Churn Modelling Tournament

Table of Contents

[Executive Summary 3](#_Toc430335416)

[Introduction 4](#_Toc430335417)

[What can be done? 4](#_Toc430335418)

[Understanding the churn modelling 4](#_Toc430335419)

[Details on the datasets: 4](#_Toc430335420)

[Methodology 5](#_Toc430335421)

[Understanding the Data 5](#_Toc430335422)

[Numerical Data 5](#_Toc430335423)

[Business Logic 8](#_Toc430335424)

[Categorical data 8](#_Toc430335425)

[Based on the Intuition/general understanding: 9](#_Toc430335426)

[Selection of variables 10](#_Toc430335427)

[Interpretation of mean decrease accuracy and mean decrease in Gini coefficient 12](#_Toc430335428)

[Algorithm Implementation 12](#_Toc430335429)

[Random Forest 13](#_Toc430335430)

[Steps of Implementation 13](#_Toc430335431)

[Model Formulation 14](#_Toc430335432)

[Artificial Neural Network 18](#_Toc430335433)

[Steps of Implementation 19](#_Toc430335434)

[Model Formulation 19](#_Toc430335435)

[Model Results 19](#_Toc430335436)

[Logistic Regression 23](#_Toc430335437)

[Steps of implementation 23](#_Toc430335438)

[Model Equation 23](#_Toc430335439)

[Model Summary 23](#_Toc430335440)

[Conclusion 27](#_Toc430335441)

[References 28](#_Toc430335442)

List of Tables

[Table 1: Variable selection output for Tree Algorithm 9](#_Toc406340478)

[Table 2: Gini Coefficient computation for RF 17](#_Toc406340479)

[Table 3: Confusion matrix for training dataset using ANN 18](#_Toc406340480)

[Table 4:Confusion matrix for validation dataset using ANN 18](#_Toc406340481)

[Table 5:Confusion Matrix for ANN using testing data 19](#_Toc406340482)

[Table 6:Gini Coefficient estimation using ANN 19](#_Toc406340483)

[Table 7: Coefficient Table for LLN 22](#_Toc406340484)

[Table 8:Gini Coefficient LLN 26](#_Toc406340485)

[Table 9: Comparison of Models 26](#_Toc406340486)

List of Figures

[Figure 1: ROC Curve for training dataset using RF 14](#_Toc406340502)

[Figure 2: ROC Curve for validation dataset using RF 15](#_Toc406340503)

[Figure 3: ROC Curve for Test data set using RF 15](#_Toc406340504)

[Figure 4: ROC using the training dataset for ANN 19](#_Toc406340505)

[Figure 5: ROC for testing dataset using ANN 20](#_Toc406340506)

[Figure 6: ROC for Validation data using ANN 21](#_Toc406340507)

[Figure 7: ROC Curve for testing Dataset using LNN 24](#_Toc406340508)

[Figure 8: ROC Curve for Training Dataset using LNN 25](#_Toc406340509)

[Figure 9: ROC Curve for Validation Dataset for LNN 25](#_Toc406340510)

# Executive Summary

The churn data assignment is a real world telecom industry challenge where the customer churn is a very big issue. The task was to create a model which could help in predicting the prospect customer who is going to churn out within a period of next 3-6months (the time period of the future data). The model was also tested for consistency across time that is if it is trained with data at time t, is it also valid for data at time period t+x or does it needs a retraining.

The initial shortlist of variables was done based on the data visualization techniques learned in the earlier sessions as well as based on the past work done in the field. After elimination of redundant and constant columns we were left with about 63 variables to work with. We applied three major classifiers viz. logistic regression, ANN and Random forest algorithm on this to understand how the data reacted to all the three classification algorithms.

We found that of all the algorithms viz. ANN, Logistic Regression and random forest, random forest algorithm performed the best with the best case scenario of 33% error or in other words 70% correct predictions for the given data set while in case of other algorithms the best case went to as high as 42% error (for ANN) and 40% error (for logistic regression) which implies for the given case they were not as good.

We also found that the data when missing data was imputed the results were slightly better as compared to the missing values. The imputation for numeric variables was done to be mean imputation while regression model was used in case of categorical variables for doing the imputation.

# Introduction

Customer retention and increasing customer lifetime value (CLV) of the customers in the business world today are no longer fancy concepts but something of great necessity for the survival of the business in a profitable manner in the long run. This is because of the increasing cost of acquiring customers especially when there are large no. of competitors with very diminishing gap in differentiation which each of them exhibit. This is true for any industry today especially for the telecom considering the similarities in the gamut of services which each of the operator provides, the traditional nature of industry being debt ridden and not to mention of the ever increasing regulations.

Customer churn is one of the biggest problems being faced by companies and a lot of effort is put in for retaining the customers. However most of the times, efforts for retaining a customer may actually not be required because they are anyways not going to leave. Hence, it is very important for the players in this type of industry to try to develop a model which could help them predict in advance about which customers are the “leaving customers” as compared to which are the loyal ones. This loss of an existing customer to a competitor is called churn and telecom industries are constantly worried about this high churn rate.

These days the problem of churn has become more difficult for the telecom operators to manage because introduction of mobile no. portability has removed even the last of the obstacles for switching.

## What can be done?

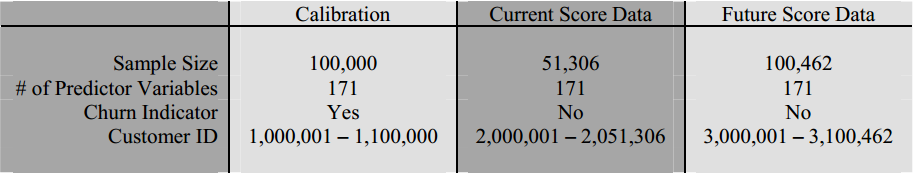
The answer to this is predictive churn management. Trying to predict beforehand which customer would be leaving and how could he be retained profitably. The process would be trying to sort the high risk customer profitably and then try to retain those using incentives like price breaks and by offering few special services to these few. However, the scope of this assignment would be restricted to predicting the churn (the customers who would be leaving) as against the customers who would not be and the determination of CLV/profitability would not be covered.

# Understanding the churn modelling

The churn modelling was used to identify and answer the question of who would be the next set of probable customers leaving. In fact the model was designed using the calibration data set given to us.

## Details on the datasets:

The data are organized into three data files: Calibration, Current Score Data, and Future Score Data.



The calibration data set contains the actual data on churn of customer a 1 indicated a churn happening a 0 indicated no churn from customers. The current score data has been taken from a time period which is same around the calibration data but no details on the churn is provided. The future score data is the 3rd set of data provided to do prediction for customers at a time which is further away from the time set.

The idea is to understand the validity of the model generated across time. I.e. if a model is designed using a data at a time frame T is it also valid for a time T+ t or does it needs a revision.

# Methodology

The methodology for the churn data assignment was as following steps:

1. Data visualization and understanding of the data fields, trying to select what is relevant using the research papers published on data churning exercises and based on common sense.
2. Few of the data predictors were also filtered off like: csa etc. which had the no. of categories much larger than what R could handle (R algorithms which were used for the process viz. Logistic Regression, Trees, Random Forest and ANN)
3. Data imputation was done using the mi package in R for the important predictors which were shortlisted in the option 2. The code snippet for the same without mi library is as below:

*churn\_master<-read.csv("calibration.csv")*

*for (i in which(sapply(churn\_master, is.numeric))) {*

*churn\_master[is.na(churn\_master[, i]), i] <- mean(churn\_master[, i], na.rm = TRUE)*

*}*

1. Each of this model was ran to find out what could be the best fit logical model using the prescribed test of lift chart and Decile implementation.
2. Performance summary table was prepared for each of the algorithms and conclusive model was achieved though refining and re-refining.
3. Considering the complexity of the code to be written for analysis, most of the analysis was done using the **rattle library of r** which provided extensive abilities to create run and test the datasets

The details for each of the algorithm used and its nitty-gritties is listed in the algorithm implementation section.

# Understanding the Data

The dataset consisted of about 172 predictors and one classification category variable (viz. churn). Of the 172 variables, about 124 variables were numerical while others were categorical.

