HR Churn Analysis

**Abstract:**

Major problem these days is the loss of revenue for companies due to employees switching from one particular organization to another due to unsatisfaction with their jobs or pay packages. This issue is known as churn which is creating a big gap of finances in business sector and ultimately affects the economy of a country because business sector forms a major part of economy these days. People leave their jobs due to many reasons, maybe they are suffering from pressure because they work for long periods of time every day or they are not satisfied with the salary or they feel that they cannot get promotion. Our project aims at finding out who will leave the company in near future based on their satisfaction level, last evaluation score, whether get promotion in the past 5 years and so on.

**Project description:**

Our dataset is from Kaggle “Human Resources Analytics” with nine predictors, one target variable and 14999 observation. And the figure – 1 below shows how our dataset looks like is.

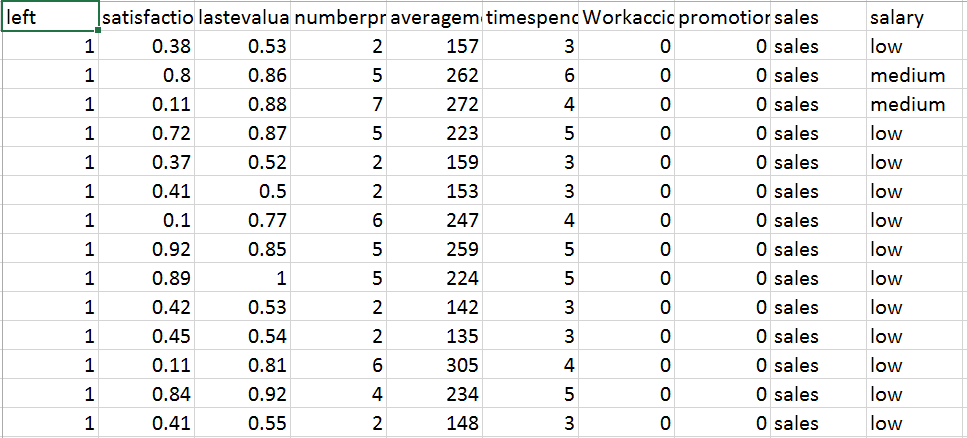


Figure – 1

We aim at building the models by different methods such as logistic regression, backwards logistic regression, Decision Tree and Random forest. We also need to figure out which model performs the best, by using Key Performance Indicators such as accuracy, misclassification error AUC value, ROC Curve and Lift Value (where applicable). One thing we need to avoid is model overfitting we don’t want to build a model with high accuracy in the training dataset but low accuracy in the real case.

**Data preparation:**

For data preparation we divided our dataset into two parts, training and validation datasets with 60% and 40% respectively. The training and validation datasets don’t have any overlap sample. For the predictor: “average monthly work hours”, it range from 100 to 300 it is very different from other predictors if we use this predictor directly it will dominate other predictors since other predictors’ value range from 0 to 1 or from 1 to 7(number of project). And in here we used min-max normalization to deal with average monthly working hours and convert the range from 100 to 300 to 0 to 1. After we normalized our dataset we worked on selecting the important features. For that we calculated the importance of each features by using Boruta feature importance package. The importance visual plots are showed in Figure - 2. In this section I am going to explain the working of Boruta Package. Its working is as follows:-

1. The First step involves adding randomness to the given data set by creating shuffled copies of all features. Shuffled copies are also known as shadow features.
2. Second step is about training a random forest algorithm on the extended data set and applying a feature importance measure. Mean Decrease Accuracy is used to evaluate the importance of each feature where higher means more important.
3. Third step is about checking whether a real feature has a higher importance than the best of its shadow features at every iteration and constantly removes features which are deemed highly unimportant.
4. At the end, the algorithm stops either when all features gets confirmed or rejected

