the dependent and independent variables that resembles an S-shaped curve, which uses an iterative process to estimate the 'most likely' values of the coefficients. This results in the use of a 'likelihood' function in fitting the equation rather than using the sum of squares approach of the regression analysis. The dependent variable is considered as the 'odds ratio' of a specific observation belonging to a particular group or category. In that sense, logistic regression estimates the probability directly. (Srinivasan, V. & Valk, R. 2008)

4.2.2 Artificial Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing

10

system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example (Maind & Wankar 2014). The ANN accepts the values of inputs by input nodes which is also called input layer. These input values are then multiplied by a set of numbers which are called weights and stored in the links. After multiplication, these values are added together to become inputs to the set of nodes that are to the right of the input nodes. This layer of nodes is usually referred to as the hidden layer. The number of hidden layers could be one to many but in our model, we will use only 1-layer ANNs. Finally, the values from the hidden layer are fed into an output node, where a special mapping or thresholding function is applied and the resulting number is mapped to the prediction class.

4.2.3 Random Forest

Random forest (RF) is an ensemble learning method that constructs multiple decision trees. Each decision tree can be implemented by a CART procedure. CART recursively partitions on a nominal target category to reach a tree structure. The input of CART can be nominal or numerical. As the decision tree grows, a feature must be identified to split on it. So, all features are compared to each other to select the best feature. This comparison can be done by the Gini index that measures pureness of feature separation. The CART stopping rule occurs when the target feature in the last separations are insignificant. RF samples randomly training data with replacement on constructing each decision tree that is called bagging. Each decision tree returns a class and then bagging combines them to reach a unique decision (Breiman & Friedman 2001).

4.3 Evaluation

To evaluate the prediction of our selected models, k-fold cross validation will be applied to reduce the bias of sampling data and ensuring model error randomness. K-fold cross validation randomly divides data into k subsets and one subset is used as testing data and k-1 subsets are used as training data. This process is repeated k times to cover all data. We will use 10-fold cross validation in our analysis.

In addition to the k-fold cross validation we will also use repeated cross validation in which a stratified partitioning will be used to split the data into train (75%) and test set (25%). Stratified partitioning splits the data in such a way that the proportion of response class values remain the same in both train and test datasets. This process will be repeated 100 times and the average value of each evaluating measure will be used to check the model performance.

To estimate the models performance, different evaluating measures can be considered. Since our problem is a 2-class classification, we will only use those measures which are considered best practices for such classification problems.

- **Overall accuracy:** is a measure that indicates the correctly predicted matches and non-matches. This may be problematic when the classes are not balanced.

$$Accuracy = \frac{\text{\# true positives } + \text{ \# true negatives}}{\text{(total \# of prediction)}}$$

- **Kappa:** statistic considers the expected error rate:

$$K = \frac{O - E}{1 - E}$$

(where O is the observed accuracy and E is the expected accuracy)

- **Sensitivity:** given that a result is truly an event, what is the probability that the model will predict an event results?

$$sensitivity = \frac{\text{\# true positives}}{\text{\# true positives } + \text{ \# false negatives}}$$

- **Specificity:** given that a result is truly not an event, what is the probability that the model will predict a negative result?

$$specificity = \frac{\text{\# true negatives}}{\text{\# true negatives } + \text{ \# false positives}}$$

- **ROC Curve:** With two classes the Receiver Operating Characteristic (ROC) curve can be used to estimate performance using a combination of sensitivity and (1-specificity). The area under the ROC curve is a common metric of performance.

4.4 Imbalanced Data Handling

As we have seen in data exploration and preprocessing part, there is a class imbalance problem with the response variable. A dataset is imbalanced if the classification categories are not approximately equally represented. The proportion of 'Left' class is very low (2-3%) in all three datasets which means our models prediction accuracy will be more biased towards the majority class. If we will not be able to predict the 'Left' class, then a concrete conclusion cannot be drawn from our analysis. To overcome the class imbalanced problem, the 'SMOTE' technique will be used.