## Numerical Data

The numerical variables were as below:

|  |  |  |
| --- | --- | --- |
| ***Interval Variables*** | ***Explanation*** | ***% Missing*** |
| [ADJMOU](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A149:E150) | Billing adjusted total minutes of use over the life of the customer | 0.000% |
| [ADJQTY](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A152:E154) | Billing adjusted total number of calls over the life of the customer | 0.000% |
| ADJREV | Billing adjusted total revenue over the life of the customer | 0.000% |
| [ATTEMPT\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A149:E150) | Mean number of attempted calls | 0.000% |
| ATTEMPT\_RANGE | Range of number of attempted calls | 0.000% |
| AVG3MOU | Average monthly minutes of use over the previous three months | 0.000% |
| [AVG3QTY](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A152:E154) | Average monthly number of calls over the previous three months | 0.000% |
| AVG3REV | Average monthly revenue over the previous three months | 0.000% |
| AVG6MOU | Average monthly minutes of use over the previous six months | 2.839% |
| [AVG6QTY](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A149:E150) | Average monthly number of calls over the previous six months | 2.839% |
| AVG6REV | Average monthly revenue over the previous six months | 2.839% |
| AVGMOU | Average monthly minutes of use over the life of the customer | 0.000% |
| [AVGQTY](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A152:E154) | Average monthly number of calls over the life of the customer | 0.000% |
| AVGREV | Average monthly revenue over the life of the customer | 0.000% |
| BLCK\_DAT\_MEAN | Mean number of blocked (failed) data calls | 0.000% |
| BLCK\_DAT\_RANGE | Range of number of blocked (failed) data calls | 0.000% |
| BLCK\_VCE\_MEAN | Mean number of blocked (failed) voice calls | 0.000% |
| BLCK\_VCE\_RANGE | Range of number of blocked (failed) voice calls | 0.000% |
| CALLFWDV\_MEAN | Mean number of call forwarding calls | 0.000% |
| CALLFWDV\_RANGE | Range of number of call forwarding calls | 0.000% |
| CALLWAIT\_MEAN | Mean number of call waiting calls | 0.000% |
| CALLWAIT\_RANGE | Range of number of call waiting calls | 0.000% |
| [CC\_MOU\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A156:E157) | Mean unrounded minutes of use of customer care (see CUSTCARE\_MEAN) calls | 0.000% |
| CC\_MOU\_RANGE | Range of unrounded minutes of use of customer care calls | 0.000% |
| [CCRNDMOU\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A156:E157) | Mean rounded minutes of use of customer care calls | 0.000% |
| CCRNDMOU\_RANGE | Range of rounded minutes of use of customer care calls | 0.000% |
| CHANGE\_MOU | Percentage change in monthly minutes of use vs previous three month average | 0.891% |
| [CHANGE\_REV](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A156:E157) | Percentage change in monthly revenue vs previous three month average | 0.891% |
| COMP\_DAT\_MEAN | Mean number of completed data calls | 0.000% |
| COMP\_DAT\_RANGE | Range of number of completed data calls | 0.000% |
| [COMP\_VCE\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A159:E161) | Mean number of completed voice calls | 0.000% |
| COMP\_VCE\_RANGE | Range of number of completed voice calls | 0.000% |
| [COMPLETE\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A163:E164) | Mean number of completed calls | 0.000% |
| [COMPLETE\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A156:E157) | Range of number of completed calls | 0.000% |
| [CUSTCARE\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A163:E164) | Mean number of customer care calls | 0.000% |
| [CUSTCARE\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A159:E161) | Range of number of customer care calls | 0.000% |
| [DA\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A166:E167) | Mean number of directory assisted calls | 0.357% |
| [DA\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A163:E164) | Range of number of directory assisted calls | 0.357% |
| [DATOVR\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A169:E171) | Mean revenue of data overage | 0.357% |
| [DATOVR\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!G50) | Range of revenue of data overage | 0.357% |
| [DROP\_BLK\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A169:E171) | Mean number of dropped or blocked calls | 0.000% |
| [DROP\_BLK\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A166:E167) | Range of number of dropped or blocked calls | 0.000% |
| [DROP\_DAT\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!G50) | Mean number of dropped (failed) data calls | 0.000% |
| [DROP\_DAT\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A169:E171) | Range of number of dropped (failed) data calls | 0.000% |
| DROP\_VCE\_MEAN | Mean number of dropped (failed) voice calls | 0.000% |
| DROP\_VCE\_RANGE | Range of number of dropped (failed) voice calls | 0.000% |
| EQPDAYS | Number of days (age) of current equipment | 0.001% |
| [INONEMIN\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A173:E174) | Mean number of inbound calls less than one minute | 0.000% |
| [INONEMIN\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!G50) | Range of number of inbound calls less than one minute | 0.000% |
| [IWYLIS\_VCE\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A169:E171) | Mean number of inbound wireless to wireless voice calls | 0.000% |
| IWYLIS\_VCE\_RANGE | Range of number of inbound wireless to wireless voice calls | 0.000% |
| MONTHS | Total number of months in service | 0.000% |
| [MOU\_CDAT\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A173:E174) | Mean unrounded minutes of use of completed data calls | 0.000% |
| MOU\_CDAT\_RANGE | Range of unrounded minutes of use of completed data calls | 0.000% |
| MOU\_CVCE\_MEAN | Mean unrounded minutes of use of completed voice calls | 0.000% |
| MOU\_CVCE\_RANGE | Range of unrounded minutes of use of completed voice calls | 0.000% |
| MOU\_MEAN | Mean number of monthly minutes of use | 0.357% |
| [MOU\_OPKD\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A201) | Mean unrounded minutes of use of off-peak data calls | 0.000% |
| [MOU\_OPKD\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A173:E174) | Range of unrounded minutes of use of off-peak data calls | 0.000% |
| MOU\_OPKV\_MEAN | Mean unrounded minutes of use of off-peak voice calls | 0.000% |
| MOU\_OPKV\_RANGE | Range of unrounded minutes of use of off-peak voice calls | 0.000% |
| MOU\_PEAD\_MEAN | Mean unrounded minutes of use of peak data calls | 0.000% |
| MOU\_PEAD\_RANGE | Range of unrounded minutes of use of peak data calls | 0.000% |
| MOU\_PEAV\_MEAN | Mean unrounded minutes of use of peak voice calls | 0.000% |
| MOU\_PEAV\_RANGE | Range of unrounded minutes of use of peak voice calls | 0.000% |
| MOU\_RANGE | Range of number of minutes of use | 0.357% |
| MOU\_RVCE\_MEAN | Mean unrounded minutes of use of received voice calls | 0.000% |
| MOU\_RVCE\_RANGE | Range of unrounded minutes of use of received voice calls | 0.000% |
| MOUIWYLISV\_MEAN | Mean unrounded minutes of use of inbound wireless to wireless voice calls | 0.000% |
| MOUIWYLISV\_RANGE | Range of unrounded minutes of use of inbound wireless to wireless voice calls | 0.000% |
| MOUOWYLISV\_MEAN | Mean unrounded minutes of use of outbound wireless to wireless voice calls | 0.000% |
| MOUOWYLISV\_RANGE | Range of unrounded minutes of use of outbound wireless to wireless voice calls | 0.000% |
| OWYLIS\_VCE\_MEAN | Mean number of outbound wireless to wireless voice calls | 0.000% |
| OWYLIS\_VCE\_RANGE | Range of number of outbound wireless to wireless voice calls | 0.000% |
| OPK\_DAT\_MEAN | Mean number of off-peak data calls | 0.000% |
| OPK\_DAT\_RANGE | Range of number of off-peak data calls | 0.000% |
| OPK\_VCE\_MEAN | Mean number of off-peak voice calls | 0.000% |
| OPK\_VCE\_RANGE | Range of number of off-peak voice calls | 0.000% |
| [OVRMOU\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A176:E178) | Mean overage minutes of use | 0.357% |
| OVRMOU\_RANGE | Range of overage minutes of use | 0.357% |
| [OVRREV\_MEAN](file:///C:\\Users\\Hardik%20Shah\\Desktop\\PFA\\Churn%20Assignment\\data_documentation.xls" \l "RANGE!A180:E181) | Mean overage revenue | 0.357% |
| OVRREV\_RANGE | Range of overage revenue | 0.357% |
| PEAK\_DAT\_MEAN | Mean number of peak data calls | 0.000% |
| [PEAK\_DAT\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A189:E190) | Range of number of peak data calls | 0.000% |
| PEAK\_VCE\_MEAN | Mean number of inbound and outbound peak voice calls | 0.000% |
| [PEAK\_VCE\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A183:E184) | Range of number of inbound and outbound peak voice calls | 0.000% |
| PLCD\_DAT\_MEAN | Mean number of attempted data calls placed | 0.000% |
| [PLCD\_DAT\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A186:E187) | Range of number of attempted data calls placed | 0.000% |
| [PLCD\_VCE\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A192:E193) | Mean number of attempted voice calls placed | 0.000% |
| [PLCD\_VCE\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A189:E190) | Range of number of attempted voice calls placed | 0.000% |
| [RECV\_SMS\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A192:E193) | Mean number of received SMS calls | 0.000% |
| RECV\_SMS\_RANGE | Range of number of received SMS calls | 0.000% |
| [RECV\_VCE\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A195:E196) | Mean number of received voice calls | 0.000% |
| [RECV\_VCE\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A192:E193) | Range of number of received voice calls | 0.000% |
| [RETDAYS](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A195:E196) | Number of days since last retention call | 96.017% |
| REV\_MEAN | Mean monthly revenue (charge amount) | 0.357% |
| REV\_RANGE | Range of revenue (charge amount) | 0.357% |
| [RMCALLS](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A195:E196) | Total number of roaming calls | 85.777% |
| RMMOU | Total minutes of use of roaming calls | 85.777% |
| [RMREV](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A192:E193) | Total revenue of roaming calls | 85.777% |
| ROAM\_MEAN | Mean number of roaming calls | 0.357% |
| ROAM\_RANGE | Range of number of roaming calls | 0.357% |
| THREEWAY\_MEAN | Mean number of three way calls | 0.000% |
| [THREEWAY\_RANGE](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A195:E196) | Range of number of three way calls | 0.000% |
| [TOTCALLS](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A198:E199) | Total number of calls over the life of the customer | 0.000% |
| TOTMOU | Total minutes of use over the life of the customer | 0.000% |
| [TOTMRC\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A198:E199) | Mean total monthly recurring charge | 0.357% |
| TOTMRC\_RANGE | Range of total monthly recurring charge | 0.357% |
| TOTREV | Total revenue | 0.000% |
| [UNAN\_DAT\_MEAN](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A198:E199) | Mean number of unanswered data calls | 0.000% |
| UNAN\_DAT\_RANGE | Range of number of unanswered data calls | 0.000% |
| UNAN\_VCE\_MEAN | Mean number of unanswered voice calls | 0.000% |
| UNAN\_VCE\_RANGE | Range of number of unanswered voice calls | 0.000% |
| VCEOVR\_MEAN | Mean revenue of voice overage | 0.357% |
| VCEOVR\_RANGE | Range of revenue of voice overage | 0.357% |