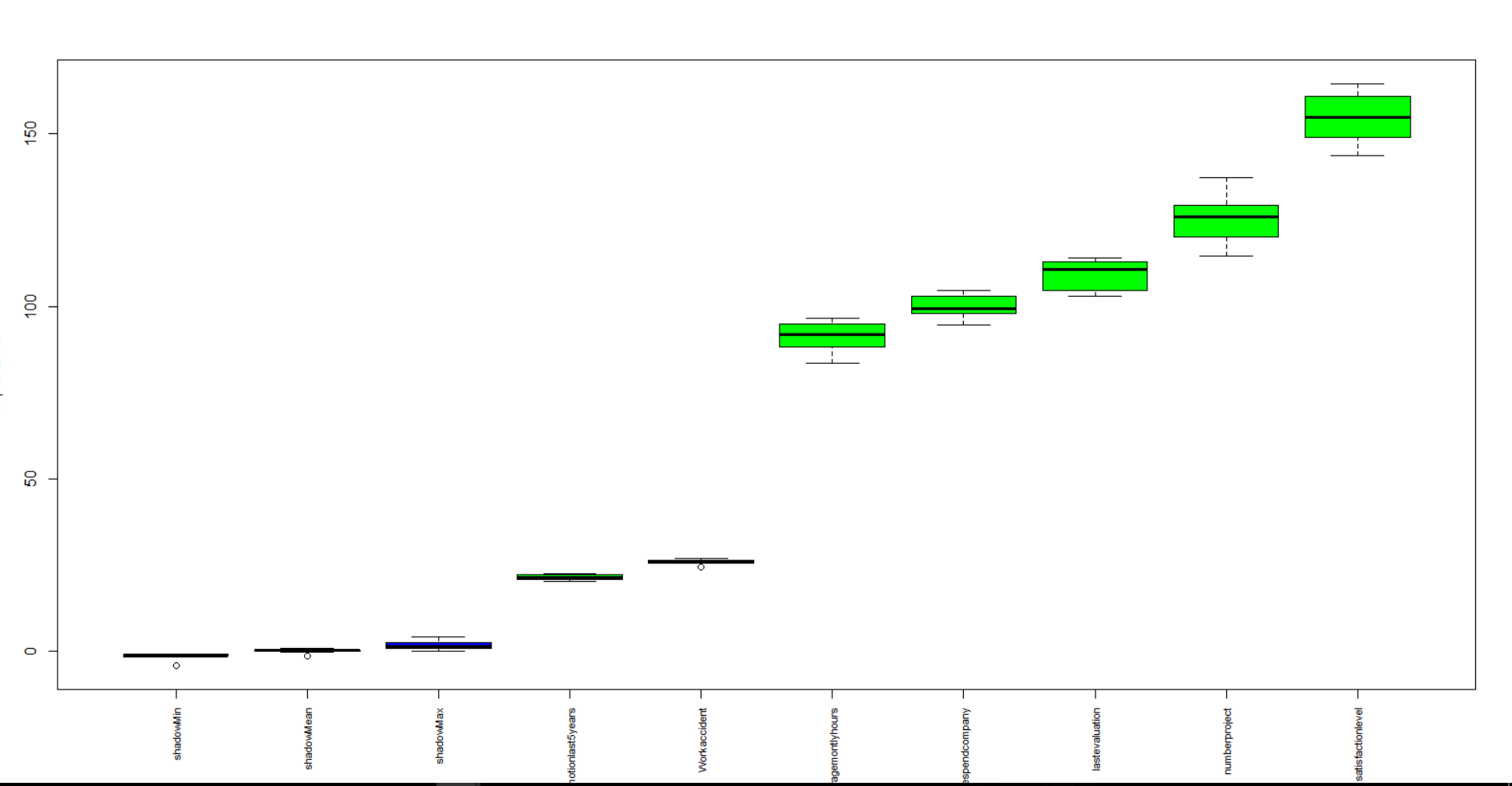


Figure - 2

The important values’ rank is as follow: satisfaction level > number of project > last evaluation > time spend in company> monthly working hours> work accident> promotion last 5 years.

Based on the importance values which we got from the plot we can choose the most important features to build the model and there are two main advantages to choose only important features to build the model.

1. Decrease the model’s complexity, the more complex model you have the higher probability you get an overfitting model.
2. Dimensionality reduction, when you only choose the important features to build the model you will also decrease the dimensions of your dataset and it will make people easier to understand your model (Suppose you have 100 coefficient for a logistic regression model and another logistic regression model with 5 coefficient the one with 5 coefficient is easier to understand.)

**Data Exploration and Visualization:**

Data exploration and visualization is one of the most important part in understanding the data. It helps in extracting some meaning out from the data set. It further helps us in analyzing the trends in the data set. We used tableau to explore data and some of the most important visualizations are as follows:

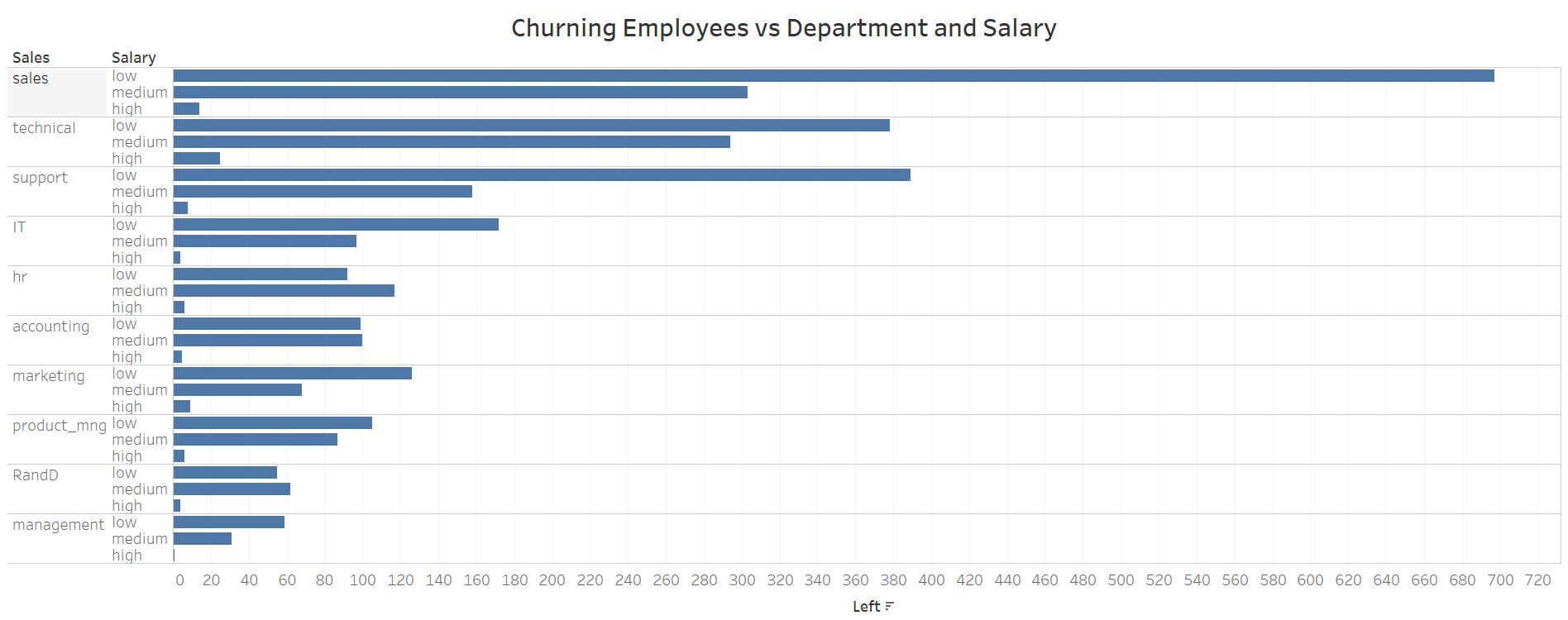
* **Satisfaction level vs Churning Employees**

