Abstract

The aim of this project is to conduct EDA and model building for churn prediction model to help a telecommunication corporation predict/understand which customer are in risk of churning. Thus, allowing the company to be able to design marketing campaign or conduct intervention in order attempt to prevent churn from occurring.

Bu.330.780.52 Data Science & Business Intelligence

Final Project

Lok Tin Kevin Chan, Hanchi Gu, Xinyi Li, Sihan Niu, Yuzi Wang, Lulu Xing

Group 6

Table of Content

[Business Understanding 2](#_Toc8212503)

[Data Understanding and Visualization 3](#_Toc8212504)

[Data Understanding 3](#_Toc8212505)

[Visualizations 4](#_Toc8212506)

[Modelling 5](#_Toc8212507)

[Logistic Regression 5](#_Toc8212508)

[Decision Tree 6](#_Toc8212509)

[Random Forest 7](#_Toc8212510)

[Linear SVM 7](#_Toc8212511)

[Threshold Analysis 8](#_Toc8212512)

[Evaluation 8](#_Toc8212513)

[Deployment 9](#_Toc8212514)

[Appendix A: Metadata Information 11](#_Toc8212515)

[Appendix B: Figures & Tables 12](#_Toc8212516)

[Appendix C: Rmd Code file 13](#_Toc8212517)

## Business Understanding

Telecommunication industry in the past couple of years have faced pricing pressure, with the disappearance of the differentiability of product and service between network has led customer to lose loyalty and become indifferent between telecommunication companies. This combined with the increase competition from cable operators and startups, has forced many telecommunication companies to come up with new strategies to attempt to retain their existing customer by offering more competitive offers, bundles and price cuts.

Our client, an incumbent in the telecommunication industry, wish to be able to identify and predict if a customer will churn. As the on average the cost to acquire a new customer is much more expensive to retain a customer. Our client thus wishes to use data science models to build a model that is able to identify potential churning customer and target them those customers with a new competitive offers/bundles or other marketing campaign in an attempt to retain them as budget is limited and they want to have most effective return on investment. Furthermore, in the process of building the model, to maybe gain some understanding to why customer churn, or characteristics that lead to customer leaving their company to a competitor.

Though there are no information about what our client has done in attempt to solve this churn issue, we believe that currently like most telecommunication companies, our client has mistakenly focused on creating new integrated services and plans in attempt to attract new clients and have not focus on retaining their existing customers. Thus, now our client wishes to explore the optionality of client retention strategy by spending a limited amount of budget to see if this can yield an increase in customer lifetime value.

## Data Understanding and Visualization

### Data Understanding

The data used to address the client’s problem was retrieved from Kaggle: Teleco Customer Churn: <https://www.kaggle.com/blastchar/telco-customer-churn> (Metadata are shown in Appendix A).

As the nature of the problem and data source, we are trying to predict whether a customer will churn in the next month, we can thus analyze the problem as a classification problem. The dataset contains 7,043 client information and 20 variables (with 16 categorical variable, 3 continuous variable and 1 target variable – Churn). This dataset is most likely cross-sectional subset of the database of customer information on a particular month, with each row of dataset to represent a customer and each column containing customer attributes of the following:

* Services that each customer has signed up for
* Customer account information
* Demographic information about customer
* Target variable: whether customer has left within the last month

First conducting a simple missing data analysis, we can observe from Appendix B: Figure 1, that out of 7043 observations, there are only 11 missing values and all of them belong to missing total charges column. Taking a detail look at these 11 observations, the commonality between them are that tenure = 0, meaning that these customers are most likely in their first month and thus have not been billed. Since none of these 11 clients has churned, and thus would have minimum impact, we remove them from our dataset.

After an initial visualization of the variables, it is evident that some categorical variables contain: “No”, “No internet Service” or “No Phone Service”, this for all analytical purpose should mean the same. We thus convert these them to be same. Furthermore, for our continuous variables since tenure, monthly charge and total charges are on different scales, we thus standardized them to reduce the bias that might result from the difference in scale.

In order to ensure that we are not overfitting our dataset, we will use a holdout validation method to validate our results. Thus, we split our 7032 observation into a training set (80%) and testing set (20%), though the following visualization analysis will be conducting on the full dataset and the holdout validation will only apply when conducting model building.

### Visualizations

First looking at the distribution of churn in our dataset (Appendix B: Figure 2), we observe that around 26% of the customer left out client within the past month. This shows that our dataset is somewhat unbalance with more data leaning towards un-churned customer.

Looking at our categorical variables with respect to churn (Appendix B: Figure 3,4,5), we identify the following trends:

1. Senior Citizen has higher churn percentage
2. Customer with dependents or partners tend to have lower churn rate
3. Customer with no online security, online backup or tech support have higher churn rate
4. Customer with monthly subscription are more likely to churn compared to longer contracts
5. Customer with electronic check payment method tend to churn more compared to other options

Most of these trends observed except for 1 and 5 are in line with our team’s intuition, as customer with no additional service are less tied in the eco-system, and thus would be more easily to switch between carriers. This logic also applies to customer with monthly subscription, as they are less tied down, and thus would be easier for them to switch to another company if provided a lower cost. Finally, customer with dependents / partner would be less likely to switch companies as it impacts not only one person, but multiple person and these customers are more likely to have less time to go and compare and find service with better rate. For trend 1 and 5, the difference in churn may result from other correlation with other factors.

Looking at the continuous variables (Appendix B: Figure 6), from the diagonal distribution plots, we can identify two trends:

1. Recent client is more likely to churn
2. Clients with higher monthly charges are more likely to churn

This appears to follow simple intuition, customer that are with the telecommunication firm for only a short period would in general have no loyalty compared to long time customer which are comfortable with the service provided and thus less willing to switch. Also, customer with higher monthly charges, will in general wish to reduce cost by seeking alternative service provider that may provide the same level of service for lower cost. From Figure 6, the scatter plots between continuous variables in general follows the trend described above and allow us to visually see if there are any obvious outliers which in this case seem to be none.

## Modelling

The business problem at hand for the telecommunication corporation is a classification problem, thus we analysis the data with different supervised data mining algorithms.