4.4.1 SMOTE

Synthetic Minority Over-Sampling Technique (SMOTE) is an approach to the construction of classifiers from imbalanced datasets. In this approach the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement. The synthetic examples are generated in a less application-specific manner, by operating in "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to

the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general (Chawla, Bowyer, Hall, Kegelmeyer 2002).

5 Results

When all the selected models are applied to imbalanced datasets, the performance of all the predicting models were very poor. Although, the logistic regression performed with an accuracy of 98% but it was biased towards the majority class (stayed) and could not predict any instance of the minority class (Left). We could not use these results to check the factors which have influence in predicting the employee leave status. So, the response class balancing was mandatory to get valuable results.

5.1 Balanced Datasets

The SMOTE technique was applied to the response (Leave type) class in each dataset to get more balanced class values. The 3:1 ratio was selected as criteria for the response class values. To bring the majority and minority class values to this ratio level, the majority class was under sampled and the minority class was oversampled using SMOTE algorithm. More weight is given to oversample the minority class and less to under sample the majority class. The majority class cannot be under sampled at higher percentage because valuable data will be lost and the model cannot learn well.

Table 9 Datasets after applying SMOTE algorithm

Dataset	No. of instances	Leave Type	
		Stayed	Left
USA	5593	75.0%	25.0%
Europe	4736	75.0%	25.0%
Total	8212	75.0%	25.0%

5.2 Models Performance with Cross-Validation

10-fold cross-validation (CV) is used to check the performance of all predicting models Random forest (Rf), Neural network (Nnet) and Logistic regression (Glm) on each dataset. Accuracy and kappa are used as evaluation measures to compare the performance of these models. The mean values of the10-fold CV for each measure are given in the tables below and the range of these values from all predicting models are given in following figures.

Table 10 10-fold CV results

Europe Dataset

Model	Performance Measure	
	Accuracy	Kappa
Rf	94.07%	83.52%
Nnet	87.2%	62.05%
Glm	88.2%	66.17%

USA Dataset

Model	Performance Measure	
	Accuracy	Kappa
Rf	86.57%	61.26%
Nnet	86.52%	61.21%
Glm	84.48%	53.56%

Total Dataset

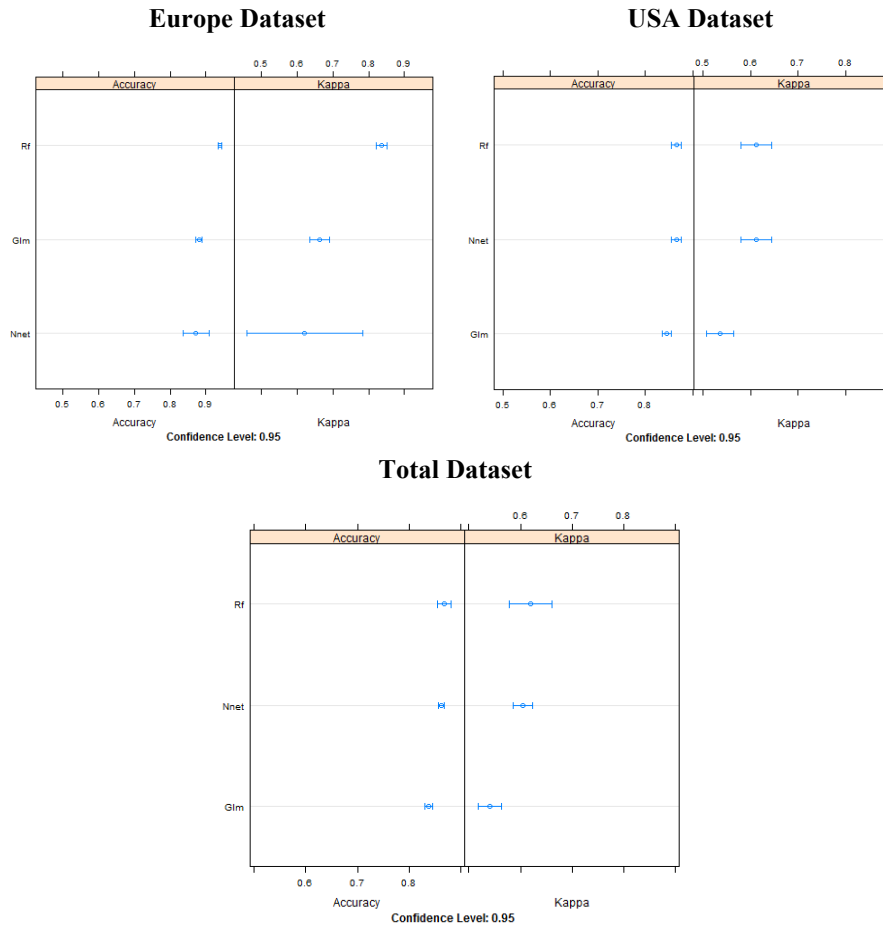| Model	Performance Measure	
	Accuracy	Kappa
Rf	86.79%	61.98%
Nnet	86.17%	60.52%
Glm	83.76%	54.01%