Of all the given data set for numerical variables, except for the following data variables, most of the data sets had sufficient data or in other words low missing data:

|  |  |  |
| --- | --- | --- |
| [RETDAYS](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A195:E196) | Number of days since last retention call | 96.017% |
| [RMCALLS](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A195:E196) | Total number of roaming calls | 85.777% |
| RMMOU | Total minutes of use of roaming calls | 85.777% |
| [RMREV](file:///C:\Users\Hardik%20Shah\Desktop\PFA\Churn%20Assignment\data_documentation.xls#RANGE!A192:E193) | Total revenue of roaming calls | 85.777% |

Business Logic: Of the above four predictors, Retdays suggest that the company felt that the specific customer was about to leave a call was made but we would not require this data because we are making a new prediction model, also the remaining three data points are on roaming data and till the time, the company plans to introduce special offers for roaming customers, but in that case the requirements of the model would be quite different and out of scope for this analysis.

Further analysis on this variables would be done in the data analysis section.

## Categorical data

Coming to the categorical variables, of the 57 variables about 27 variables have data missing as large as 38-85%. With this humongous loss of data for these variables, it would not be advisable to keep them in the model as prediction power of the model would be highly restricted and missing data imputation for such kind of variables would be again difficult because of the inherent limitations of data. Hence for the initial analysis we would be restricting ourselves from using these predictors. This does not mean that these predictors are not significant for defining the churning but the case is that the inclusion of this will not make the model actionable.

These predictors are:

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Description** | **Missing Value** |
| MAILFLAG | DMA: Do not mail flag | 0.98523 |
| SOLFLAG | Infobase no phone solicitation flag | 0.98039 |
| CRTCOUNT | Adjustments made to credit rating of individual | 0.965 |
| TOT\_ACPT | Total offers accepted from retention team | 0.96017 |
| TOT\_RET | Total calls into retention team | 0.96017 |
| REF\_QTY | Total number of referrals | 0.95545 |
| WRKWOMAN | Working woman in household | 0.87491 |
| EDUC1 | Education of first household member | 0.86478 |
| PCOWNER | PC owner dummy variable | 0.81534 |
| DIV\_TYPE | Division type code | 0.81459 |
| OCCU1 | Occupation of first household member | 0.73353 |
| PROPTYPE | Property type detail | 0.71788 |
| CARTYPE | Dominant vehicle lifestyle | 0.68412 |
| CHILDREN | Children present in household | 0.65928 |
| MAILORDR | Mail order buyer | 0.64363 |
| MAILRESP | Mail responder | 0.62889 |
| LAST\_SWAP | Date of last phone swap | 0.58 |
| PRE\_HND\_PRICE | Previous handset price | 0.57515 |
| NUMBCARS | Known number of vehicles | 0.49366 |
| DWLLSIZE | Dwelling size | 0.38308 |
| HHSTATIN | Premier household status indicator | 0.37923 |
| OWNRENT | Home owner/renter status | 0.33706 |
| DWLLTYPE | Dwelling unit type | 0.31909 |
| LOR | Length of residence | 0.3019 |
| INCOME | Estimated income | 0.25436 |
| ADULTS | Number of adults in household | 0.23019 |
| INFOBASE | InfoBase match | 0.22079 |

Of the remaining predictors, we would be using the below two models:

1. Based on the domain expert views (from the research papers etc. which were studied for the effort and based on our past experience with the telecom industry) we would be dropping of few of the variables
2. Or allow the tree method to do this

### Based on the Intuition/general understanding:

The following variables highlighted were used while the one in yellow were ignored for the purpose of study.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Explanation** | **% Missing Value** |
| AGE1 | Age of first household member | 0.01732 |
| AGE2 | Age of second household member | 0.01732 |
| CAR\_BUY | New or used car buyer | 0.01732 |
| CREDITCD | Credit card indicator | 0.01732 |
| ETHNIC | Ethnicity roll-up code | 0.01732 |
| FORGNTVL | Foreign travel dummy variable | 0.01732 |
| KID0\_2 | Child 0 - 2 years of age in household | 0.01732 |
| KID3\_5 | Child 3 - 5 years of age in household | 0.01732 |
| KID6\_10 | Child 6 - 10 years of age in household | 0.01732 |
| KID11\_15 | Child 11 - 15 years of age in household | 0.01732 |
| KID16\_17 | Child 16 - 17 years of age in household | 0.01732 |
| MARITAL | Marital status | 0.01732 |
| MTRCYCLE | Motorcycle indicator | 0.01732 |
| RV | RV indicator | 0.01732 |
| TRUCK | Truck indicator | 0.01732 |
| HND\_PRICE | Current handset price | 0.00847 |
| AREA | Geographic area | 0.0004 |
| DUALBAND | Dualband | 0.00001 |
| HND\_WEBCAP | Handset web capability | 0.00001 |
| MODELS | Number of models issued | 0.00001 |
| PHONES | Number of handsets issued | 0.00001 |
| REFURB\_NEW | Handset: refurbished or new | 0.00001 |
| ACTVSUBS | Number of active subscribers in household | 0 |
| ASL\_FLAG | Account spending limit | 0 |
| CHURN | Instance of churn between 31-60 days after observation date | 0 |
| CRCLSCOD | Credit class code | 0 |
| CSA | Communications local service area | 0 |
| CUSTOMER\_ID | Unique tournament specific customer ID for scoring purposes | 0 |
| NEW\_CELL | New cell phone user | 0 |
| UNIQSUBS | Number of unique subscribers in the household | 0 |

# Selection of variables

As seen from the table, we have both categorical and numerical variables which are important from the point of study and hence, in order to shortlist from the given set of variables we would use a simple tree algorithm. From that we would try to apply more complex algorithm to see if we can make a better prediction model for the same.

