

**team: ISSSR**

Ishrath Sharffuddin, Roshan zubair, sandeep matta, sneha prabhu, suyash tiwari

Telecom Customer Churn

Project Report

**INSTRUCTOR:**

Shun-YanG lee

**Teaching ASSISTANT:**

Svecha Vangala

**Contents**

[Executive Summary](#_Executive_Summary) 2

[Business Case](#_Business_Case) 2

[Problem Statement](#_Problem_Statement) 2

[Data Overview](#_Data_Overview:) 2

[Source](#_Source) 2

[Description](#_Description) 2

[Key Variables](#_Key_Variables) 3

[Exploratory Data Analysis](#_Exploratory_Data_Analysis:) 3

[Univariate Analysis](#_Univariate_Analysis)  [3](file:///C:\Users\psneh\Downloads\Yashaswini.docx#_Toc468212543)

[Multivariate Analysis](#_Bivariate_Analysis) 4

[Correlation Matrix](#_Correlation_Matrix) 5

[Preprocessing and Variable Transformation](#_Preprocessing_and_Variable) 5

[Dimension/Variable Reduction](#_Dimension/Variable_Reduction) 6

[Feature Analysis](#_Feature_Analysis) 6

[Adaptive LASSO](#_Adaptive_LASSO_(generalized) 6

[Weight of Evidence](#_Weight_of_Evidence) 7

[Modeling Analysis](#_Modeling_Analysis) 8

[Data Sampling](#_Data_sampling_(partition) 8

[Nominal Logistic Analysis](#_Fitting_the_Logistic) 8

[Decision Tree Analysis](#_Decision_Tree) 9

[Random Forest Analysis](#_Random_Forest) 10

[Gradient Boosting Analysis](#_Gradient_Boosting_Model) 11

[Grid Search](#_Grid_Search) 11

[Ensemble Analysis](#_Ensemble_Model) 12

[Model Evaluation](#_Model_Evaluation) 12

[Business Insights and Conclusion](#_Business_Insights) 13

# **Executive Summary**

## **Business Case**

The business case brings together the benefits, disadvantages, costs, and risks of the telecom company. The predictions majorly targets tackling the pre-existing problem of customer churn. It also helps enhance the future vision so that executive management can decide how the customer churning rate can be reduced and also improve customer service by targeted approach methodology. The benefits and limitations describe the purpose of the predictions to improve quality, save costs through efficiencies, generate revenue for the telecom company, remain competitive, and majorly improve customer service.

## **Problem Statement**

The data from the repository contained data with null values and it required some pre-processing such as the null value treatment, variable reduction, outlier analysis, lasso regression and clustering to create a model that helps us predict the customer churn. The large set of variables being categorical, was a major challenge and hence also required dummy variable creation and manipulation. The accuracy was enhanced to predict the customer churn rate probability better.

# **Data Overview:**

## **Source**

This public dataset is fetched from EmcienScan (link provided below) and contains the customer records of a telecommunication industry interested in analyzing whether the customer will churn or not.

<http://www.scan-support.com/help/sample-data-sets>

## **Description**

The dataset contains the details of customers with close to **7043** observations and **21** variables(Features). The variables provide customer personal information and a detailed description on current preference of the services used (contract duration, paperless billing, payment method, internet service etc.), the location, date/time, vehicle details, driver details etc.