If you see the following dashboard it is pretty clear that maximum employees who are leaving have low satisfaction level. Maximum leaving employees lie in range 0.05 to 0.10. Then if you see in the middle there is also an increase in employees leaving and that could be due to employees in search of new challenge. Then again at the end if you see there is an increase in employees leaving even the satisfaction level is very high and that could be due to those particular employees being very senior employees and are in search of better managerial level positions in a better company. This visualization is very clear and explaining trends of data in a very good way.



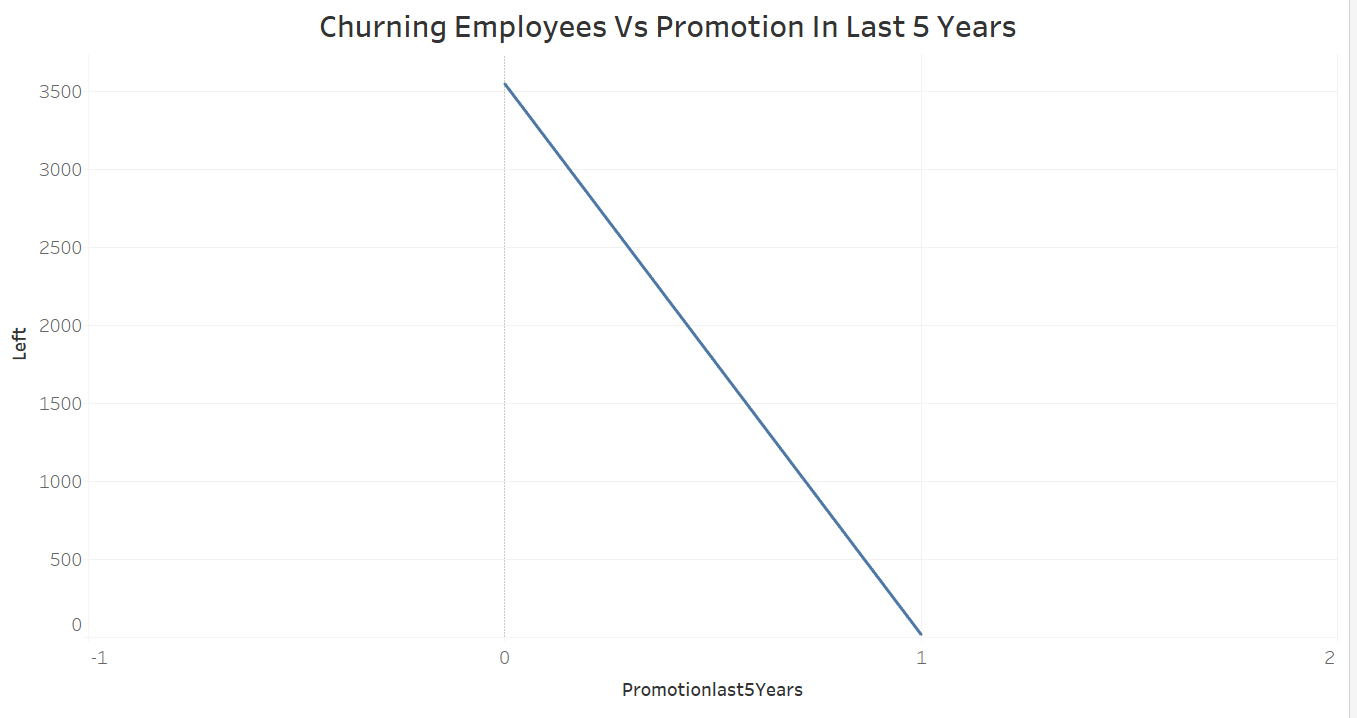
* **Churning Employees vs Department and Salary:**

The following dashboard is giving us information about the churning employees according to their department and salaries. So if you notice in every department the maximum number of churning employees are those who have the least salaries and it makes sense that less salary could be the main reason of employee churn. Organizations can stop these employees from leaving by giving them pay hikes.