The below is the table (importance table) generated from using a tree algorithm for **“variable selection”. The logic for selection of the variables and interpretation of mean decrease in efficiency and mean decrease in Gini coefficient is after the table.**

Table : Variable selection output for Tree Algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **0** | **1** | **Mean decrease in accuracy** | **Mean decrease in Gini coefficient** |
| IMN\_eqpdays | 4.39 | 3.46 | 5.84 | 5.35 |
| TFC\_hnd\_price | 3.73 | 2.14 | 5.83 | 5.58 |
| IMN\_change\_mou | 0.81 | 2.65 | 5.03 | 4.19 |
| IMN\_months | 3.53 | 1.67 | 4.96 | 3.59 |
| IMN\_mou\_Mean | 0.73 | 1.26 | 4.15 | 3.55 |
| IMN\_ovrrev\_Mean | 1.97 | -0.24 | 3.98 | 1.72 |
| IMN\_avgmou | 2.28 | -1.42 | 3.85 | 2.74 |
| IMN\_ovrmou\_Mean | -0.43 | 1.38 | 3.62 | 1.65 |
| IMN\_opk\_vce\_Mean | -0.34 | 1.38 | 3.62 | 2.57 |
| IMN\_avg3mou | 1.03 | 0.49 | 3.59 | 3.28 |
| IMN\_age1 | 1.3 | 0.65 | 3.44 | 2.25 |
| area | 3.38 | -0.72 | 3.4 | 16.59 |
| IMN\_adjrev | 0.97 | 0.57 | 3.38 | 3.23 |
| IMN\_complete\_Range | 1.35 | -0.4 | 3.33 | 1.74 |
| IMN\_custcare\_Mean | 2.17 | -1.25 | 3.3 | 1.02 |
| IMN\_adjqty | 2.2 | -1.75 | 3.28 | 2.69 |
| IMN\_totmou | 1.47 | -0.9 | 3.16 | 2.98 |
| IMN\_inonemin\_Range | 1.63 | -1.02 | 3.14 | 2.12 |
| IMN\_totmrc\_Mean | 1.57 | 0.15 | 3.09 | 2.61 |
| IMN\_avgqty | 1.95 | -1.53 | 3.09 | 2.72 |
| IMN\_attempt\_Mean | 0.91 | -0.17 | 3.06 | 2.13 |
| IMN\_mou\_rvce\_Mean | -1.68 | 2.39 | 3.01 | 2.58 |
| IMN\_mou\_opkv\_Mean | -1.38 | 2.21 | 3 | 2.03 |
| IMN\_mou\_cvce\_Mean | -0.33 | 1.32 | 2.99 | 3 |
| IMN\_ovrmou\_Range | 1.3 | -0.57 | 2.91 | 1.98 |
| IMN\_avgrev | 2.09 | -1.48 | 2.91 | 3.23 |
| IMN\_change\_rev | -0.2 | 1.29 | 2.86 | 3.45 |
| IMN\_totrev | 0.19 | 1.54 | 2.86 | 3.24 |
| IMN\_mou\_Range | 0.26 | 0.63 | 2.83 | 2.77 |
| IMN\_peak\_vce\_Mean | -0.43 | 1.11 | 2.82 | 2.31 |
| IMN\_mou\_peav\_Range | -0.08 | 1.14 | 2.78 | 2.71 |
| IMN\_avg3qty | -0.53 | 1.32 | 2.76 | 2.5 |
| IMN\_blck\_vce\_Mean | -1.41 | 2.05 | 2.68 | 1.76 |
| IMN\_retdays | 2.3 | 2.8 | 2.68 | 0.43 |
| IMN\_comp\_vce\_Mean | -0.29 | 0.99 | 2.66 | 1.85 |
| IMN\_ovrrev\_Range | 1.22 | -0.31 | 2.63 | 1.83 |
| IMN\_da\_Mean | -1.62 | 2.4 | 2.62 | 1.17 |
| IMN\_comp\_vce\_Range | -0.85 | 1.52 | 2.55 | 2.17 |
| IMN\_mouowylisv\_Mean | 1.03 | -0.46 | 2.53 | 2.28 |
| IMN\_mou\_rvce\_Range | -1.29 | 1.83 | 2.53 | 2.37 |
| IMN\_plcd\_vce\_Range | 0.14 | 0.63 | 2.51 | 2.25 |
| IMN\_recv\_vce\_Mean | 0.4 | 0.37 | 2.5 | 2.4 |
| IMN\_vceovr\_Range | 0.39 | 0.38 | 2.48 | 1.66 |
| IMN\_callwait\_Range | -1.01 | 1.32 | 2.48 | 0.89 |
| IMN\_avg6qty | -0.33 | 0.72 | 2.47 | 2.47 |
| IMN\_adjmou | 0.73 | -0.13 | 2.44 | 2.91 |
| IMN\_owylis\_vce\_Mean | -0.74 | 1.25 | 2.35 | 2.08 |
| IMN\_mou\_peav\_Mean | 0.45 | 0.21 | 2.32 | 2.61 |
| IMN\_attempt\_Range | 1.4 | -0.74 | 2.31 | 2.15 |
| IMN\_lor | 2.04 | -1.73 | 2.3 | 2.08 |
| IMN\_vceovr\_Mean | 0.61 | 0.43 | 2.27 | 1.45 |
| IMN\_complete\_Mean | -2.01 | 2.66 | 2.26 | 2.45 |
| IMN\_cc\_mou\_Mean | 1.16 | -0.38 | 2.25 | 1.6 |
| IMN\_inonemin\_Mean | 0.32 | 0.46 | 2.23 | 2.39 |
| IMN\_hnd\_price | 1.33 | 0.74 | 2.22 | 1.58 |
| IMN\_recv\_vce\_Range | 1.53 | -0.75 | 2.21 | 2.45 |
| IMN\_mouiwylisv\_Range | -1.06 | 1.54 | 2.21 | 2.21 |
| IMN\_uniqsubs | 2.33 | -1.41 | 2.21 | 0.81 |
| asl\_flag | 1.3 | 0.86 | 2.08 | 0.6 |
| IMN\_totcalls | -0.05 | 0.59 | 2.08 | 2.97 |
| IMN\_drop\_blk\_Range | 0.46 | 0.1 | 2.07 | 1.92 |
| IMN\_tot\_ret | 2 | 1.7 | 2.07 | 0.3 |
| IMN\_mou\_cvce\_Range | 1.22 | -0.59 | 2.01 | 2.67 |

From the above table variables were selected based on the value of mean decrease in accuracy and the mean decrease in Gini coefficient.

## Interpretation of mean decrease accuracy and mean decrease in Gini coefficient

The more the accuracy of the random forest decreases due to the addition of a single variable, the more important the variable is deemed, and therefore variables with a large mean decrease in accuracy are more important for classification of the data.

The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. Each time a particular variable is used to split a node, the Gini coefficient for the child nodes are calculated and compared to that of the original node. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). The changes in Gini are summed for each variable and normalized at the end of the calculation. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient.

For the purpose of this research, based on our study we are putting a cutoff of >=2 for mean decrease in accuracy. From the assumed cutoffs we are getting the above 61 variables to be of significant for further analysis. ***We would try to apply random forest algorithm, Logistic regression model and ANN on these predictors and see which model performs the best in terms of classification and further move with the same.***

# Algorithm Implementation

As discussed for the variables shortlisted above, we would be using 3 algorithms as below:

* Random Forest Algorithm
* Logistic Regression
* Artificial Neural Network

However, before going into the algorithms, we would try to analyze the data and minimize the missing data for numeric variables using the **mean imputation method. For few of the categorical variables used, we would using the mi package function and develop a linear regression model to be plugged into the mi.categorical for imputation.**

## Random Forest

They are the ensemble learning method that operate by constructing a multitude of trees at training time with each tree giving a vote for the specific class of classification. The accuracy of the algorithm does not depend on individual tree but as the collection of trees as a whole. The process of formulation of a tree by the algorithm is as below:

1. If the number of cases in the training set is N, sample n cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M (generally m~sqrt(M)) is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown as large as possible with limit of about 30levels.