Total Observations: 7043

Total Variables: 21

## **Key Variables**

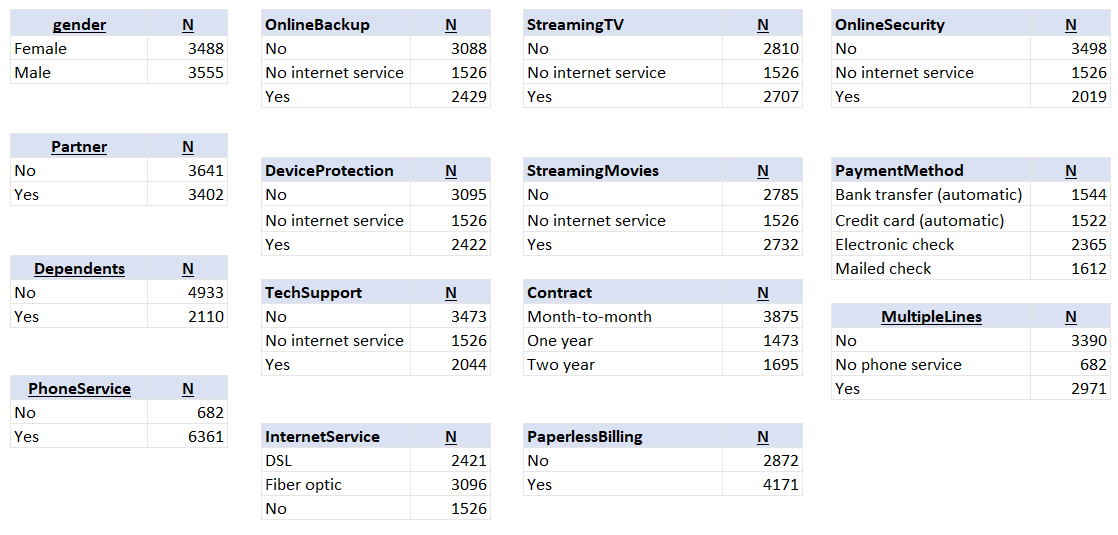
There are the variables which are used as the predictors to build the model. Based on these variables the target variable is predicted using the model.

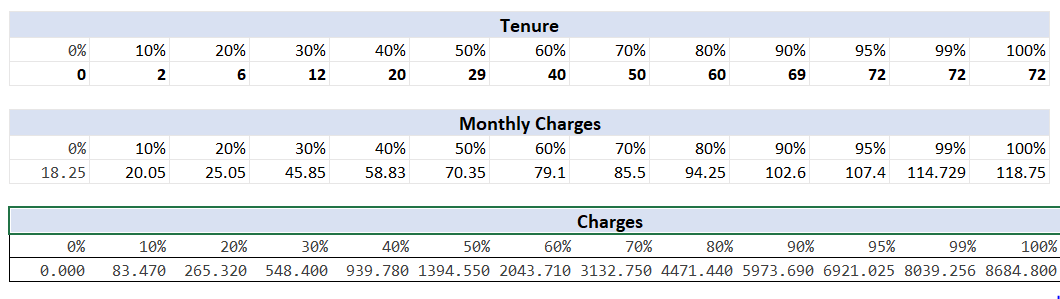
* Customer information:
  + Senior Citizen
* Telecom Service used:
  + Online Security
  + Tech Support
* Duration of service:
  + Tenure
  + Contract
* Service Charges:
  + Monthly Charges

# **Exploratory Data Analysis:**

## **Univariate Analysis**

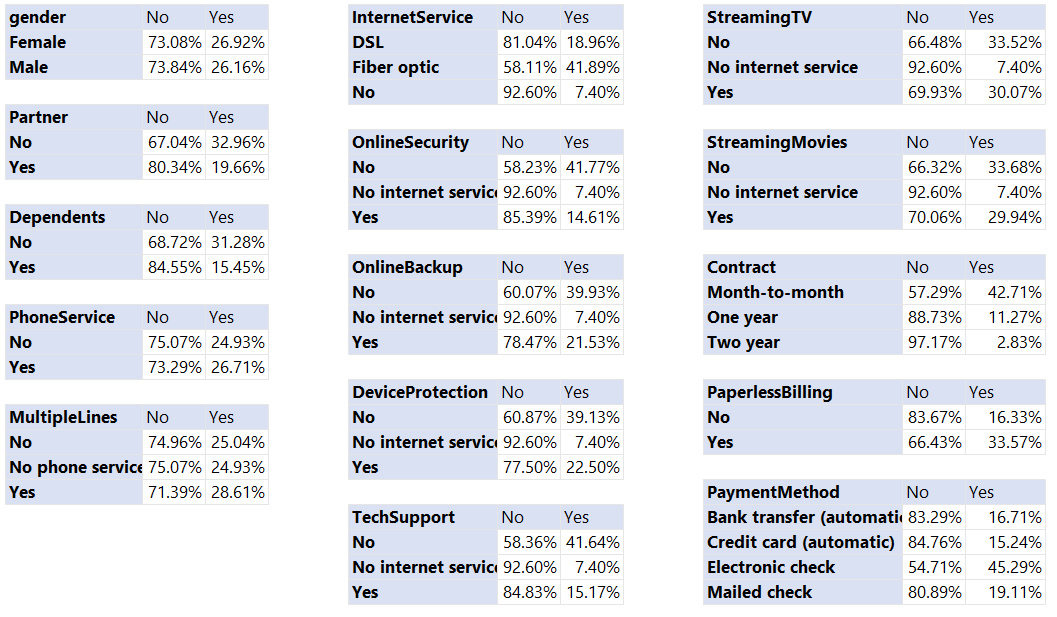
The summarization of individual variables aided in understanding the distribution of data. The information such as the count of records in various levels of the variable, minimum and maximum values of a continuous variable helped in preliminary analysis such as outlier detection, missing values etc.