Figure 2 – Range of values for each dataset

For the Europe dataset, the random forest worked very well and accuracy is very high as compared to the other two models. Moreover, the neural network sometimes performed better than the logistic regression but the overall mean values from logistic regression are higher than the neural network because neural networks sometimes get stuck in local minima and perform poorly. The spread of neural networks values is big.

For the USA dataset, the results are lower than the Europe dataset because fewer variables were available and data cannot be explained well with few variables. Random forest worked almost the same as the neural network and the neural network values are also stable for this dataset. Logistic regression has given the lowest accuracy in this case.

For the Total dataset, the results are almost the same as for the USA dataset. The addition of two more variables as compared to the USA dataset does not make a significant

difference. Random forest and neural network have given approximately the same accuracy and the spread of the random forest values is bigger than the other two models.

5.3 Models Performance with repeated Cross-Validation

The three datasets are randomly divided into a training and testing part using stratified partitioning and each model is trained and tested on these training and testing data parts respectively. This process is repeated 100 times and the mean values of evaluating measures for each model are calculated. The performance of models with repeated cross validation is almost the same as with 10-fold cross validation. The mean values are given in the tables below and the range of these values are given in following figure.

Table 11 Repeated CV results

Europe Dataset

Model	Performance Measure			
	Accuracy	Kappa	Sensitivity	Specificity
Rf	93.53%	82.03%	81.17%	97.64%
Nnet	87.57%	58.82%	59.94%	96.78%
Glm	88.13%	66.27%	67.04%	95.16%

USA Dataset

Model	Performance Measure			
	Accuracy	Kappa	Sensitivity	Specificity
Rf	86.64%	61.53%	61.04%	95.36%
Nnet	86.36%	61.11%	61.84%	94.72%
Glm	84.76%	54.66%	52.89%	95.61%

Total Dataset

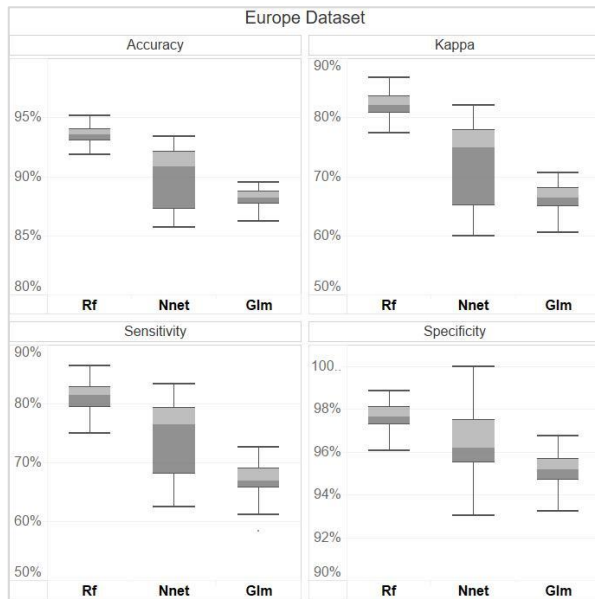| Model	Performance Measure			
	Accuracy	Kappa	Sensitivity	Specificity
Rf	86.65%	61.76%	61.78%	95.13%
Nnet	86.32%	61.56%	63.80%	94.00%
Glm	83.69%	53.70%	57.18%	92.72%