### Logistic Regression

Though logistic regression relies on linear parametric assumptions and is less flexible compared to other models, it will allow us to not only predict but also understand the relationship of the variables with respect to churn. Applying a forward stepwise logit regression to our data, will yield results shown in Appendix B: Figure 7.

From the above model parameters, the larger the coefficient value, the more impact it has on customer churn (more specifically the log odds of a customer churning). Among the nine variables the forward stepwise selected logit model: contract, internet service, and tenure seem to have greatest impact on churn rate.

* Customer with longer contract tends to be more viscid.
* Customer with fiber optic service has higher churn rate than that with DSL service.
* Recent clients are more likely to churn.

All this interpretive are in line with what we observed in the visualization of the data.

Because the variance inflation factor (VIF) of total charges (value of 4.2) is high, it means that our model suffers from multi-collinearity problem, we thus remove this variable from our model. The final logit model has 0.763 accuracy, and 0.780 sensitivity from our holdout-validation testing.

### Decision Tree

Given the ease of visualization, and interpretation decision tree can allow us to get a good understanding of our data, though it is prone to high variance, which can cause the model to perform quite poorly.

The pruned classification tree can be see in Appendix B: Figure 8.

- Entropy – split at the top most information gained is variable Contract.

From our classification tree, customer will churn (Yes at end node) if:

1. Client has one year contract, not using DSL internet service, and tenuring shorter than 14.5.
2. Client has one year contract, not using DSL internet service, tenuring longer than 14.5, using electronic check, and with total charges less than 3076.67.

This is in line with what we observed with visualization and logic:

* Clients with one-year contract are more likely to churn compared to two-year contracts.
* Clients using DSL are less likely to churn.
* Recent clients are more likely to churn.

### Random Forest

To relief the classification tree problem of high variance, we also applied random forest (which is a collection of trees) to our dataset. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. In this model, we use 500 trees and mtry of 2 in our random forests.

And then we build a confusion matrix to calculate the accuracy and sensitivity of our testing data. The result is pretty good, and the accuracy is 0.7809111. Random forest also allows us to compute the importance of each variance (Appendix B: Figure 9). From the figure, we able to observe the follow variables has the highest importance: tenure, charges, contract, payment method and internet service. This in general goes in line with difference in churn in our data visualization.

### Linear SVM

Lastly, we use the algorithms of Linear SVM. SVM is to find a plane that has the maximum margin. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. However, since our data is not linearly separable, we applied a loss penalty of 0.01.

We use the linear SVM algorithms to calculate the accuracy and sensitivity of our testing data. From the confusion matrix we can get the accuracy is 0.7968185 with the sensitivity of 0.5055249. And then we use the hyperparameter tuning to find the best number for the loss penalty. The result shows that the best cost is 0.1 and the gamma is 0.03225806. After that, we apply the best SVM algorithms and the sensitivity rise to 0.5082873.

### Threshold Analysis

As we are attempting to identify customer that are going to churn, we thus need to focus on sensitivity metric compared to accuracy. As it is comparatively more expensive to acquire customer than retain customer, thus we are not as concern with false positive, but rather concerned with false negative. We would ideally like a model that is able to successful target all customer that are going to churn, and it should matter less if we have a higher number of false positive to us a telecommunication company. Thus, we should have a lower threshold value than 0.5, using a more objective method to set out threshold value. From Appendix B: Figure 10, we are going to use the intercept point of accuracy, sensitivity and specificity as our cutoff point for the evaluation of our models yielding a threshold value of 0.2893.

## Evaluation

In general, we can access the performance of our models using Confusion Matrix and Area Under the ROC Curve (AUC). The confusion matrix is a two by two table formed by four outcomes of a binary classifier, of which the four outcomes are True Positive, True Negative, False Positive, and False Negative. From this confusion matrix, we can derive the following metrics: accuracy and sensitivity. Sensitivity is a better metric to use when describing a model which shows the fraction of positives predicted correctly. As shown in Appendix B: Figure 11, logistic regression model clearly has a higher sensitivity ratio which means that this model can give us more users who keep using the company’s service given that the prediction is correct. The reason that SVM model is not listed is because we faced some difficulties with SVM model. We were not able to change the cutoff value for SVM model, because cutoff value is adjusted by deviations which is not a direct comparison with probabilities. So, because of that it becomes hard to directly compare SVM’s performance with other model using confusion matrix metrics.

AUC is another performance measurement which measures the quality of a probability estimate model. The larger AUC, the better quality the model is. According to our result (Appendix B: Figure 12), logistic regression also has the highest value which means it has higher chance of predict the correct probability.

Combine the above evaluation together, we believe that logistic regression model is the best prediction model for our purpose. To further evaluate how well our logistic regression model can help telecommunication industry retain customers and improve their revenue, more analysis need to be conducted, as the model only helps our client to choose which customer, the actual impact will ultimately depend on the effectiveness of the marketing campaign/promotion. Measurements such as Customer Life Value and Return on Marketing Investment can be applied to calculate and project the future changes in telecommunication business.

## Deployment

Through previous discussion of finding the best prediction model for evaluating customer churn, we will further focus on what telecommunication company could do to attract targeting users. The main purpose of the model building is to create a model that our client can use to target its existing customers. Using the results, we can then select customer from those predicted to churn (from highest probability to lowest probability) till we run out of budget. To measure the impact of the targeting, we can theoretically compare the customer lifetime value of the targeted customer before and after the campaign over couple month and see whether the return on marketing investment (ROMI) yields positive value.

Since our best model is logistic regression, we can further investigate the relationship of variables to churn. Based on the prediction model we built, we found the two of the most important variables that influence customers’ decisions on whether to churn or not are tenure and contract. Under this circumstance, the company can retain long-tenure customers by providing welfares like free on-site service, premium for contract renewal. Also, it will always help if company builds a close communication relationship with customers by sending letter, special offers, emails and follow-ups. Besides, since two-year contract customers are more likely to stay at current service, the companies can attach more preferential policy on this contract, expecting more new customers will go for two-year contract. For example, if customers choose two-year contract, they can also get a big reduction in total charges. Or, telecom companies can bundle their service with partner firms to provide extra benefit for customers.