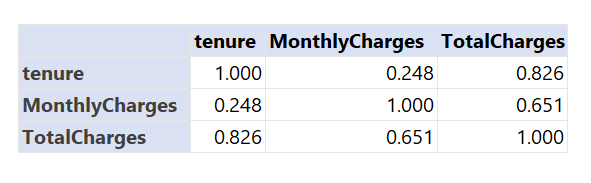
## **Bivariate Analysis**

The relationship between the target variable and the input variable is meaningful to capture any hidden trend. The data is easy to interpret and is in a readable format. A visible trend observed after analysis, is that, the churning is very likely when a customer is having a month-to-month contract than any other contract durations and also when there is no tech support taken/or provided.



## **Correlation Matrix**

It is essential to remove variables having high correlation as it can affect model’s prediction ability. The Monthly Charges and Total Charges are correlated. So, one of them will be removed from the model. We removed Total Charges.



## **Preprocessing and Variable Transformation**

Looking at the variables, we needed to do some data transformation to generate new features that might help improve the model accuracy. Below are the findings and solutions:

* No duplicate data in the data set.
* Missing values were found only in one variable **TotalCharges**. Value was imputed with product of MonthlyCharges and tenure

|  |  |
| --- | --- |
| Questions | Results(Numeric) |
| Total number of observations in the dataset | 7043 |
| Total number of variables in the dataset | 21 |
| Total missing values in the dataset | 11 |

* No outliers were found
* Bring the data in the correct format.
  + **SeniorCitizen** variable was changed to factor

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | None |
| Dummy creation | 1. Partner 2. PhoneService 3. Contract 4. PaperlessBilling 5. PaymentMethod 6. InternetService 7. OnlineSecurity 8. TechSupport |
| Binning of variables | 1. Tenure |

Added a new data column to convert the categorical target column “**Churn**” into numerical variable “**churn\_number**”

# **Dimension/Variable Reduction**

## **Feature Analysis**

The total number of independent variable in our model were 21. Using all the 21 variables to build a predictive model would increase the model complexity. The variables that affect the dependent variable majorly were selected using the Univariate and Bivariate analysis. So, out of original 21 variables we concluded that only 7 variables are required to build an effective predictive model.

## **Adaptive LASSO** (generalized regression for churn =Yes)

This dimension reduction technique used to select the important predictors. JMP has an inbuilt algorithm that uses the suitable penalty value while reducing the coefficients of unimportant predictors to zero.



Based on this result, the variables (Tenure, Internet Service, Online security, TechSupport, Contract, Payment Method) having very small p-value are significant for model building. Used JMP SAS Pro for this anaylsis.

## **Weight of Evidence**

This method is a powerful technique to assess the predictive power of categorical, ordinal and continuous together. The goal is to understand the uncertainty involved in predicting the outcome of random events given varying degrees of knowledge of other variables. It compares the strength of continuous and categorical variables without creating dummy variables. The weight of evidence (WOE) and information value (IV) provide a great framework for exploratory analysis and variable screening for binary classifiers.