* **Churning Employees vs Promotion Last 5 years:**

The following chart tells us that majority employees who leave have not been promoted in last 5 years and this churn rate decreases as the employees are getting promotions. So this could be a major reason too for employee churn. Organization can work to on this issue to solve churn problem.



**Models:**

For model application we divided the dataset into two parts which we got from Kaggle. The division ratio is 60:40. 60% is being used as training data to train the algorithm and remaining 40% is being used as validation data set to test the fitness of the algorithm.

**Resultant (dependent) and Predictors (independent) Variables:**

The variable which is need to be predicted is Left variable which in other words is the churn variable while all other variables such as satisfaction level, last evaluation, number of projects, average monthly hours, time spend company, Work accident and promotion last 5 years are used as predictors to predict churn value.

The algorithms and comparisons which we used to build our churn model are as follows:

1. Logistic Regression
2. Backward Logistic Regression
3. Decision Tree
4. Random Forest

**Logistic Regression:**

We started with a logistic regression model to predict churn ratio using all predictors mentioned above. The confusion matrix, fitness of model, accuracy, ROC Curve, AUC value, Lift value and misclassification error of data set using logistic regression is as follows:

* **Confusion Matrix:**

Following is the confusion matrix for logistic Regression.

observed

pred 0 1

0 3924 489

1. 619 968

* **Accuracy:**

Through Logistic regression model we were able to achieve accuracy of about **81.53%**. It means that model is fit more than 81% which is a very respectable result.



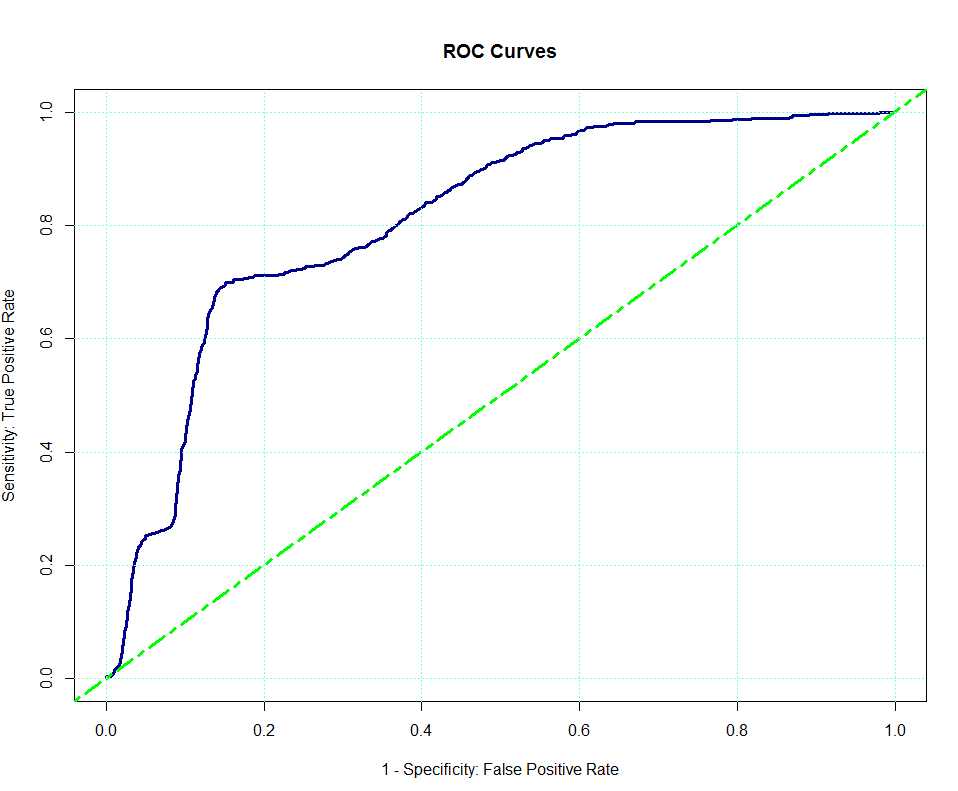
* **Misclassification Error:**

The misclassification error is around **18.46%.** It means that the model is misclassifying churn by around 18 percent which is not so bad considering this our first model.



* **ROC Curve:**

If you see here the ROC curve is giving us a very good representation as the curve is increasing and is tending to reach 1 which depicts that the model is giving us good predictions. The closer the ROC curve to 1 is the fitter the model is. In our case this condition is being satisfied.