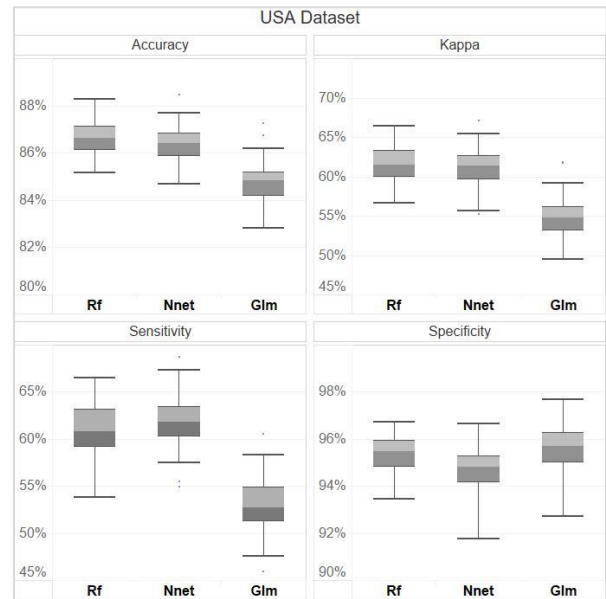**Figure 3** – Range of values for Europe dataset

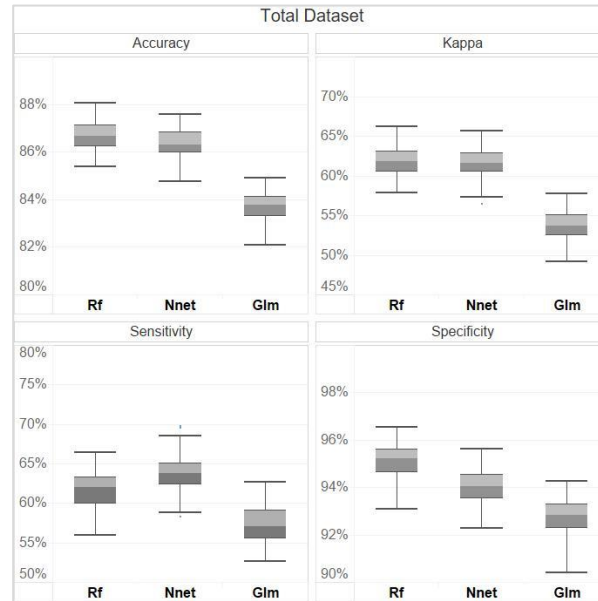**Figure 4** – Range of values for USA dataset

Figure 5 – Range of values for Total dataset

For the Europe dataset, the random forest again worked best and there is a clear difference of performance by all evaluation measures for each model. The neural network worked poorly because of local minima limitation. As it can be seen from the above tables that the neural network has the lowest 'Sensitivity' values as compared to other models because when a neural network get stuck in local minima, it cannot predict the minority class at all which affects the sensitivity. In the neural network case the overall accuracy is not a reliable measure. It is more biased towards a majority class. Moreover, the boxplots show that random forest and logistic regression values do not change a lot in each iteration but the spread of neural network values is very high because the values are not stable.