The variables are categorized based on their Information Value:

< 0.02 : unpredictive

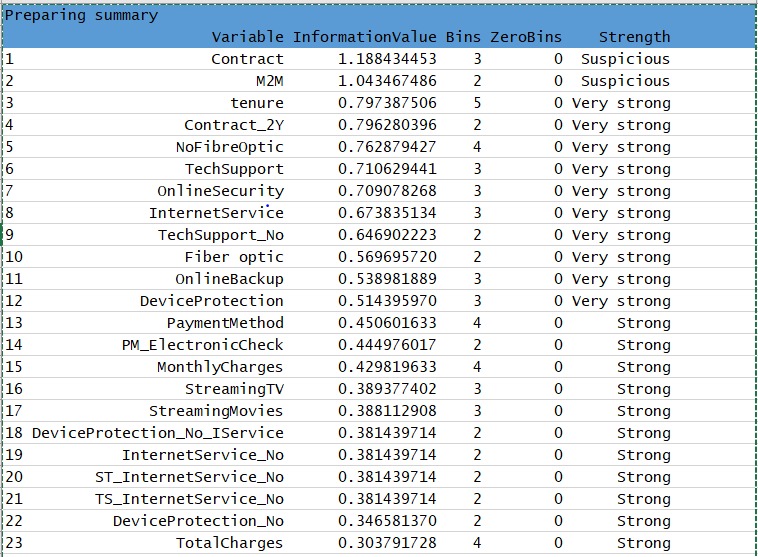
0.02 to 0.1 : weak

0.1 to 0.3 : medium

0.3 to 0.5 : strong

> 0.5 : suspicious

Based on this result, the variables (Contract, M2M, tenure, Contract\_2Y, NoFibreOptic, OnlineSecurity) having very small p-value are significant for model building.



# **Modeling Analysis**

## **Data sampling (partition into test and training dataset)**

For building a model we’ll want dataset which can be used to evaluate the model’s predictive power. Sampling the data set into train-test data will help us build the model on training data set and test the model’s performance on testing dataset.

A 75/25 split seems reasonable.75% is 5282 observations is randomly sampled to make a training data set and rest 25% i.e. 1761 records are marked as test data. Finally, we are ready to create the model.

## **Build Model**

## **Fitting the Logistic Regression Model: -**

As our dependent variable is categorical. We converted the target variable “**Churn**” – Yes/No to a numerical variable “**churn\_number**” - 1/0. We created a logistic model including only the important variables that we filter using the feature analysis. Variables with less significance and low VIF are retained till model cannot be bettered.

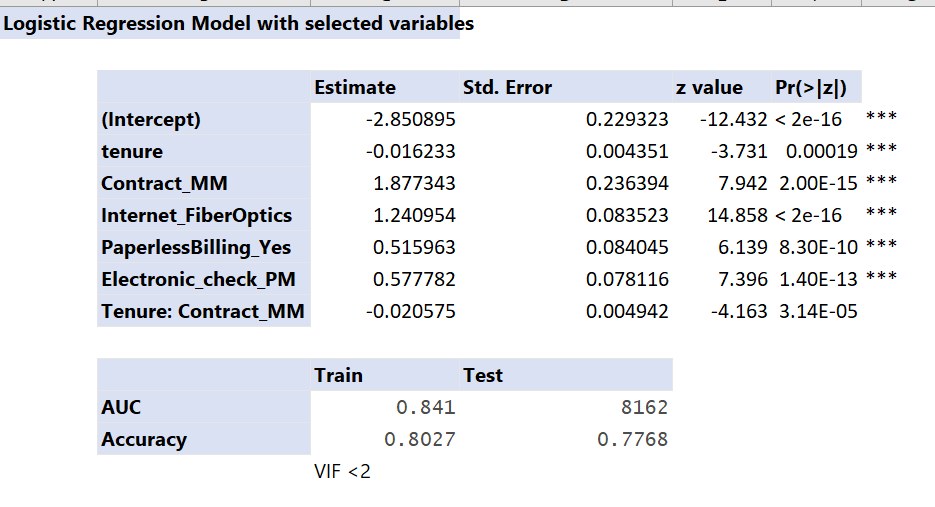
glm(formula = train$churn\_number ~ tenure \* Contract\_MM + Internet\_FiberOptic + PaperlessBilling\_Yes + Electronic\_check\_PM, family = binomial("logit"),

data = train)

|  |  |  |
| --- | --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** | **Interpretation** |
| Intercept | -2.850895 | Negative value |
| tenure | -0.016233 | Negative impact on churn variable. |
| Contract\_MM | 1.877343 | Positive impact on churn variable. |
| Internet\_FiberOptics | 1.240954 | Positive impact on churn variable. |
| PaperlessBilling\_Yes | 0.515963 | Positive impact on churn variable. |
| Electronic\_check\_PM | 0.577782 | Positive impact on churn variable. |
| Tenure: Contract\_MM | -0.020575 | Negative impact on churn variable. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| AUC (C-statistic) | 0.841 | AUC (C-statistic) | 8162 |
| Accuracy | 0.8027 | Accuracy | 0.7768 |
| Model Evaluation (write Accept or Reject) | | Accept | |