The deployment process also has some potential risks. In order to keep in touch with customers, we will gather detailed and personal data from customers. With the high volume of information transmission, customers become more protective and vigilant about information safety. Shady practice like selling and buying personal information or using customer’s data without permission can make customers insecure and worry. Therefore, when gathering data to figure out potential targeting customers, telecommunication company need to avoid leakage of information and use it in a responsible way.

Another challenge telecom companies may face when using data retain customers is that the employees in companies may fail to put these suggestions into practice. If the employees keep thinking in the old way or keep implementing the old strategy to customers, the company may end up creating a fragmented customer experience.  In order to mitigate this risk, the company better builds a good communication system to constantly remind the employees of the necessity to carry out new loyalty system.

## Appendix A: Metadata Information

**customerID**: Customer ID

**gender**: Whether the customer is a male or a female

**SeniorCitizen**: Whether the customer is a senior citizen or not (1, 0)

**Partner**: Whether the customer has a partner or not (Yes, No)

**Dependents**: Whether the customer has dependents or not (Yes, No)

**Tenure**: Number of months the customer has stayed with the company

**PhoneService**: Whether the customer has a phone service or not (Yes, No)

**MultipleLines**: Whether the customer has multiple lines or not (Yes, No, No phone service)

**InternetService**: Customer’s internet service provider (DSL, Fiber optic, No)

**OnlineSecurity**: Whether the customer has online security or not (Yes, No, No internet service)

**OnlineBackup**: Whether the customer has online backup or not (Yes, No, No internet service)

**DeviceProtection**: Whether the customer has device protection or not (Yes, No, No internet service)

**TechSupport**: Whether the customer has tech support or not (Yes, No, No internet service)

**StreamingTV**: Whether the customer has streaming TV or not (Yes, No, No internet service)

**StreamingMovies**: Whether the customer has streaming movies or not (Yes, No, No internet service)

**Contract**: The contract term of the customer (Month-to-month, One year, Two year)

**PaperlessBilling**: Whether the customer has paperless billing or not (Yes, No)

**PaymentMethod**: The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

**MonthlyCharges**: The amount charged to the customer monthly

**TotalCharges**: The total amount charged to the customer

**Churn**: Whether the customer churned or not (Yes or No)

## Appendix B: Figures & Tables

Figure 1: Missing Data Analysis

A screenshot of a cell phone

Description automatically generated

A picture containing screenshot

Description automatically generated

Figure 2: Distribution of Churn

A screenshot of a cell phone

Description automatically generated

Figure 3: Binary variable categorical variable distributionA screenshot of a cell phone

Description automatically generated

Figure 4: 3 factor level categorical variable distribution

A screenshot of a cell phone

Description automatically generated

Figure 5: 4 factor level categorical variable distribution

A screenshot of a cell phone

Description automatically generated

Figure 6: Continuous variable distribution scatterplot

A close up of a map

Description automatically generated

Figure 7: Logistic Regression summary output

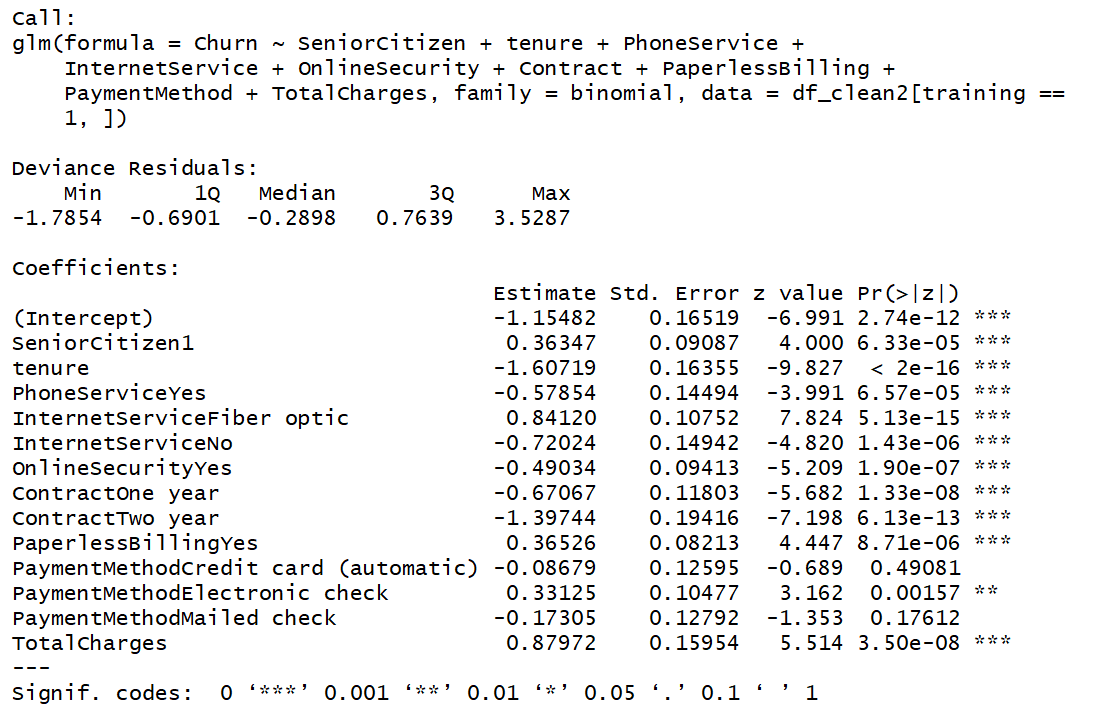


Figure 8: Pruned Decision Tree

A picture containing screenshot

Description automatically generated

Figure 9: Variable Importance Chart

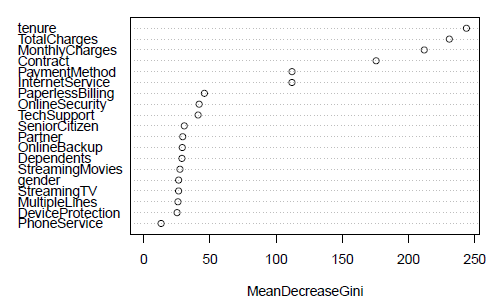


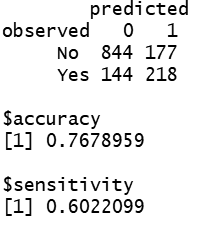
Figure 10: Threshold Analysis

A close up of a map

Description automatically generated

Figure 11: Confusion Matrix for Logistic Regression, Decision Tree, Random Forest

|  |  |  |
| --- | --- | --- |
| Logistic Regression | Decision Tree | Random Forest |