For the USA dataset, random forest performed better than the other two models but again the difference in measures is very small as compared to other models. However, logistic regression has given the lowest accuracy. The boxplots for this data show that the values spread is almost the same for all models. Moreover, the accuracy of random forest on this dataset is lower than the Europe dataset because fewer variables were available for this dataset.

For the Total dataset, the performance measures of each model are almost the same as in the USA dataset. The addition of two variables (Location, FTE) in the Total dataset does not make a big difference because all the instances that belong to USA have a single value for these variables. A little improvement in performance of these models on this dataset is due to the location variable instances that belong to the Europe dataset. Moreover, the boxplots show that the spread of values is almost the same for all the models on this dataset.

5.4 Evaluation by ROC Curve

The performance of predicting models is also evaluated using ROC curve. The datasets are divided into training and testing sets using stratified partitioning and each model is trained on a training set and evaluated on a testing set. ROC curves obtained from these models for every dataset are given in below figures.
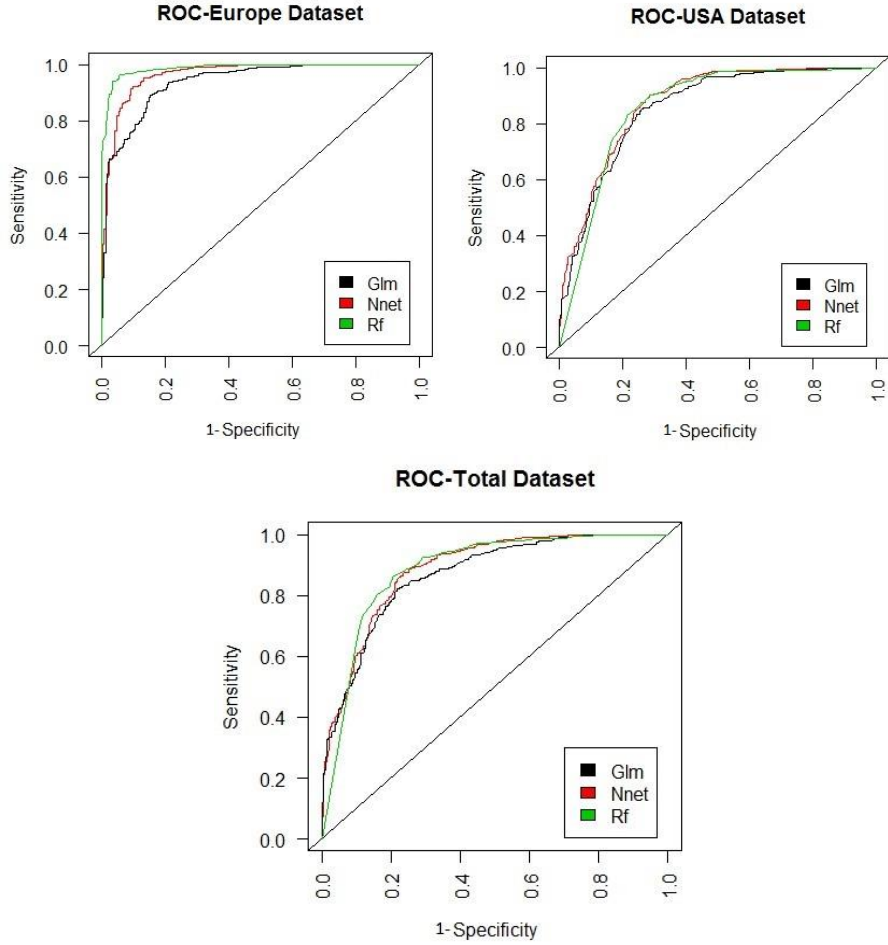

Figure 6 – ROC curves for each dataset

For the Europe dataset, the ROC curves verifies that the random forest worked best on this dataset. The neural network performance is better than the logistic regression but we already analyzed that neural networks can get stuck in local minima so different results can also be expected.