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7768 |
| Sensitivity | 0.7958 |
| Specificity | 0.6555 |

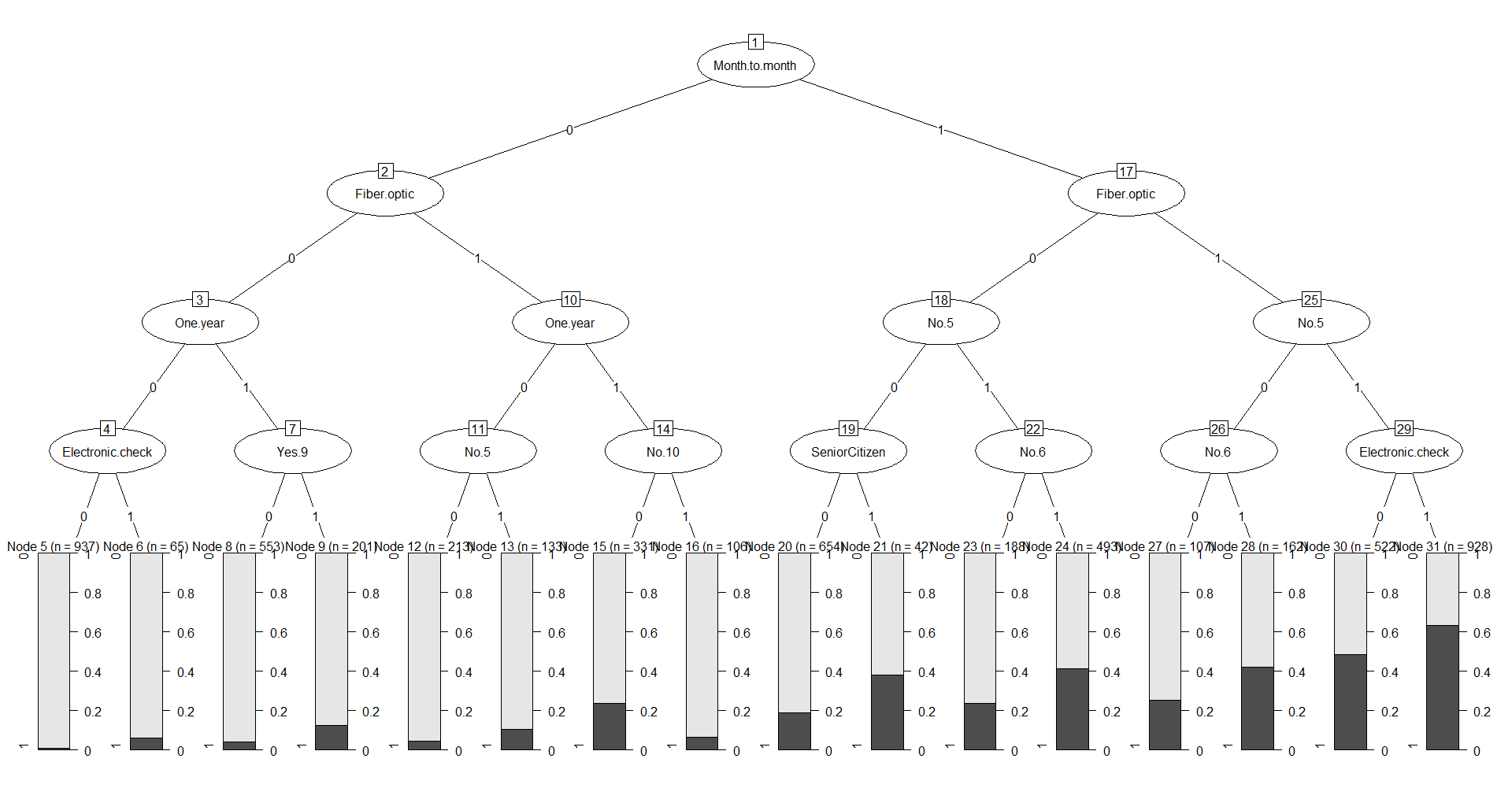


## 

## **Decision Tree**

CHAID Tree -  CHAID is most useful for analysis. Our goal here was to describe or understand the relationship between a response variable and a set of explanatory variables.