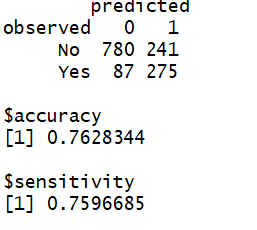
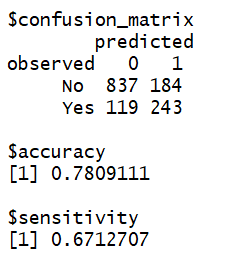


Figure 12: ROC Curve

A close up of a map

Description automatically generated



## Appendix C: Rmd Code file

Data Science and Business Intel Project

library(tidyverse)  
library(ggplot2)  
library(ggROC)  
library(cowplot)  
library(reshape2)  
library(car)  
library(caret)  
library(leaps)  
library(bestglm)  
library(plotly)  
library(webshot)  
library(DataExplorer)  
library(purrr)  
library(rpart)  
library(rpart.plot)  
library(randomForest)  
library(e1071)  
library(pROC)  
source("./[5]Script/Confusion\_matrix.R")  
source("./[5]Script/cutoff.R")

# Read in raw data  
ds <- (read.csv("./[4]source/WA\_Fn-UseC\_-Telco-Customer-Churn.csv"))  
ds$SeniorCitizen <- as.factor(ds$SeniorCitizen)  
#####################################################################################  
# Metadata  
##################################################################################### Customer Churn: Whether customer has left within the last month   
# Service that each customer has signed up for  
# Demographic information   
# Customer Account Information  
#####################################################################################  
# Data Type:  
#####################################################################################  
# 16 Categorical Variables:  
# - 6 Binary Variables (Gender, Senior Citizen, Partner, Dependents, Phone Service, Paperless Billing)  
# - 9 3-Factor level Variable (Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract)  
# - 1 4-Factor level Variable (Payment Method)  
#####################################################################################  
# 3 Continious Variables:  
# - Tenure, Monthly Charge, Total Charge  
#####################################################################################  
# 1 Target Variables:  
# - Churn  
#####################################################################################

## Data Cleaning

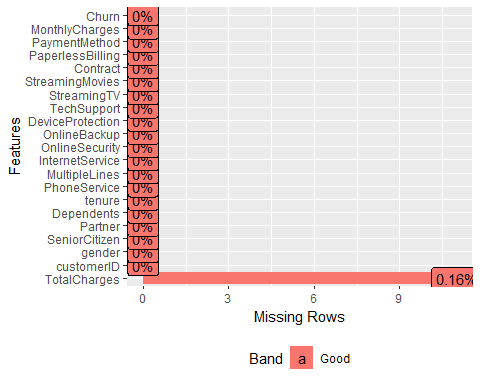
# missing data analysis  
# 1) We first check if missing exist within our dataset  
print(paste0("The dataset contains missing data: ", any(is.na(ds))))

## [1] "The dataset contains missing data: TRUE"

if (any(is.na(ds)) == "TRUE"){  
 print(paste0("The total number of missing data(s) are: ", sum(is.na(ds))))  
 print(paste0("The variable(s) with missing data(s) are: ", colnames(ds)[colSums(is.na(ds))>0]))  
}

## [1] "The total number of missing data(s) are: 11"  
## [1] "The variable(s) with missing data(s) are: TotalCharges"

plot\_missing(ds)



# 2) Filter the missing data into a its own dataset for further analysis  
df\_na <- ds[rowSums(is.na(ds))>0,]  
df\_na[c("gender","tenure","PhoneService","InternetService","Contract","MonthlyCharges","TotalCharges","Churn")]

## gender tenure PhoneService InternetService Contract MonthlyCharges  
## 489 Female 0 No DSL Two year 52.55  
## 754 Male 0 Yes No Two year 20.25  
## 937 Female 0 Yes DSL Two year 80.85  
## 1083 Male 0 Yes No Two year 25.75  
## 1341 Female 0 No DSL Two year 56.05  
## 3332 Male 0 Yes No Two year 19.85  
## 3827 Male 0 Yes No Two year 25.35  
## 4381 Female 0 Yes No Two year 20.00  
## 5219 Male 0 Yes No One year 19.70  
## 6671 Female 0 Yes DSL Two year 73.35  
## 6755 Male 0 Yes DSL Two year 61.90  
## TotalCharges Churn  
## 489 NA No  
## 754 NA No  
## 937 NA No  
## 1083 NA No  
## 1341 NA No  
## 3332 NA No  
## 3827 NA No  
## 4381 NA No  
## 5219 NA No  
## 6671 NA No  
## 6755 NA No

From the above missing data analysis, we are able to see out of the 7043 observation of 21 variables there are only 11 missing values and they are belong to the TOTAL CHARGES column(.16%), hence we are working with a pretty clean dataset.

An possible explaination for this mssing values is: (1) These customer never paid anything to the company (2) Tenure for all these customer are 0, thus meaning that this may be their first month with the company and thus the company hasn’t charged them.