For the USA and Total datasets, the ROC curves show that all three models worked almost the same on these two datasets. The area under the curve for random forest is

18

higher than the other two models but this difference is not very significant. The sensitivity is lower which means models do not predict the minority class very well for these datasets.

5.5 Important Factors

It is clear in models performance analysis that the random forest gives the highest accuracy, kappa, sensitivity and specificity measures so it is better to check the factors/variables which played an important role to predict the response class. The importance of these variables is checked by the mean decrease of accuracy (MDA) measure which is a global variable importance measure. It is a mean decrease of accuracy over all out-of-bag cross validated predictions, when a given variable is permuted after training, but before prediction. It is easier to understand and robust as it is averaged over all predictions. The following figures show the variables in order of importance for model accuracy on each dataset.
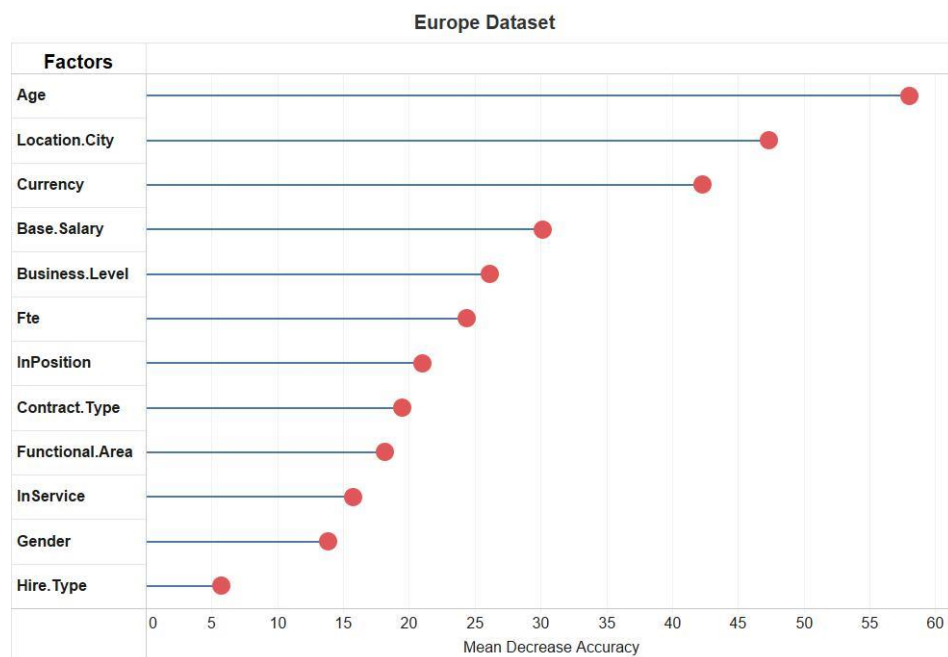

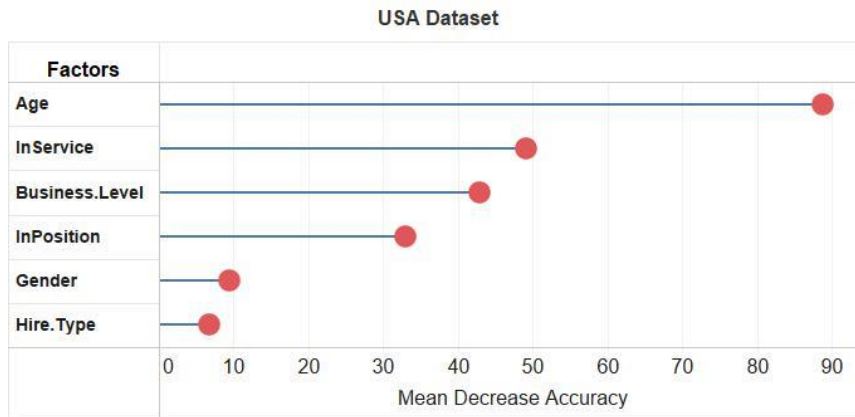**Figure 7** – Importance of factors for Europe dataset