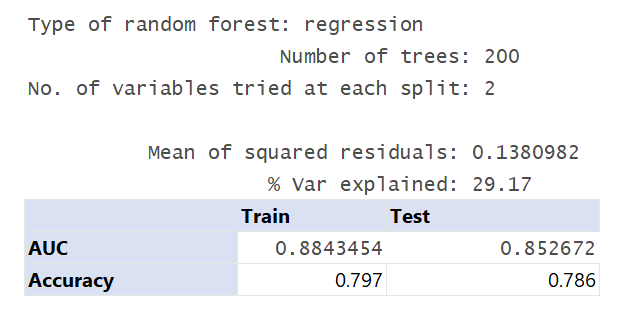
For illustration purpose, we are going all variables for plotting Decision Trees, they are “Month\_to\_Month”, “Fibre Optics”, “OneYear”, “Electronic Check”, “Senior Citizen”.



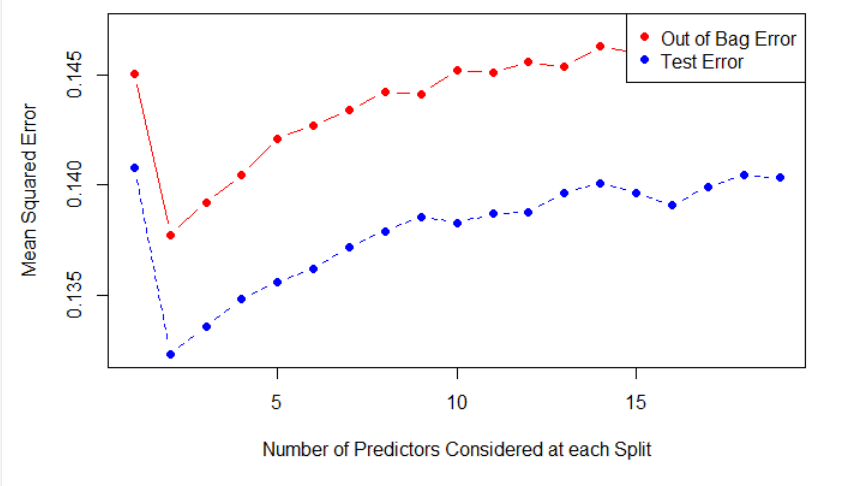
## **Random Forest**

Random Forests the idea is to decorrelate the several trees which are generated on the different bootstrapped samples from training Data. And then we simply reduce the Variance in the Trees by averaging them.

Averaging the Trees helps us to reduce the variance and also improve the Performance of Decision Trees on Test Set and eventually avoid Overfitting.