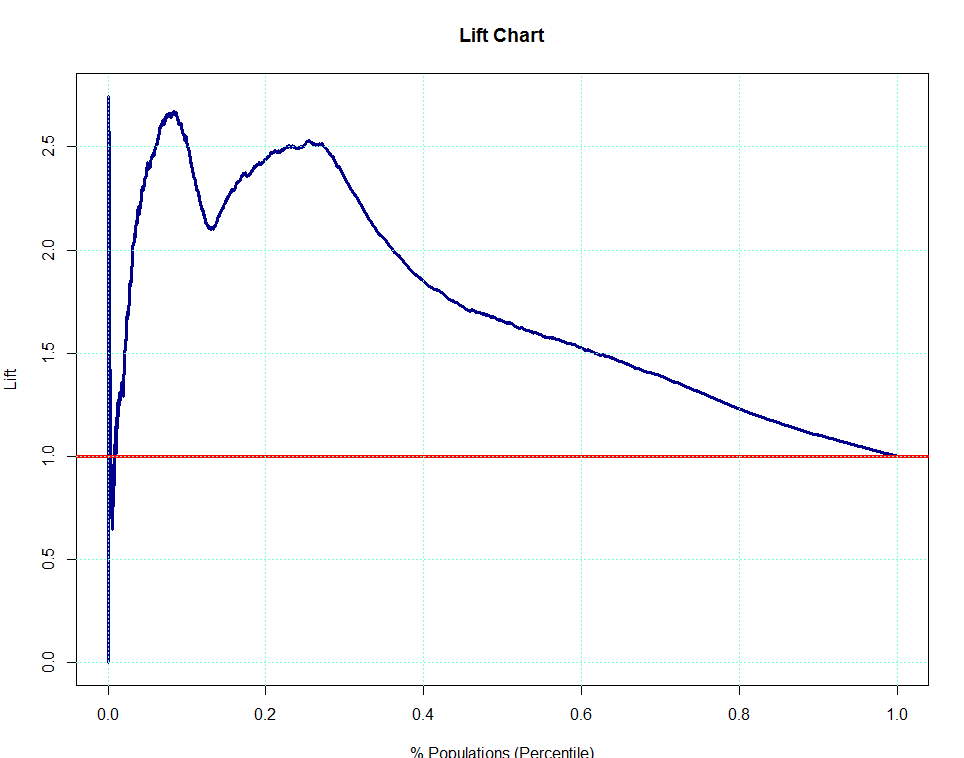
* **Area Under the Curve (AUC) Value:**

AUC value for logistic Regression Model is 0.812 which is closer to 1. It means that model is very fit. If AUC value is above 0.5 then model is considered as good model and this condition is also being satisfied in our case.



* **Lift Value:**

If you see the following lift chart it is representing a very high lift value. The higher the lift value the better the model is. This condition is also being satisfied in our case.



**Backward Logistic Regression:**

Second model which we used to predict churn ratio using all predictors was backward logistic regression model. The results of backward logistic regression are exactly the same as forward logistic regression because in forward logistic regression we are only using the most important variables and that’s exactly the case with backward regression. It uses just the most important one in backward direction. The confusion matrix, fitness of model, accuracy, ROC Curve, AUC value, Lift value and misclassification error of data set using logistic regression is as follows:

* **Confusion Matrix:**

Following is the confusion matrix for backward logistic Regression.

observed

pred 0 1

1. 3924 489

1 619 968

* **Accuracy:**

Through backward Logistic regression model we were able to achieve accuracy of about **81.53%**. It means that model is fit more than 81% which is a very respectable result.



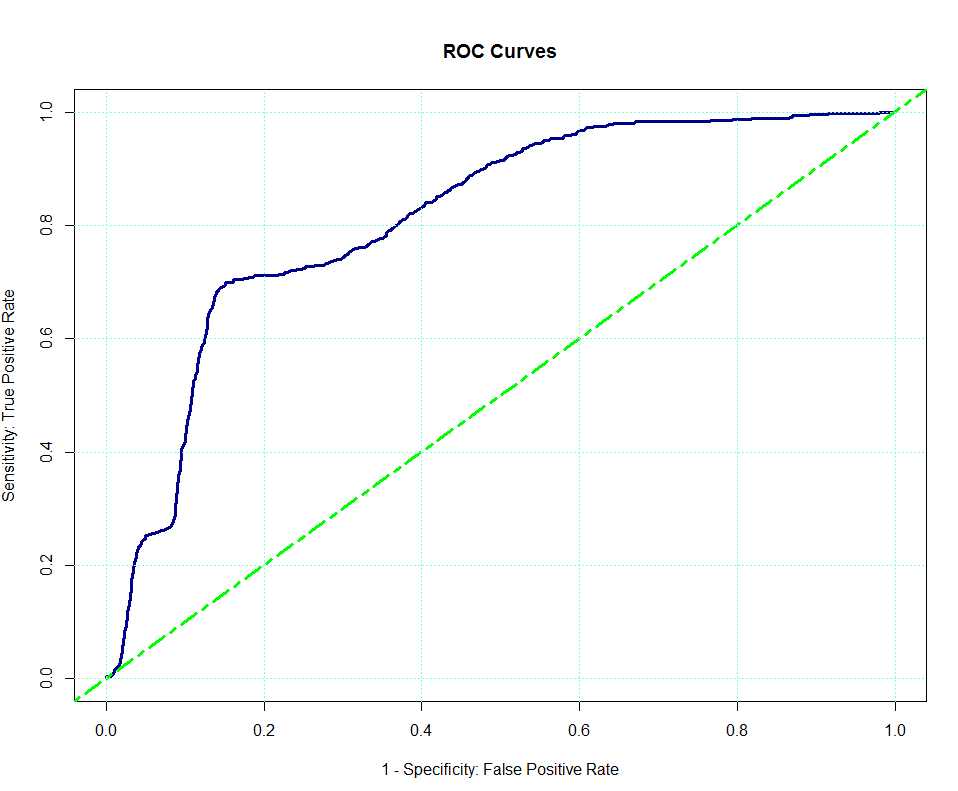
* **Misclassification Error:**

The misclassification error is around **18.46%.** It means that the model is misclassifying churn by around 18 percent which is not so bad either.