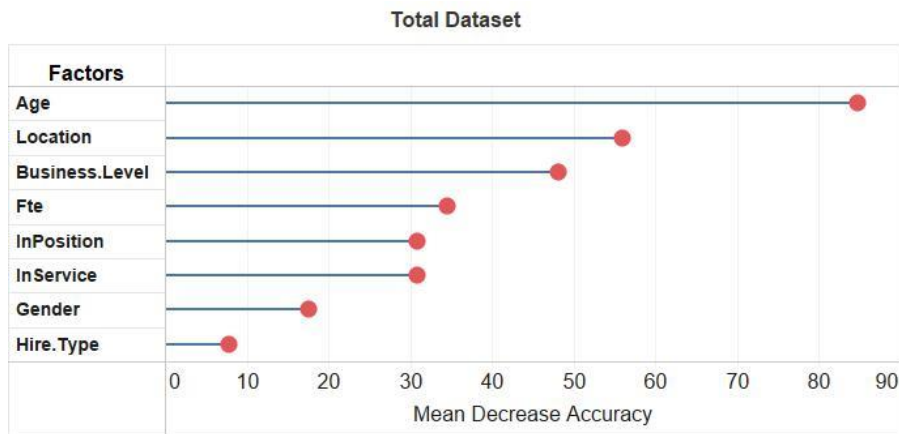**Figure 8** – Importance of factors for USA dataset

Figure 9 – Importance of factors for Total dataset

The above figures for the importance of factors show that Age is one of the most important factor in deciding the model accuracy for all three datasets. Location is the second best for both the Europe and Total datasets. Business-Level, In-Position and In-Service are also common in all three datasets which have high impact on model predictions. However, in the Europe dataset the Currency and Base-Salary have a very large influence on the model accuracy. It is noticeable that these two variables are not available for the rest of the two datasets and the model prediction accuracy for these two datasets is quite low as compared to the Europe dataset.

6 Conclusion & Discussion

The results of our predicting models indicate that the random forest works best on these datasets. It gives the highest accuracy and sensitivity values which means that it can predict the employee turnover and minority class (Left) more precisely. The highest achieved overall accuracy is (94.07%) on the Europe dataset and the highest sensitivity value is (81.17%) which means this model can predict the minority class with 81% accuracy and it will make only 3% (1-specificity) mistakes to predict the majority class as minority class. On the USA and Total datasets, the sensitivity value is low because fewer number of variables were available for these datasets.

The factors which have the highest influence on employee turnover are Age, Location, Currency, Base salary, Business level, FTE, In-position and In-service. The factors which have less impact in predicting the employee turnover are Hire type, Gender, Contract type and Functional area. The results also reveal that the Currency and Base Salary are among the most important factors because these factors were not available in the USA and Total datasets, so the accuracy of our model was quite low on these datasets. The location factor has also some effect because it was not available in the USA dataset and the accuracy of model was the lowest for this dataset.

Demographical factors like age and location are strong predictors of employee turnover because the younger employees from age 18-25 are more likely to turnover than older employees. Since younger employees leave in early stages so in-position and in-service also have some effect on turnover. These results are consistent with a study on turnover rates conducted by Hill and Associates which found that young undergraduates, graduates and post graduates in the outsourcing business had changed their jobs at least once in the past three years (Banerjee 2008). The location in our analysis has an impact because most of the employees who leave the company are from UK.

Since the currency is one of the most important predictors of employee turnover, it can be explained by the fact that people who are working in UK and not getting salary in GBP are more likely to leave the company. Moreover, the people who are working as a part-time employee are more likely to leave the company as compared to a full-time employee. The salary of the part-time employee is also lower than the full-time employee.