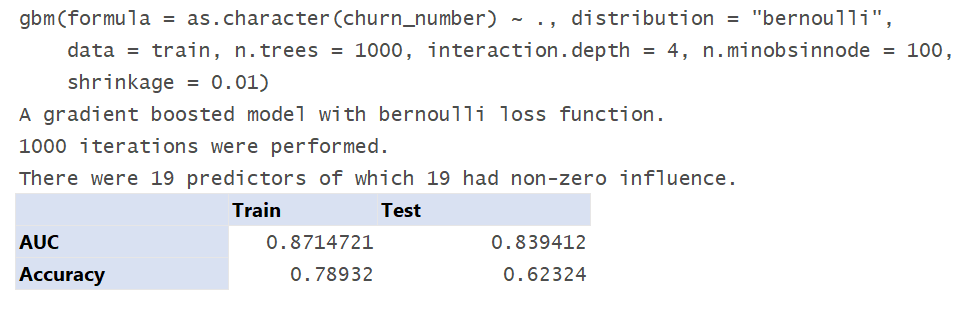


### **Plotting both Test Error and Out of Bag Error**



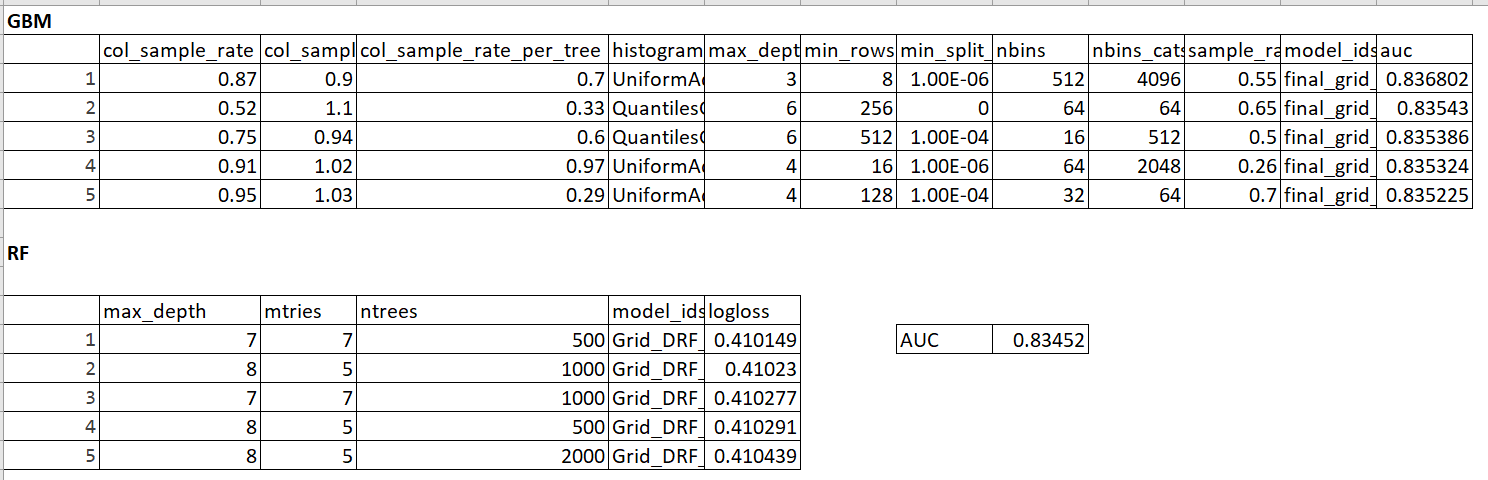
## **Gradient Boosting Model**

The Gradient Boosting model was built using all the independent variables. Fitted a gbm (gradient boosted model) to a subset of the data (training data) to generate a list describing how each variable reduced the squared error.



## **Grid Search**

Grid search was performed, involves running a model many times with combinations of various hyperparameters. The point is to identify which hyperparameters are likely to work best.



## **Ensemble Model**

This technique was performed by combining the algorithms of logistic regression, random forests and K-nearest neighbors. The average probability was then calculated to get the accuracy and AUC of the ensembled model.

|  |  |  |
| --- | --- | --- |
| Algorithm | Accuracy | AUC |
| Random Forest | 0.8069 | 0.726831 |
| K-nearest neighbor | 0.8075 | 0.728573 |
| Logistic Regression | 0.8085 | 0.724479 |
| **Ensembled model 1 - Average of Probabilities** | **0.80854** | **0.846516** |
| **Ensembled model 2 -Weighted Average of Probabilities** | **0.81449** | **0.847777** |

## **Model Evaluation**

K-Fold Cross Validation

To evaluate the model performance, we used k-fold cross validation technique to assess how well our model performs. We did the most common variation i.e. 10-fold cross validation. The result seems to be consistent with the result we got from a normal sampling.

AUC (C-statistic) - 0.8443475

# **Business Insights**

* Monthly plan customers are more likely to churn
* Person having no dependents are churning more
* No device protecting - more likely to churn
* Whoever using fiber optic internet services, they are more likely to churn
* Person having no online backup, online security is more likely to churn
* Unmarried people are churning more as compare to married people
* Percentage of churning on senior citizen is higher
* No tech support, Streaming movies, Streaming TV more likely to churn