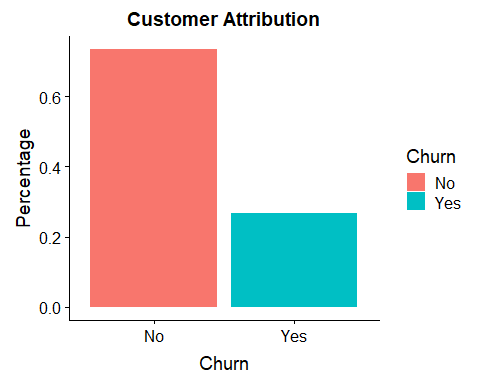
For these 11 missing data, we can either: (1) Impute the total charge value (2) Set total charge value to be zero (3) Remove them from the data set

Since we have a relatively large dataset, and that none of the customer with missing value have churn, thus for convience of the analysis, we will drop the 11 observation with missing TOTAL CHARGE. ## Data Exploration

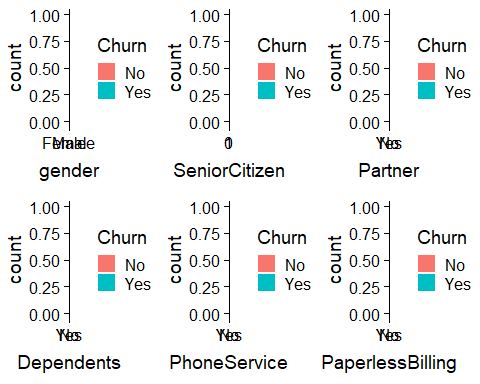
df\_clean <- ds %>%   
 na.omit() %>%   
 select(-1)  
  
summary(df\_clean)

## gender SeniorCitizen Partner Dependents tenure   
## Female:3483 0:5890 No :3639 No :4933 Min. : 1.00   
## Male :3549 1:1142 Yes:3393 Yes:2099 1st Qu.: 9.00   
## Median :29.00   
## Mean :32.42   
## 3rd Qu.:55.00   
## Max. :72.00   
## PhoneService MultipleLines InternetService  
## No : 680 No :3385 DSL :2416   
## Yes:6352 No phone service: 680 Fiber optic:3096   
## Yes :2967 No :1520   
##   
##   
##   
## OnlineSecurity OnlineBackup   
## No :3497 No :3087   
## No internet service:1520 No internet service:1520   
## Yes :2015 Yes :2425   
##   
##   
##   
## DeviceProtection TechSupport   
## No :3094 No :3472   
## No internet service:1520 No internet service:1520   
## Yes :2418 Yes :2040   
##   
##   
##   
## StreamingTV StreamingMovies  
## No :2809 No :2781   
## No internet service:1520 No internet service:1520   
## Yes :2703 Yes :2731   
##   
##   
##   
## Contract PaperlessBilling PaymentMethod   
## Month-to-month:3875 No :2864 Bank transfer (automatic):1542   
## One year :1472 Yes:4168 Credit card (automatic) :1521   
## Two year :1685 Electronic check :2365   
## Mailed check :1604   
##   
##   
## MonthlyCharges TotalCharges Churn   
## Min. : 18.25 Min. : 18.8 No :5163   
## 1st Qu.: 35.59 1st Qu.: 401.4 Yes:1869   
## Median : 70.35 Median :1397.5   
## Mean : 64.80 Mean :2283.3   
## 3rd Qu.: 89.86 3rd Qu.:3794.7   
## Max. :118.75 Max. :8684.8

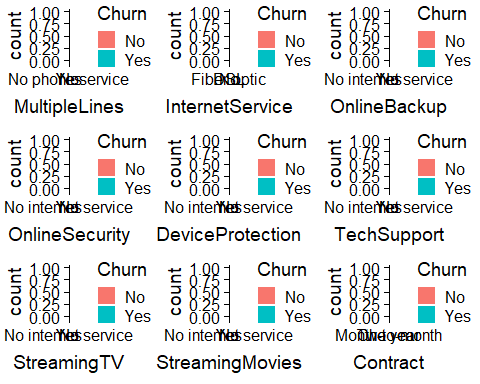
# Binary variable distribution in Customer attribution  
ggplot(data = df\_clean, aes(x = Churn, y = (..count..)/sum(..count..), fill = Churn))+  
 geom\_bar()+  
 ggtitle("Customer Attribution")+  
 ylab("Percentage")

 Of our dataset, 26% of the customer has left the platform within the past month

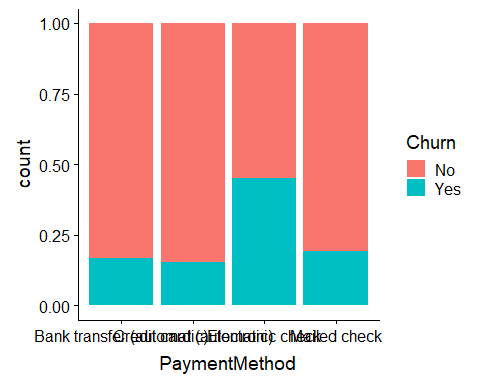
# Categorical Variable Analysis  
# Binary binary variables Analysis  
options(repr.plot.width = 12, repr.plot.height = 8)  
plot\_grid(  
 ggplot(data = df\_clean, aes(gender, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(SeniorCitizen, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(Partner, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(Dependents, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(PhoneService, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(PaperlessBilling, fill = Churn))+geom\_bar(position = "fill")  
)



plot\_grid(  
 ggplot(data = df\_clean, aes(MultipleLines, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(InternetService, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(OnlineBackup, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(OnlineSecurity, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(DeviceProtection, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(TechSupport, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(StreamingTV, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(StreamingMovies, fill = Churn))+geom\_bar(position = "fill"),  
 ggplot(data = df\_clean, aes(Contract, fill = Churn))+geom\_bar(position = "fill")  
)

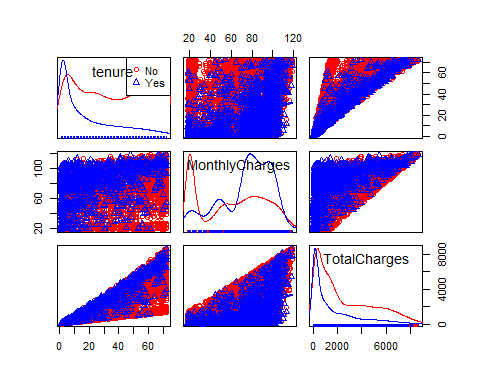


ggplot(data = df\_clean, aes(x=PaymentMethod, fill=Churn))+  
 geom\_bar(position = "fill")

 Some Trends - Senior Citizens churn percentage are higher - Customer with dependents or partners tend to have lower churn rate compared to counterparts - Customer with paperless billing have higher churn rate - Customer with Fiber Optic Internet Service have significant higher churn rate - Customer with No online security, or online backup or tech support have higher churn rate - Customer with monthly subscription are more likely to churn compared to customer with one- or two-year contract - Customer with Electronic Check payment method tend to leave our client more compared to other options.