* **ROC Curve:**

If you see here the ROC curve is giving us a very good representation as the curve is increasing and is tending to reach 1 which depicts that the model is giving us good predictions. The closer the ROC curve to 1 is the fitter the model is. In our case this condition is being satisfied.



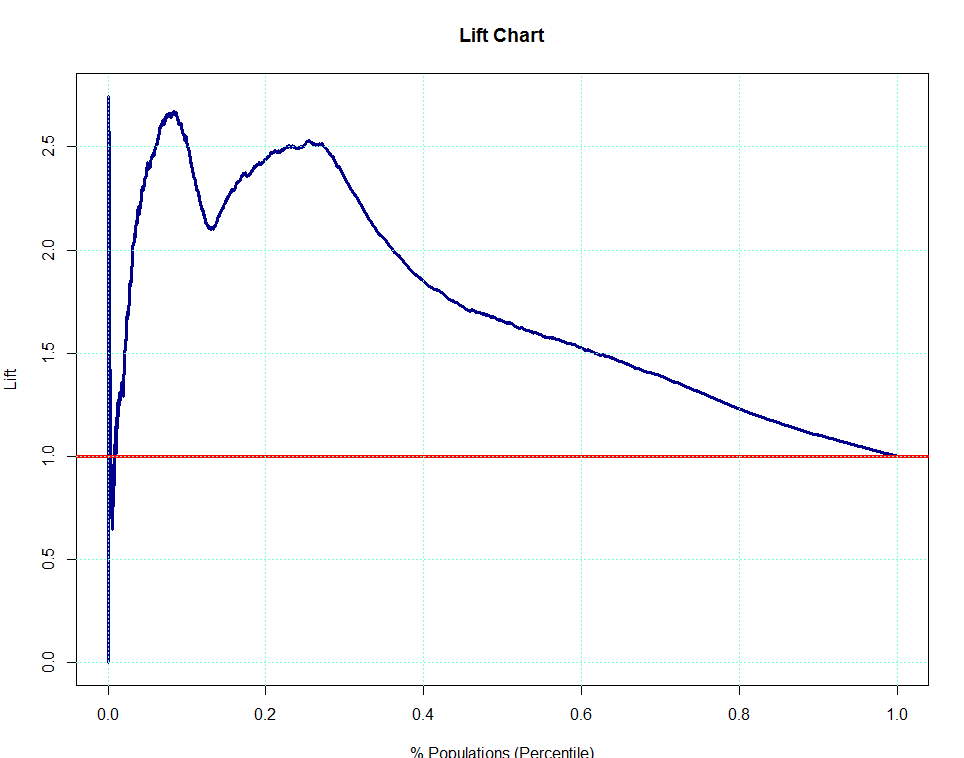
* **Area Under the Curve (AUC) Value:**

AUC value for logistic Regression Model is 0.812 which is closer to 1. It means that model is very fit. If AUC value is above 0.5 then model is considered as good model and this condition is also being satisfied in our case.