The forest overall over rate depends on the following factors:

1. Correlation in between the trees of the forest: Higher the correlation, higher is the error
2. Strength of each of the individual trees in the forest.

One of the major advantages of the Random Forest Algorithm is that it has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

### Steps of Implementation

* Install the rattle package in R using: install.packages(“rattle”) 🡪 onetime activity
* Call to library(rattle)
* To start rattle 🡪 rattle()
* Import the data from the data tab, browse the csv data file and then execute. The size of training, validation and testing partition can be changed from the partition option given.
* Once imported, the data can be used directly to create the model or else can be subjected to imputation.
* For imputation: Go to transform 🡪 select the impute option and then select the method of imputation. Rattle only allows imputation of numeric variables for categorical variables imputation is done using the r package mi and the command mi.categorical.
* The model can be built and configured in the models tab. In this case we would select forest and the parameters – No. of trees – 500 and the no. of variables as 7 (this is square root of the number of predictors but can be increased or decreased), sample size of about 1000 samples and then execute.
* Once done, in the Evaluate tab the model can be validated for all the other partitions and even for a new source data file

### Model Formulation

The model for the Random forest developed as part of this exercise is as below:

#### *Model Equation*

Call:

randomForest(formula = as.factor(churn) ~ .,

data = crs$dataset[crs$sample, c(crs$input, crs$target)],

ntree = 500, mtry = 7, sampsize = c(1000), importance = TRUE, replace = FALSE, na.action = na.roughfix)

Definition of crs$input, crs$categoric, crs$risk, crs$target:

crs$input <- c("area", "hnd\_price", "eqpdays", "change\_mou",

"months", "mou\_Mean", "change\_rev", "avg3mou",

"totrev", "adjrev", "avgrev", "mou\_cvce\_Mean",

"totmou", "totcalls", "adjmou", "mou\_Range",

"avgmou", "avgqty", "mou\_peav\_Range", "adjqty",

"mou\_cvce\_Range", "totmrc\_Mean", "mou\_peav\_Mean", "mou\_rvce\_Mean",

"opk\_vce\_Mean", "avg3qty", "avg6qty", "complete\_Mean",

"recv\_vce\_Range", "recv\_vce\_Mean", "inonemin\_Mean", "mou\_rvce\_Range",

"peak\_vce\_Mean", "mouowylisv\_Mean", "age1", "plcd\_vce\_Range",

"mouiwylisv\_Range", "comp\_vce\_Range", "attempt\_Range", "attempt\_Mean",

"inonemin\_Range", "owylis\_vce\_Mean", "lor", "mou\_opkv\_Mean",

"ovrmou\_Range", "drop\_blk\_Range", "comp\_vce\_Mean", "ovrrev\_Range",

"blck\_vce\_Mean", "complete\_Range", "ovrrev\_Mean", "vceovr\_Range",

"ovrmou\_Mean", "cc\_mou\_Mean", "vceovr\_Mean", "da\_Mean",

"custcare\_Mean", "callwait\_Range", "uniqsubs", "asl\_flag",

"retdays", "tot\_ret")

crs$numeric <- c("hnd\_price", "eqpdays", "change\_mou", "months",

"mou\_Mean", "change\_rev", "avg3mou", "totrev",

"adjrev", "avgrev", "mou\_cvce\_Mean", "totmou",

"totcalls", "adjmou", "mou\_Range", "avgmou",

"avgqty", "mou\_peav\_Range", "adjqty", "mou\_cvce\_Range",

"totmrc\_Mean", "mou\_peav\_Mean", "mou\_rvce\_Mean", "opk\_vce\_Mean",

"avg3qty", "avg6qty", "complete\_Mean", "recv\_vce\_Range",

"recv\_vce\_Mean", "inonemin\_Mean", "mou\_rvce\_Range", "peak\_vce\_Mean",

"mouowylisv\_Mean", "age1", "plcd\_vce\_Range", "mouiwylisv\_Range",

"comp\_vce\_Range", "attempt\_Range", "attempt\_Mean", "inonemin\_Range",

"owylis\_vce\_Mean", "lor", "mou\_opkv\_Mean", "ovrmou\_Range",

"drop\_blk\_Range", "comp\_vce\_Mean", "ovrrev\_Range", "blck\_vce\_Mean",

"complete\_Range", "ovrrev\_Mean", "vceovr\_Range", "ovrmou\_Mean",

"cc\_mou\_Mean", "vceovr\_Mean", "da\_Mean", "custcare\_Mean",

"callwait\_Range", "uniqsubs", "retdays", "tot\_ret")

crs$categoric <- c("area", "asl\_flag")

crs$target <- "churn"

crs$risk <- NULL

crs$ident <- c("X", "Customer\_ID")

#### Model Results

**Data**: as.numeric(crs$rf$predicted) in 35261 controls (crs$rf$y 0) < 34739 cases (crs$rf$y 1).

Area under the curve: **0.6133**

**95% CI**: 0.6097-0.6169 (DeLong) – this is the 95% confidence interval for the overall model, we would be more interested in TP and FN primarily because that would be scenarios of predicted hit vs. actual churn and predicted miss vs. actual churn.

##### ROC for Training dataset: (Optimistic)

Value of AOC is optimistic because this is the same data used to train the model

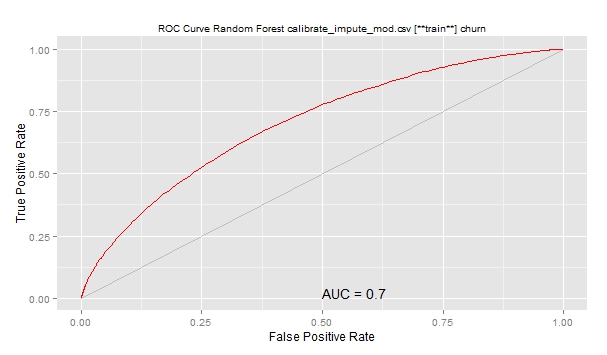


Figure : ROC Curve for training dataset using RF

##### ROC for the Validation dataset

For the validation partition data of calibration.csv, used to validate the model

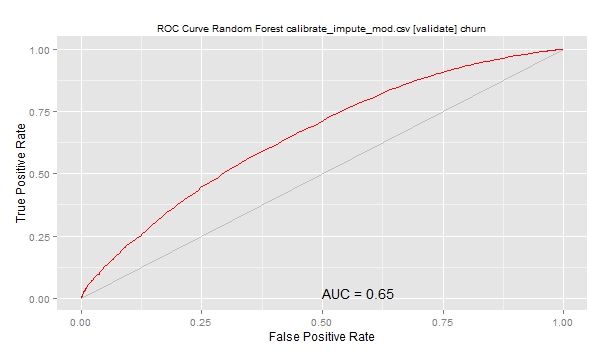


Figure : ROC Curve for validation dataset using RF

ROC for the testing dataset for the calibration.csv partition: the third partition created from the calibration.csv

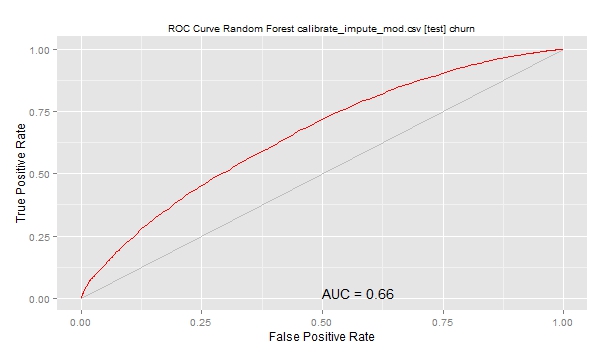


Figure : ROC Curve for Test data set using RF

##### Confusion Matrix & Its interpretation

Confusion matrix in rattle is also called as the error matrix. To generate an error matrix go to evaluate 🡪 error matrix and then execute.