# # Continous Variable Analysis  
# p <- plot\_ly(df\_clean,  
# x = ~MonthlyCharges,  
# y = ~TotalCharges,  
# z = ~tenure,  
# color = ~Churn,  
# marker = list(  
# size = 2)) %>%  
# add\_markers() %>%  
# layout(scene = list(  
# xaxis = list(title = "Monthly Charges"),  
# yaxis = list(title = "Total Charges"),  
# zaxis = list(title = "Tenure")  
# ))  
# p

# Correlation matrix of continous variable analysis (thank you very much)  
scatterplotMatrix(~ tenure + MonthlyCharges + TotalCharges|Churn, data = df\_clean, col = c("red","blue"))

 This appears to follow simple intuition, customer that are with the telecommunication firm for only a short period would in general have no loyalty compared to long time customer which are comfortable with the service provided and thus less willing to switch. Also, customer with higher monthly charges, will in general wish to reduce cost by seeking alternative service provider that may provide the same level of service for lower cost. From Figure 6, the scatter plots between continuous variables in general follows the trend described above and allow us to visually see if there are any obvious outliers which in this case seem to be none.

# Data Cleaning and Standardization  
df\_clean2 <- df\_clean %>%   
 mutate(MultipleLines=replace(MultipleLines,MultipleLines=="No phone service", "No")) %>%  
 mutate(OnlineSecurity=replace(OnlineSecurity,OnlineSecurity=="No internet service","No")) %>%   
 mutate(DeviceProtection=replace(DeviceProtection,DeviceProtection=="No internet service","No")) %>%   
 mutate(TechSupport=replace(TechSupport,TechSupport=="No internet service","No")) %>%   
 mutate(StreamingTV=replace(StreamingTV,StreamingTV=="No internet service","No")) %>%   
 mutate(StreamingMovies=replace(StreamingMovies,StreamingMovies=="No internet service","No")) %>%  
 mutate(OnlineBackup=replace(OnlineBackup,OnlineBackup=="No internet service","No")) %>%   
 mutate(tenure=scale(tenure)) %>%   
 mutate(MonthlyCharges=scale(MonthlyCharges)) %>%   
 mutate(TotalCharges=scale(TotalCharges))

# Split data into training and validation split  
set.seed(1994)  
training <- sample(2,nrow(df\_clean2),replace=TRUE,prob=c(.8,.2))

# GLM Analysis  
# Still need to look at threshold analysis  
# Full GLM  
df\_clean.fulllogit <- glm(Churn~.,   
 family = binomial,  
 data = df\_clean2[training==1,])  
  
getinfo(df\_clean.fulllogit,df\_clean2)[c("confusion\_matrix", "accuracy","sensitivity")]

## $confusion\_matrix  
## predicted  
## observed 0 1  
## No 770 251  
## Yes 84 278  
##   
## $accuracy  
## [1] 0.757773  
##   
## $sensitivity  
## [1] 0.7679558

# Using Forward Approach to search for GLM model with lowest BIC   
  
# tmp.modelsearch <- bestglm(df\_clean2[training==1,],IC = "BIC", family = binomial, method = "forward")  
  
  
# Takes a long while (>= 4 to 6 hours)  
# tmp.modelsearch$BestModels  
# tmp.modelsearch$BestModel  
  
# Best GLM Model  
df\_clean.bestlogit <- glm(Churn~   
 SeniorCitizen +   
 tenure +   
 PhoneService +   
 InternetService +   
 OnlineSecurity +   
 Contract +   
 PaperlessBilling +   
 PaymentMethod +   
 TotalCharges,   
 family = binomial,  
 data = df\_clean2[training==1,])  
  
summary(df\_clean.bestlogit)

##   
## Call:  
## glm(formula = Churn ~ SeniorCitizen + tenure + PhoneService +   
## InternetService + OnlineSecurity + Contract + PaperlessBilling +   
## PaymentMethod + TotalCharges, family = binomial, data = df\_clean2[training ==   
## 1, ])  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7854 -0.6901 -0.2898 0.7639 3.5287   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.15482 0.16519 -6.991 2.74e-12  
## SeniorCitizen1 0.36347 0.09087 4.000 6.33e-05  
## tenure -1.60719 0.16355 -9.827 < 2e-16  
## PhoneServiceYes -0.57854 0.14494 -3.991 6.57e-05  
## InternetServiceFiber optic 0.84120 0.10752 7.824 5.13e-15  
## InternetServiceNo -0.72024 0.14942 -4.820 1.43e-06  
## OnlineSecurityYes -0.49034 0.09413 -5.209 1.90e-07  
## ContractOne year -0.67067 0.11803 -5.682 1.33e-08  
## ContractTwo year -1.39744 0.19416 -7.198 6.13e-13  
## PaperlessBillingYes 0.36526 0.08213 4.447 8.71e-06  
## PaymentMethodCredit card (automatic) -0.08679 0.12595 -0.689 0.49081  
## PaymentMethodElectronic check 0.33125 0.10477 3.162 0.00157  
## PaymentMethodMailed check -0.17305 0.12792 -1.353 0.17612  
## TotalCharges 0.87972 0.15954 5.514 3.50e-08  
##   
## (Intercept) \*\*\*  
## SeniorCitizen1 \*\*\*  
## tenure \*\*\*  
## PhoneServiceYes \*\*\*  
## InternetServiceFiber optic \*\*\*  
## InternetServiceNo \*\*\*  
## OnlineSecurityYes \*\*\*  
## ContractOne year \*\*\*  
## ContractTwo year \*\*\*  
## PaperlessBillingYes \*\*\*  
## PaymentMethodCredit card (automatic)   
## PaymentMethodElectronic check \*\*   
## PaymentMethodMailed check   
## TotalCharges \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6553.1 on 5648 degrees of freedom  
## Residual deviance: 4715.9 on 5635 degrees of freedom  
## AIC: 4743.9  
##   
## Number of Fisher Scoring iterations: 6

vif(df\_clean.bestlogit)