This study investigated the factors of influence on employee turnover using the data mining techniques. Three important conclusions can be made from this research study. First, it finds the importance of prediction models which can be used to predict the employee turnover. By considering different models, it is investigated that best prediction is possible using random forest. Secondly, the identification of the important factors like age, location, currency, business level, in-position, in-service and FTE are significant from a research perspective. Lastly, as this analysis is specific to the given dataset so these predictive accuracies can be used by the company who provided this

dataset and they can identify those employees who have turnover intentions even before they had made their final decision to leave.

This research study also reveals several issues for future research. First, the future research could get more balanced data from large sample so there should not be any need to generate data synthetically. Secondly, the available factors to understand the employee turnover behavior were very few. The factors like potential score, talent status, employee grade and retention risk etc. could be useful to understand the turnover phenomena deeply. More data can be collected for these factors as well as for those factors where data was missing. More data will help to do a more rigorous analysis and refine the prediction model. Lastly, more research should be conducted on different samples to check the validation of the prediction models proposed in this study. Additionally, other prediction models can also be tested to check the performance on this dataset.

This research is useful for HR professionals and managers. For every company the human resources are the source of competitive advantage in a current global economy. Manpower planning is one of the most important responsibility of the HR professionals. Therefore, tools and models that enhance understanding and prediction of factors which have influence on employee turnover can bring significant value to HR professionals. In recent years, various authors have urged human resource professionals to play the role of a strategic partner (Ulrich & Brockbank 2005).

References

Barney, J. B., & Wright, P. M. (1998). On becoming a strategic partner: The role of human resources in gaining competitive advantage. *Human Resource Management*, 37(1), 31-46.

Bartlett, C. A., & Ghoshal, S. (1997). The myth of the generic managers: new personal competencies for new management roles. California Management Review, 40(1), 92-116.

Pfeffer, (1994) Competitive advantage through people: Unleashing the power of the work force. Boston: Harvard Business School Press.

Price, J.L & Mueller, C.W (1981). A causal model of turnover for nurses, Academy of Management Journal, 24:543-565.

Schwab, D. P. (1991). Contextual variables in employee performance-turnover relationships. Academy of Management Journal. 34, 966-975.

Kevin MM, Joan LC, Adrian JW (2004). "Organizational change and employee turnover" Personnel Rev. 33 (2):161-166.

Denvir A, .McMahon F (1992). Labour turnover in London hostels and the cost effectiveness of preventive measures int. J. Hosp. Manage, 11 (2): 143-540.

Wasmuth WJ, Davis S W (1993). "Managing employee turnover: why employees leave", The Cornell HRA Quarterly, pp.11-18.

Kramer MW, Callister RR, Turban DB (1995). "Information-receiving and information-giving during job transitions", West. J. Commun. (59):151-70.

Saks AM (1996). "The relationship between the amount of helpfulness of entry training and work outcomes", Hum. Rel. 49: 429-451.

Nagadevara, V., Srinivasan, V. & Valk, R. (2008). Establishing a Link between Employee Turnover and Withdrawal Behaviours: Application of Data Mining Techniques, Research and Practice in Human Resource Management, 16(2), 81-99.

Booth, S., & Hamer, K. (2007). Labour turnover in the retail industry: Predicting the role of individual, organisational and environmental factors labour turnover in the retail industry. International Journal of Retail & Distribution Management, 35(4), 289-307.

Mobley, W. H. (1982). Employee turnover: causes, consequences, and control. Philippines: Addison-Wesley Publishing.

Ulschak, F.L., & Snowantle, S.M. (1992). Managing employee turnover; A guide for health care executives. Chicago, Illinois: American Hospital Publishing.

Maind, Sonali B., and Priyanka Wankar (2014). "Artificial Neural Network." SpringerReference (2014): n. pag 1-4 Web.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees: CRC press.

N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer Journal-ref: Journal Of Artificial Intelligence Research, Volume 16, pages 321-357, 2002.

Banerjee, I. (2008). Attrition: From corporate nightmare to competitive advantage. HRM Review, 68-72.