**Testing data part of calibration.csv (entire results stored in testing\_data\_calibration.xlsx): Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Cutoff Probability** | |
| **Confusion Matrix at:** | | | **0.4** |
|  |  | **Predicted** | |
|  |  | 0 | 1 |
| **Actual** | 0 | 1294 | 6284 |
| 1 | **414** | 7008 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Cutoff Probability** | |
| **Confusion Matrix at:** | | | **0.5** |
|  |  | **Predicted** | |
|  |  | 0 | 1 |
| **Actual** | 0 | 4176 | 3402 |
| 1 | **2455** | 4967 |

**Training data part of calibration.csv (entire results stored in Training\_data\_calibration.xlsx): Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix at:** | | | **Cutoff Probability** |
|  |  |  | **0.4** |
|  |  | **Predicted** | |
|  |  | 0 | 1 |
| **Actual** | 0 | 6783 | 28478 |
| 1 | **1683** | 33056 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix at:** | | | **Cutoff Probability** |
|  |  |  | **0.5** |
|  |  | **Predicted** | |
|  |  | 0 | 1 |
| **Actual** | 0 | 21027 | 14234 |
| 1 | **10622** | 24117 |

**Validation data part of calibration.csv (entire results stored in validate\_data\_calibration.csv): Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | **Cutoff Probability** |
| **Confusion Matrix at:** | | | **0.4** |
|  |  | **Predicted** | |
|  |  | 0 | 1 |
| **Actual** | 0 | 1363 | 6236 |
| 1 | **418** | 6983 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | **Cutoff Probability** |
| **Confusion Matrix at:** | | | **0.5** |
|  |  | **Predicted** | |
|  |  | 0 | 1 |
| **Actual** | 0 | 4203 | 3396 |
| 1 | **2517** | 4884 |

**Insights:** As we can see there is a significant decline in the no. of False Negative as the cutoff probability comes down from 0.5 to 0.4 and considering the cost of acquisition vs. cost of a false offering we can determine the cutoff**.**

##### Computation of Gini coefficient

The Gini coefficient is a measure of the inequality of a distribution, a value of 0 expressing total equality and a value of 1 maximal inequality. For our purpose since we are using a classification, we want the value of Gini coefficient to be as high as possible which in other words would imply the separation between the two classes.

Table : Gini Coefficient computation for RF

|  |  |
| --- | --- |
| **Model: Random Forest** | **Gini Coefficient (2\*AOC – 1)** |
| Training Data Set | **2\*0.7-1 = 0.4** |
| Validation Data set | **2\*0.65-1 = 0.3** |
| Testing Data set | **2\*.66-1 = 0.32** |

**Prediction for current\_score data and future\_score data:**

The random forest model developed above was applied to the future data and current data and the results for the same are **RF\_future\_impute\_score\_idents.csv** and **RF\_current\_impute\_score\_idents.csv.**

## Artificial Neural Network

They belong to the family of statistical learning algorithms and used best when we want to determine an approximate function for a dataset with large number of predictors.

The ANN was applied on the predictors which were selected as part of the variable shortlist process above and the results for the below model formed were then analyzed.

### Steps of Implementation

Same as random forest with only difference being in the setting of parameters: here instead of no. of trees, one has to define the number of nodes.

Model Formulation:

ANN being mostly a black box model does not give a very friendly view of model formed except for the weights on each node.

#### Model Equation

**Inputs**: areaCALIFORNIA NORTH AREA, areaCENTRAL/SOUTH TEXAS AREA, areaCHICAGO AREA, areaDALLAS AREA, areaDC/MARYLAND/VIRGINIA AREA, areaGREAT LAKES AREA, areaHOUSTON AREA, areaLOS ANGELES AREA, areaMIDWEST AREA, areaNEW ENGLAND AREA, areaNEW YORK CITY AREA, areaNORTH FLORIDA AREA, areaNORTHWEST/ROCKY MOUNTAIN AREA, areaOHIO AREA, areaPHILADELPHIA AREA, areaSOUTH FLORIDA AREA, areaSOUTHWEST AREA, areaTENNESSEE AREA, hnd\_price, eqpdays, change\_mou, months, mou\_Mean, change\_rev, avg3mou, totrev, adjrev, avgrev, mou\_cvce\_Mean, totmou, totcalls, adjmou, mou\_Range, avgmou, avgqty, mou\_peav\_Range, adjqty, mou\_cvce\_Range, totmrc\_Mean, mou\_peav\_Mean, mou\_rvce\_Mean, opk\_vce\_Mean, avg3qty, avg6qty, complete\_Mean, recv\_vce\_Range, recv\_vce\_Mean, inonemin\_Mean, mou\_rvce\_Range, peak\_vce\_Mean, mouowylisv\_Mean, age1, plcd\_vce\_Range, mouiwylisv\_Range, comp\_vce\_Range, attempt\_Range, attempt\_Mean, inonemin\_Range, owylis\_vce\_Mean, lor, mou\_opkv\_Mean, ovrmou\_Range, drop\_blk\_Range, comp\_vce\_Mean, ovrrev\_Range, blck\_vce\_Mean, complete\_Range, ovrrev\_Mean, vceovr\_Range, ovrmou\_Mean, cc\_mou\_Mean, vceovr\_Mean, da\_Mean, custcare\_Mean, callwait\_Range, uniqsubs, asl\_flagY, retdays, tot\_ret.

Model Results:

Output: as.factor(churn).

Sum of Squares Residuals: 29617.6384.

No. of nodes: 5

##### Confusion Matrix for the training dataset

The detailed scores are in the csv sheets:

* For training partition: ANN\_Training\_data\_calibration.csv
* For validation partition: ANN\_Validate\_data\_Calibration.csv
* For testing partition: ANN\_Testing\_data\_calibration.csv
* For current score: ANN\_current\_impute\_score\_idents.csv
* For future score: ANN\_future\_impute\_score\_idents.csv

The confusion matrix for the training dataset results in an error rate of about 42% for the average class.

Table : Confusion matrix for training dataset using ANN

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | |  |
| Actual |  | 0 | 1 |
|  | 0 | 20845 | 14404 |
|  | 1 | 15214 | 19510 |

Also, the confusion matrix for the validation results in an error rate of about 42% for the average class.

Table :Confusion matrix for validation dataset using ANN

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | |  |
| Actual |  | 0 | 1 |
|  | 0 | 4485 | 3113 |
|  | 1 | 3212 | 4185 |

Also, the confusion matrix for the testing results in an error rate of about 43% for the average class.

Table :Confusion Matrix for ANN using testing data

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | |  |
| Actual |  | 0 | 1 |
|  | 0 | 4432 | 3145 |
|  | 1 | 3255 | 4160 |

##### Computation of Gini coefficient

The Gini coefficient is a measure of the inequality of a distribution, a value of 0 expressing total equality and a value of 1 maximal inequality. For our purpose since we are using a classification, we want the value of Gini coefficient to be as high as possible which in other words would imply the separation between the two classes.