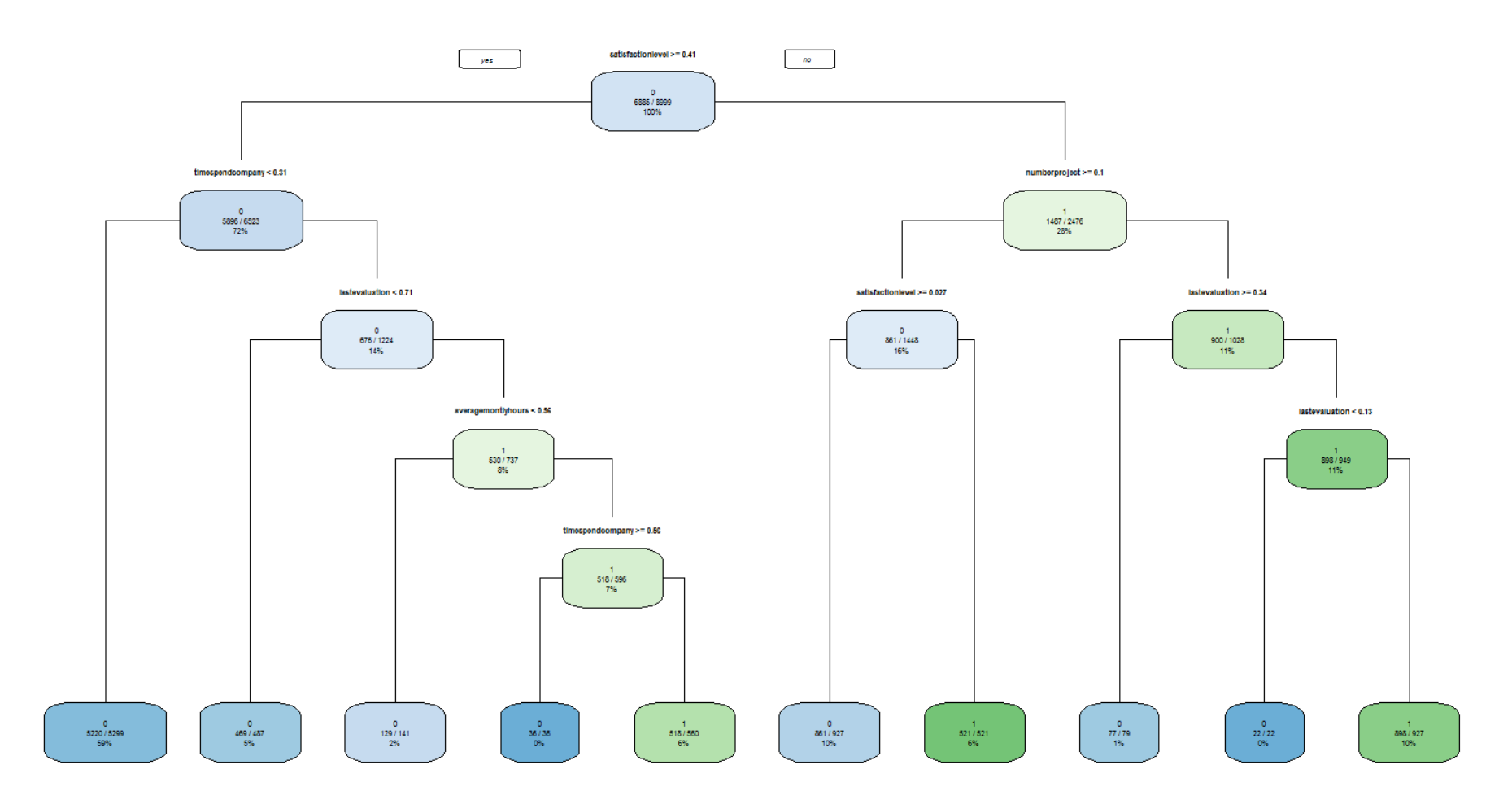
* **Lift Value:**

If you see the following lift chart it is representing a very high lift value. The higher the lift value the better the model is. This condition is also being satisfied in our case.



**Decision Trees:**

Third model which we used to predict churn ratio using all predictors was Decision Tree model. The results of Decision Tree model was much better than logistic regressions models. The Decision Tree plot, confusion matrix, fitness of model, accuracy, misclassification error and analysis of results are as follows:



If you analyze the above decision tree plot. Top node is branching on the basis of satisfaction level as being the most important variable. Employees with satisfaction level of greater than 0.41 are being branched to the left side of tree while less than 0.41 are being branched to the right side. Tree branches on the basis of most important features and continues to branch until every employee is branched into a node and at the end we get a final classification matrix which is as follows:

* **Confusion Matrix:**

Following is the confusion matrix for Decision Tree Model.

predtree

0 1

0 4490 53

1 134 1323

* **Accuracy:**

By Decision Tree Model we were able to achieve accuracy of about **96.88%**. It means that model is fit more than 96% which is a very good result and much better than both logistic regressions.



* **Misclassification Error:**

The misclassification error is around **3.11%.** It means that the model is misclassifying churn by around 3 percent which is a very reliable result.