## GVIF Df GVIF^(1/(2\*Df))  
## SeniorCitizen 1.080908 1 1.039667  
## tenure 14.617958 1 3.823344  
## PhoneService 1.401480 1 1.183841  
## InternetService 2.402199 2 1.244951  
## OnlineSecurity 1.127523 1 1.061849  
## Contract 1.521245 2 1.110580  
## PaperlessBilling 1.113757 1 1.055347  
## PaymentMethod 1.343729 3 1.050474  
## TotalCharges 16.492869 1 4.061141

getinfo(df\_clean.bestlogit,df\_clean2)[c("confusion\_matrix", "accuracy", "sensitivity")]

## $confusion\_matrix  
## predicted  
## observed 0 1  
## No 773 248  
## Yes 82 280  
##   
## $accuracy  
## [1] 0.7613883  
##   
## $sensitivity  
## [1] 0.7734807

# Remove Total Charges due to high VIF value (>2, thus multi-colinearity effect)  
  
df\_clean.bestlogit2 <- glm(Churn~   
 SeniorCitizen +   
 tenure +   
 PhoneService +   
 InternetService +   
 OnlineSecurity +   
 Contract +   
 PaperlessBilling +  
 PaymentMethod,   
 family = binomial,  
 data = df\_clean2[training==1,])  
summary(df\_clean.bestlogit2)

##   
## Call:  
## glm(formula = Churn ~ SeniorCitizen + tenure + PhoneService +   
## InternetService + OnlineSecurity + Contract + PaperlessBilling +   
## PaymentMethod, family = binomial, data = df\_clean2[training ==   
## 1, ])  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8218 -0.6731 -0.3068 0.7653 3.1154   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.31487 0.15882 -8.279 < 2e-16  
## SeniorCitizen1 0.37046 0.09132 4.057 4.98e-05  
## tenure -0.78160 0.05612 -13.927 < 2e-16  
## PhoneServiceYes -0.40597 0.13731 -2.957 0.003111  
## InternetServiceFiber optic 1.07104 0.09861 10.862 < 2e-16  
## InternetServiceNo -0.85564 0.14626 -5.850 4.91e-09  
## OnlineSecurityYes -0.44134 0.09379 -4.706 2.53e-06  
## ContractOne year -0.63739 0.11659 -5.467 4.58e-08  
## ContractTwo year -1.30311 0.19059 -6.837 8.07e-12  
## PaperlessBillingYes 0.37475 0.08173 4.585 4.53e-06  
## PaymentMethodCredit card (automatic) -0.08768 0.12574 -0.697 0.485626  
## PaymentMethodElectronic check 0.35246 0.10472 3.366 0.000763  
## PaymentMethodMailed check -0.11915 0.12660 -0.941 0.346609  
##   
## (Intercept) \*\*\*  
## SeniorCitizen1 \*\*\*  
## tenure \*\*\*  
## PhoneServiceYes \*\*   
## InternetServiceFiber optic \*\*\*  
## InternetServiceNo \*\*\*  
## OnlineSecurityYes \*\*\*  
## ContractOne year \*\*\*  
## ContractTwo year \*\*\*  
## PaperlessBillingYes \*\*\*  
## PaymentMethodCredit card (automatic)   
## PaymentMethodElectronic check \*\*\*  
## PaymentMethodMailed check   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6553.1 on 5648 degrees of freedom  
## Residual deviance: 4748.6 on 5636 degrees of freedom  
## AIC: 4774.6  
##   
## Number of Fisher Scoring iterations: 6

vif(df\_clean.bestlogit2)

## GVIF Df GVIF^(1/(2\*Df))  
## SeniorCitizen 1.081676 1 1.040037  
## tenure 1.662398 1 1.289340  
## PhoneService 1.358571 1 1.165577  
## InternetService 1.924459 2 1.177815  
## OnlineSecurity 1.113718 1 1.055328  
## Contract 1.481751 2 1.103300  
## PaperlessBilling 1.108524 1 1.052865  
## PaymentMethod 1.318029 3 1.047098

getinfo(df\_clean.bestlogit2,df\_clean2)[c("confusion\_matrix", "accuracy", "sensitivity")]

## $confusion\_matrix  
## predicted  
## observed 0 1  
## No 780 241  
## Yes 87 275  
##   
## $accuracy  
## [1] 0.7628344  
##   
## $sensitivity  
## [1] 0.7596685

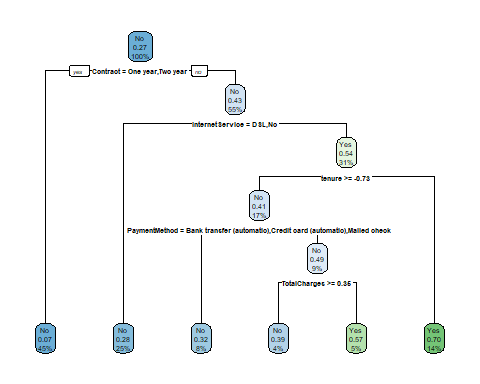
# Decision Tree Analysis  
df\_clean.fulltree <- rpart(Churn ~.,   
 data = df\_clean2[training==1,], method = "class",  
 control = rpart.control(cp=0))   
  
getinfo(df\_clean.fulltree,df\_clean2)[c("confusion\_matrix", "accuracy", "sensitivity")]

## $confusion\_matrix  
## predicted  
## observed 0 1  
## No 811 210  
## Yes 127 235  
##   
## $accuracy  
## [1] 0.7563268  
##   
## $sensitivity  
## [1] 0.6491713