Table :Gini Coefficient estimation using ANN

|  |  |
| --- | --- |
| **Model: ANN** | **Gini Coefficient (2\*AOC – 1)** |
| Training Data Set | **2\*0.58-1 = 0.16** |
| Validation Data set | **2\*0.57-1 = 0.14** |
| Testing Data set | **2\*.58-1 = 0.16** |

##### ROC Curves for Calibration datasets:

**Training dataset ROC for ANN**:

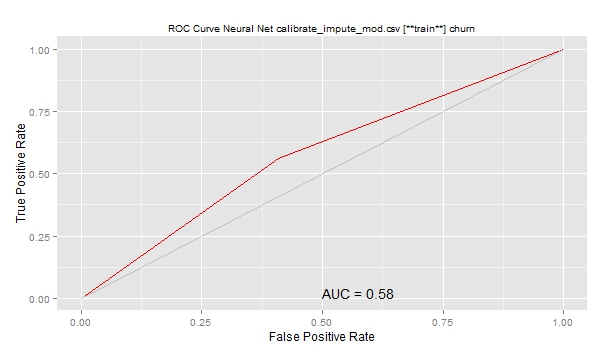


Figure : ROC using the training dataset for ANN

**Testing dataset ROC for ANN**:

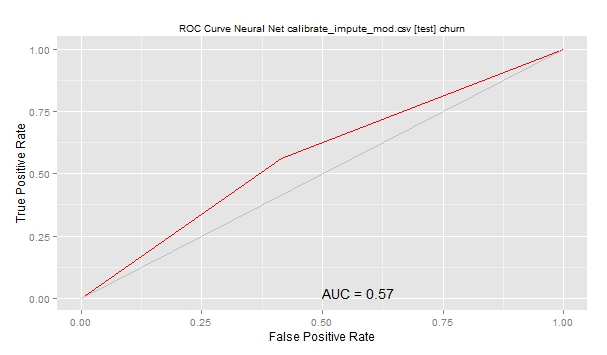


Figure : ROC for testing dataset using ANN

**Validation data ROC for ANN**:

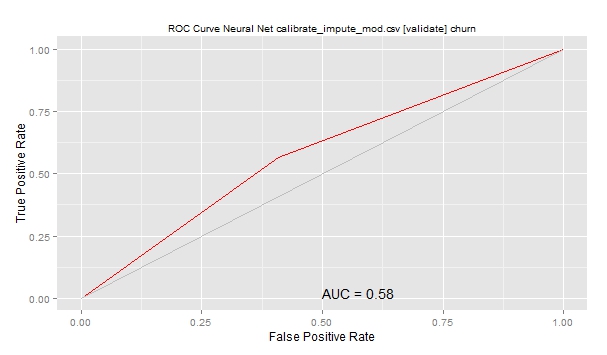


Figure : ROC for Validation data using ANN

Hence as we can see that for the ANN algorithm the performance is not as good as Random Forest for this set of data with an AUC of about 0.58 as compared to about .65 for Random Forest.

## Logistic Regression

The major advantage of using logistic regression in this case is because of the fact that most of the assumptions of the regression model are relaxed in case of logistic regression. The assumptions like that of linearity, normality, homoscedasticity and measurement levels.

### Steps of implementation

Similar to random forest except here we need to select the type of GLM which we want to use.

### Model Equation

Call: glm(formula = churn ~ ., family = binomial(link = "logit"), data = crs$dataset[crs$train,

c(crs$input, crs$target)])

The definitions of the crs$input, crs$target and crs$train are all the same as that of RF.

### Model Summary

**Deviance Residuals:**

Min 1Q Median 3Q Max

-2.8933 -1.1357 -0.6917 1.1357 2.6724

Table : Coefficient Table for LLN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** |
| (Intercept) | 0.279944136 | 0.154714217 | 1.809 | 0.070385 |
| areaCALIFORNIA NORTH AREA | 0.055225336 | 0.045033791 | 1.226 | 0.220082 |
| areaCENTRAL/SOUTH TEXAS AREA | -0.029957289 | 0.049071212 | -0.61 | 0.54154 |
| areaCHICAGO AREA | -0.010220052 | 0.046883613 | -0.218 | 0.827439 |
| areaDALLAS AREA | -0.034162501 | 0.045764284 | -0.746 | 0.455373 |
| areaDC/MARYLAND/VIRGINIA AREA | -0.151792395 | 0.043595928 | -3.482 | 0.000498 |
| areaGREAT LAKES AREA | -0.051600385 | 0.047898497 | -1.077 | 0.281352 |
| areaHOUSTON AREA | -0.060247333 | 0.04899525 | -1.23 | 0.218826 |
| areaLOS ANGELES AREA | 0.016963752 | 0.044324471 | 0.383 | 0.701929 |
| areaMIDWEST AREA | -0.125130993 | 0.044211226 | -2.83 | 0.00465 |
| areaNEW ENGLAND AREA | 0.039098864 | 0.045892571 | 0.852 | 0.394234 |
| areaNEW YORK CITY AREA | 0.046390725 | 0.039996717 | 1.16 | 0.246104 |
| areaNORTH FLORIDA AREA | 0.111459186 | 0.048641795 | 2.291 | 0.021939 |
| areaNORTHWEST/ROCKY MOUNTAIN AREA | 0.326360434 | 0.049135213 | 6.642 | 3.09E-11 |
| areaOHIO AREA | -0.136686885 | 0.047638569 | -2.869 | 0.004114 |
| areaPHILADELPHIA AREA | 0.089882501 | 0.058799386 | 1.529 | 0.126356 |
| areaSOUTH FLORIDA AREA | 0.186251974 | 0.053289313 | 3.495 | 0.000474 |
| areaSOUTHWEST AREA | 0.052709892 | 0.044612334 | 1.182 | 0.2374 |
| areaTENNESSEE AREA | -0.118253872 | 0.056405148 | -2.097 | 0.036037 |
| hnd\_price | -0.002254671 | 0.00014939 | -15.092 | < 2e-16 |
| eqpdays | 0.000817677 | 0.000041122 | 19.884 | < 2e-16 |
| change\_mou | -0.000407093 | 0.000067103 | -6.067 | 1.31E-09 |
| months | -0.007093784 | 0.001606955 | -4.414 | 1.01E-05 |
| mou\_Mean | -0.000575664 | 0.0001637 | -3.517 | 0.000437 |
| change\_rev | 0.001989572 | 0.000289518 | 6.872 | 6.33E-12 |
| avg3mou | -0.000462869 | 0.000184411 | -2.51 | 0.012074 |
| totrev | 0.000430028 | 0.000185461 | 2.319 | 0.020411 |
| adjrev | -0.000599327 | 0.000191639 | -3.127 | 0.001764 |
| avgrev | 0.00173953 | 0.000750543 | 2.318 | 0.020466 |
| mou\_cvce\_Mean | 0.005050699 | 0.004822501 | 1.047 | 0.294952 |
| totmou | -0.000087748 | 0.00010442 | -0.84 | 0.400718 |
| totcalls | 0.000319857 | 0.000264347 | 1.21 | 2.26E-01 |
| adjmou | 0.000109822 | 0.00010513 | 1.045 | 2.96E-01 |
| mou\_Range | 0.000211163 | 0.000034866 | 6.056 | 1.39E-09 |
| avgmou | 0.000552283 | 0.000108287 | 5.1 | 3.39E-07 |
| avgqty | -0.000007384 | 0.000284732 | -0.026 | 0.979311 |
| mou\_peav\_Range | 0.00018246 | 0.000124911 | 1.461 | 0.144093 |
| adjqty | -0.000331528 | 0.000266124 | -1.246 | 0.21285 |
| mou\_cvce\_Range | -0.000139178 | 0.00009062 | -1.536 | 0.124575 |
| totmrc\_Mean | -0.001662555 | 0.000555151 | -2.995 | 0.002746 |
| mou\_peav\_Mean | -0.005292744 | 0.00482139 | -1.098 | 0.272308 |
| mou\_rvce\_Mean | 0.005196037 | 0.004820786 | 1.078 | 0.281105 |
| opk\_vce\_Mean | -0.001614676 | 0.000645566 | -2.501 | 0.012378 |
| avg3qty | 0.000767321 | 0.00030762 | 2.494 | 0.012618 |
| avg6qty | 0.000030506 | 0.000180224 | 0.169 | 0.865585 |
| complete\_Mean | -0.000803107 | 0.00143255 | -0.561 | 0.575061 |
| recv\_vce\_Range | -0.000455842 | 0.000744481 | -0.612 | 0.540343 |
| recv\_vce\_Mean | 0.001340986 | 0.001138656 | 1.178 | 0.23892 |
| inonemin\_Mean | -0.001157924 | 0.001155047 | -1.002 | 0.316107 |
| mou\_rvce\_Range | 0.000041077 | 0.000144956 | 0.283 | 0.77689 |
| peak\_vce\_Mean | -0.001798126 | 0.000641101 | -2.805 | 0.005036 |
| mouowylisv\_Mean | 0.000574882 | 0.00035544 | 1.617 | 0.105796 |
| age1 | -0.00482694 | 0.000378907 | -12.739 | < 2e-16 |
| plcd\_vce\_Range | -0.004548874 | 0.002958614 | -1.538 | 0.12417 |
| mouiwylisv\_Range | -0.000906866 | 0.000276563 | -3.279 | 0.001042 |
| comp\_vce\_Range | 0.006215093 | 0.003249214 | 1.913 | 0.055774 |
| attempt\_Range | 0.003704134 | 0.002955553 | 1.253 | 0.210104 |
| attempt\_Mean | 0.001520595 | 0.000397866 | 3.822 | 0.000132 |
| inonemin\_Range | 0.000753697 | 0.000970878 | 0.776 | 4.38E-01 |
| owylis\_vce\_Mean | -0.000296892 | 0.000615993 | -0.482 | 0.629825 |
| lor | -0.014831546 | 0.002025263 | -7.323 | 2.42E-13 |
| mou\_opkv\_Mean | -0.004848605 | 0.004827709 | -1.004 | 0.31522 |
| ovrmou\_Range | 0.00019598 | 0.000333291 | 0.588 | 0.556523 |
| drop\_blk\_Range | 0.002255792 | 0.000800387 | 2.818 | 0.004827 |
| comp\_vce\_Mean | 0.000177155 | 0.001522542 | 0.116 | 0.907372 |
| ovrrev\_Range | -0.003374896 | 0.002937194 | -1.149 | 0.250548 |
| blck\_vce\_Mean | 0.000843364 | 0.001188365 | 0.71 | 0.4779 |
| complete\_Range | -0.005070927 | 0.003235834 | -1.567 | 0.117088 |
| ovrrev\_Mean | 0.004561026 | 0.007223526 | 0.631 | 0.527771 |
| vceovr\_Range | 0.004203113 | 0.002829361 | 1.486 | 0.137402 |
| ovrmou\_Mean | 0.000389414 | 0.000651424 | 0.598 | 0.549981 |
| cc\_mou\_Mean | -0.000588222 | 0.00118364 | -0.497 | 0.619217 |
| vceovr\_Mean | -0.002094097 | 0.007080959 | -0.296 | 0.767431 |
| da\_Mean | 0.006006197 | 0.004379299 | 1.371 | 0.17022 |
| custcare\_Mean | -0.002757452 | 0.002496853 | -1.104 | 0.269432 |
| callwait\_Range | -0.004649221 | 0.002600694 | -1.788 | 0.073827 |
| uniqsubs | 0.09100679 | 0.009268532 | 9.819 | < 2e-16 |
| asl\_flagY | -0.410246193 | 0.02547057 | -16.107 | < 2e-16 |
| retdays | -0.000982579 | 0.000185074 | -5.309 | 1.10E-07 |
| tot\_ret | 0.08116423 | 0.125253104 | 0.648 | 0.516984 |