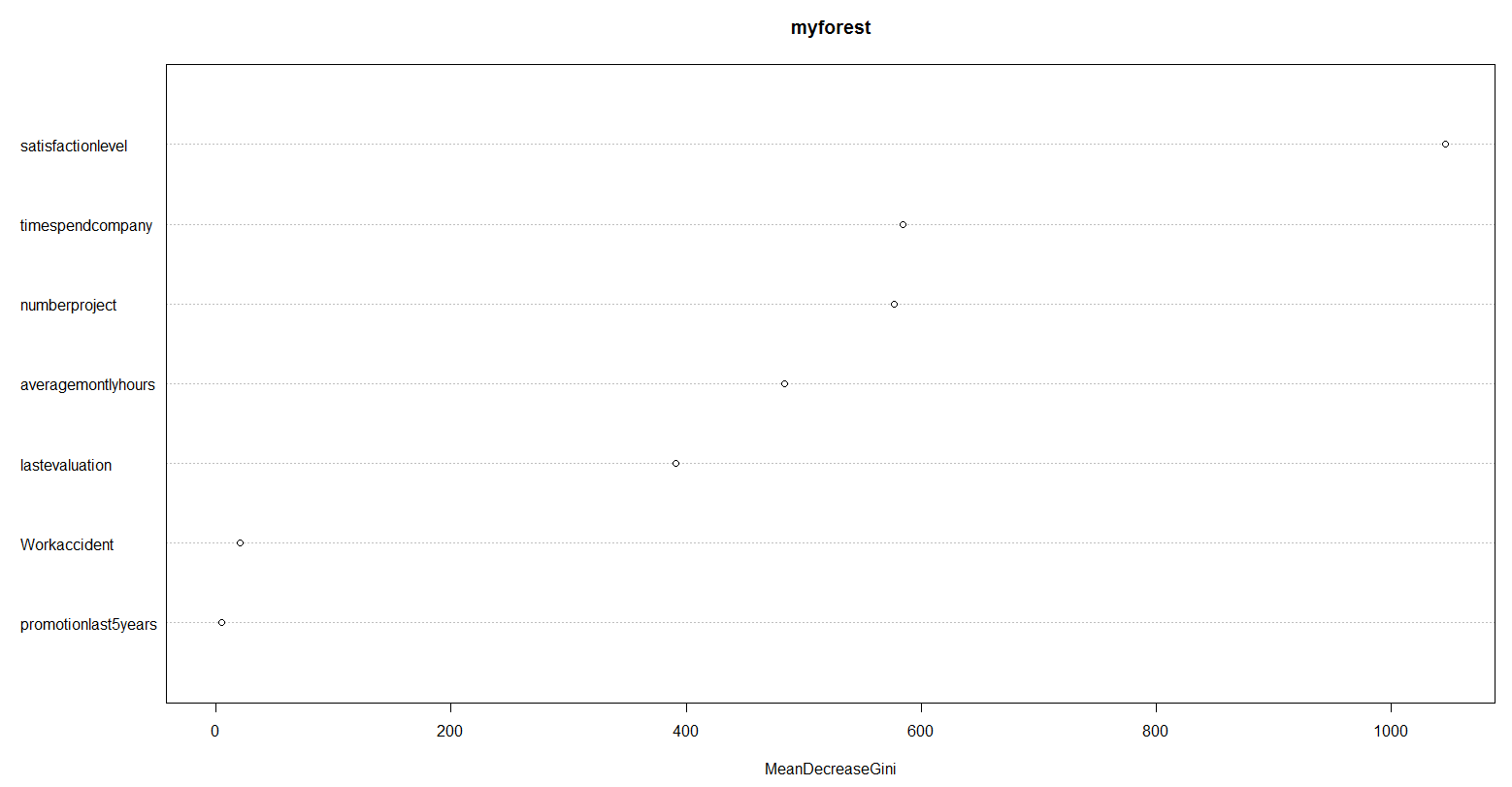


**Random Forest:**

Fourth model which we used to predict churn ratio using all predictors was Random Forest model. The results of Random Forest model was much better than logistic regressions and Decision Tree models. The feature importance plot, confusion matrix, fitness of model, accuracy and misclassification error is as follows:

* **Importance Plot:**

Following plot was generated by random forest algorithm which tells us the importance of attribute used randomly to build model. The most important attribute is satisfaction level and so on. The random number of features are taken as default value. We have not defined any value for mtry in our model.



* **Confusion Matrix:**

Following is the confusion matrix for Decision Tree Model.

obs

pred 0 1

0 4537 79

1 6 1378

* **Accuracy:**

By Decision Tree Model we were able to achieve accuracy of about **98.58%**. It means that model is fit more than 98% which is an excellent result and much better than both logistic regression and decision tree models.



* **Misclassification Error:**

The misclassification error is around **1.41%.** It means that the model is misclassifying churn by around 1.5 percent which is a very great result.



By analysis of all the algorithms stated above we came to a conclusion that **Random Forest** algorithm produced the most accurate results and can be considered as the best option to predict churn ratio because it has the maximum accuracy and the least misclassification error.

**Ways to Increase Accuracy of Logistic Regression Model:**

For logistic regression, first we used all features to train the model and the accuracy we achieved was 76%. It is very apparent by comparisons of Logistic Regression with other algorithms we used in this project that it has a very low accuracy. So we tried to figure out ways to increase accuracy of the algorithm. After plotting left (dependent variable) vs number of people we got to know that this dataset was imbalanced with 23% of people leaving. So what we did to increase the accuracy is that we applied over sampling on the original dataset and after doing that when we again checked our accuracy, it was improved. The reason an imbalanced data decreases the accuracy is that the models are normally more sensitive to the major class but less sensitive to the minor class. It means that when you use the model to predict new testing sample it will tend more to classify the sample as the major class but actually it is the minor class. This is the motivation of balancing our dataset.

Oversampling perform better than under sampling for this dataset, the reason might be when we applied under sampling in our dataset it decreased the number of majority class entries and made the number of majority class and minority class’s samples as closer as possible. In this way we lost some useful samples, which ultimately meant we were using less samples to train our model. In most cases the more the training samples the better the model we have.

Oversampling duplicated the minority class’s samples and the reason why this method worked well was that it increased the weight of the minor class and removed the bias of the model.

Then we also used five most important features to build our model again and as a result we got an increased accuracy. Another way was to adjust the threshold to a proper level. After we did all this our logistic regression model’s accuracy increased from 76% to 81.5%.

Another way to increase accuracy could be to split the data set into more features such as splitting the salary feature into three separate features to remove correlations among data and this can help in further improving the accuracy.

**Conclusion:**

So in conclusion we got to know that Random Forest algorithms works the best on this data set as it is giving us the maximum accuracy of around **98.5%** and the least misclassification error. On top of that we learnt that implementing a model is not an issue. The actual issue is to get best and optimized results from the model and for that a lot of input is required for preprocessing of the data and model optimization. Another thing which we learnt from this project was to extract a meaning out of a data set. If you are not able to extract a trend or meaning then there is no point in modeling a system. The first step to reach any solution is to understand the problem and analyze its potential solutions. The last step is modeling the best solution and interpretation of results from all that analysis.