##### ROC Curves for Testing, training and validation datasets

**ROC curve for testing dataset using LLN:**

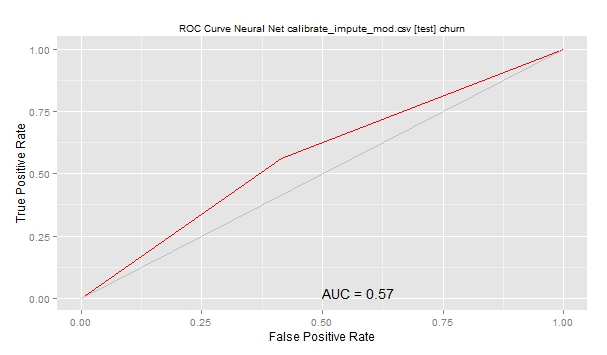
****

Figure : ROC Curve for testing Dataset using LNN

**ROC Curve for Training Dataset using LNN:**

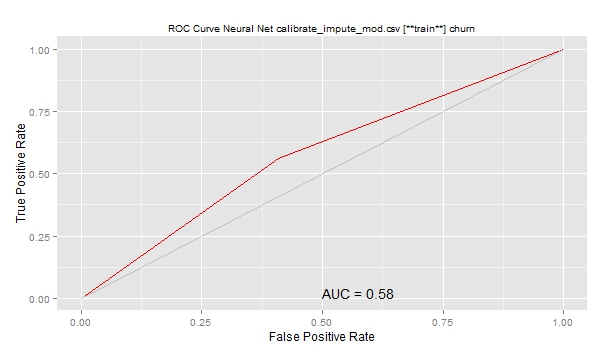
****

Figure : ROC Curve for Training Dataset using LNN

**ROC Curve for Validation Dataset using LNN:**

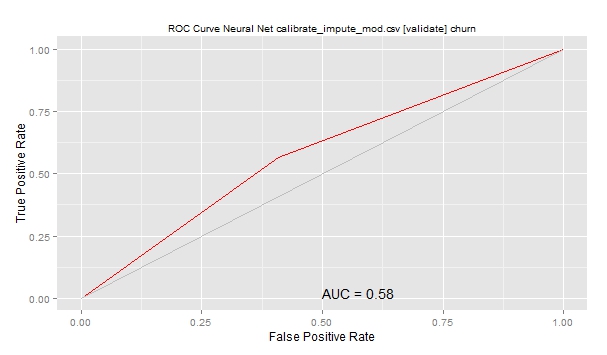
****

Figure : ROC Curve for Validation Dataset for LNN

**Insights**: As clear from the ROC curve the Logistic model is little improvement as compared to the random algorithm with ~0.57 AUC but is worst performer when compared to the random forest.

*Computation of Gini coefficient*

The Gini coefficient is a measure of the inequality of a distribution, a value of 0 expressing total equality and a value of 1 maximal inequality. For our purpose since we are using a classification, we want the value of Gini coefficient to be as high as possible which in other words would imply the separation between the two classes.

Table :Gini Coefficient LLN

|  |  |
| --- | --- |
| **Model: LLN** | **Gini Coefficient (2\*AOC – 1)** |
| Training Data Set | 2\*0.58-1 = 0.16 |
| Validation Data set | 2\*0.57-1 = 0.14 |
| Testing Data set | 2\*.58-1 = 0.16 |

Prediction for current\_score.csv and future\_score.csv are in the files: LNN\_current\_impute\_score\_idents.csv and LLN\_future\_impute\_score\_idents.csv

# Conclusion

The churn of customers in telecom industry remains to be an interesting field of study considering the value that the churn prediction adds to the business. The prediction algorithms used and their comparison as summarized below.

Table : Comparison of Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method/Algorithm** | **AOC** | **Gini Coefficient** | **Dataset Phase** | **Confusion Matrix Error** |
| Random Forest | 0.70 | 0.40 | Training | 30.22% |
| ANN | 0.58 | 0.16 | Training | 43.32% |
| Logistic Regression | 0.57 | 0.14 | Training | 42.02% |
|  |  |  |  |  |
| Random Forest | 0.65 | 0.30 | Validation | 34.00% |
| ANN | 0.58 | 0.16 | Validation | 42.00% |
| Logistic Regression | 0.58 | 0.16 | Validation | 43.00% |
|  |  |  |  |  |
| Random Forest | 0.66 | 0.32 | Testing | 33.00% |
| ANN | 0.57 | 0.14 | Testing | 44.00% |
| Logistic Regression | 0.58 | 0.16 | Testing | 40.00% |

From the study made, we can say that the process of churn prediction is highly dependent on the following factors:

1. A prior data visualization exercise as well as background study helps in such cases to reduce the number of variables
2. Prediction accuracy highly depends upon the following factors:
   1. The algorithm selected and the data (predictors) used
   2. No. of trees or iterations in case of Random Forest or ANN Respectively
3. **A RF algorithm with a cutoff probability at 0.4 gives a fairly good model for the prediction with an error for TP of about 29%.**
4. Use of prediction algorithms can help save companies huge dollars in terms of not approaching the wrong customer who is never going to leave.
5. Process of data imputation is quite useful in reducing the overall error rate for the algorithm to work and hence it is useful provided when done with care because over imputation is misleading.

# References

* Shmueli, Patel & Bruce, Data Mining and Business Intelligence 1st edition, Wiley publication
* Junxiang Lu, Predicting Customer Churn in the Telecommunications, SUGI data mining techniques
* Shyam V Nath, Customer Churn Analysis in the Wireless Industry: A data mining approach
* Scott A. Neslin, Sunil Gupta, Wagner Kamakura, Defection detection: measuring & understanding the predictive accuracy of Customer churn models