# Hyperparameter Tuning  
  
# plotcp(df\_clean.fulltree)  
printcp(df\_clean.fulltree)

##   
## Classification tree:  
## rpart(formula = Churn ~ ., data = df\_clean2[training == 1, ],   
## method = "class", control = rpart.control(cp = 0))  
##   
## Variables actually used in tree construction:  
## [1] Contract Dependents DeviceProtection gender   
## [5] InternetService MonthlyCharges MultipleLines OnlineBackup   
## [9] OnlineSecurity PaperlessBilling Partner PaymentMethod   
## [13] PhoneService SeniorCitizen StreamingMovies StreamingTV   
## [17] TechSupport tenure TotalCharges   
##   
## Root node error: 1507/5649 = 0.26677  
##   
## n= 5649   
##   
## CP nsplit rel error xerror xstd  
## 1 0.07011723 0 1.00000 1.00000 0.022058  
## 2 0.01360319 3 0.78965 0.79761 0.020412  
## 3 0.00398142 5 0.76244 0.81287 0.020553  
## 4 0.00265428 10 0.73723 0.80027 0.020437  
## 5 0.00248839 18 0.71267 0.79429 0.020381  
## 6 0.00232250 22 0.70272 0.79429 0.020381  
## 7 0.00199071 31 0.67750 0.79496 0.020387  
## 8 0.00176952 32 0.67551 0.79695 0.020406  
## 9 0.00165893 49 0.63504 0.79695 0.020406  
## 10 0.00149303 61 0.61447 0.79628 0.020400  
## 11 0.00132714 65 0.60849 0.80226 0.020455  
## 12 0.00110595 75 0.59456 0.80624 0.020492  
## 13 0.00099536 81 0.58792 0.80823 0.020510  
## 14 0.00088476 91 0.57664 0.80956 0.020523  
## 15 0.00082946 107 0.55740 0.80956 0.020523  
## 16 0.00079628 113 0.55209 0.81221 0.020547  
## 17 0.00066357 120 0.54612 0.82681 0.020679  
## 18 0.00049768 126 0.54214 0.82681 0.020679  
## 19 0.00044238 137 0.53417 0.84472 0.020838  
## 20 0.00033179 140 0.53285 0.84472 0.020838  
## 21 0.00026543 142 0.53218 0.85468 0.020924  
## 22 0.00022119 147 0.53086 0.86662 0.021027  
## 23 0.00016589 150 0.53019 0.86662 0.021027  
## 24 0.00013271 159 0.52820 0.86662 0.021027  
## 25 0.00000000 164 0.52754 0.86662 0.021027

tmp <- df\_clean.fulltree$cptable[which.min(df\_clean.fulltree$cptable[,"xerror"]),]  
  
# Prune the tree  
df\_clean.besttree <- prune(df\_clean.fulltree,cp = 0.01)  
rpart.plot(df\_clean.besttree)



getinfo(df\_clean.besttree,df\_clean2)[c("confusion\_matrix", "accuracy", "sensitivity")]

## $confusion\_matrix  
## predicted  
## observed 0 1  
## No 844 177  
## Yes 144 218  
##   
## $accuracy  
## [1] 0.7678959  
##   
## $sensitivity  
## [1] 0.6022099

# Random Forest  
set.seed(1994)  
df\_clean.rforest <- randomForest(Churn~.,  
 data = df\_clean2[training==1,],  
 ntree=500, # dataset  
 cutoff=c(0.5,0.5),   
 mtry=2,  
 importance=TRUE)   
df\_clean.rforest

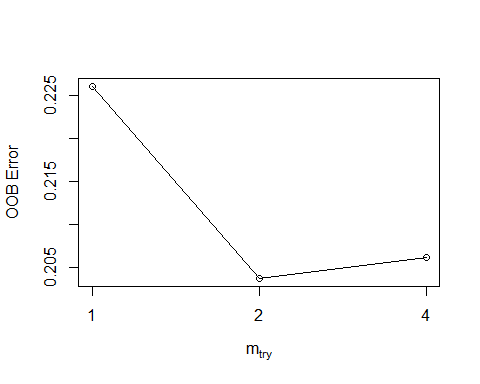
##   
## Call:  
## randomForest(formula = Churn ~ ., data = df\_clean2[training == 1, ], ntree = 500, cutoff = c(0.5, 0.5), mtry = 2, importance = TRUE)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 20.46%  
## Confusion matrix:  
## No Yes class.error  
## No 3789 353 0.08522453  
## Yes 803 704 0.53284672

# Confusion Matrix Test  
getinfo(df\_clean.rforest,df\_clean2)[c("confusion\_matrix", "accuracy", "sensitivity")]

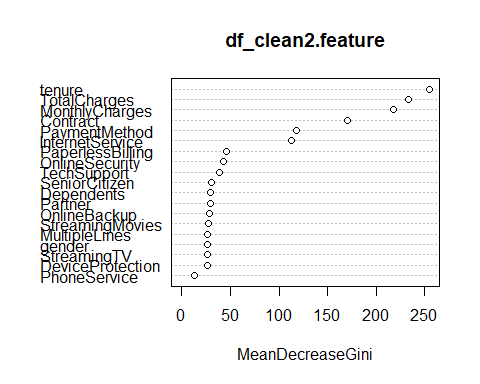
## $confusion\_matrix  
## predicted  
## observed 0 1  
## No 837 184  
## Yes 119 243  
##   
## $accuracy  
## [1] 0.7809111  
##   
## $sensitivity  
## [1] 0.6712707

# Hyperparameter Tuning  
set.seed(1994)  
rforest.tune <- tuneRF(x = df\_clean2[training==1,]%>%select(-Churn),  
 y = df\_clean2[training==1,]$Churn,mtryStart=2,  
 ntreeTry = 500)

## mtry = 2 OOB error = 20.38%   
## Searching left ...  
## mtry = 1 OOB error = 22.61%   
## -0.10947 0.05   
## Searching right ...  
## mtry = 4 OOB error = 20.62%   
## -0.01216334 0.05



# Feature Importance Analysis  
# generateFilterValuesData(task, "randomForest.importance")  
df\_clean2.feature <- randomForest(Churn~., data = df\_clean2, importance = FALSE, ntree = 500, mtry = 2, do.trace=FALSE)  
  
varImpPlot(df\_clean2.feature)



# SVM  
df\_clean.svm <- svm(Churn~.,  
 data = df\_clean2[training==1,],  
 kernel = "linear",  
 cost = 0.01,  
 proability = TRUE)  
  
getinfo(df\_clean.svm,df\_clean2)[c("confusion\_matrix", "accuracy", "sensitivity")]

## $confusion\_matrix  
## predicted  
## observed No Yes  
## No 919 102  
## Yes 179 183  
##   
## $accuracy  
## [1] 0.7968185  
##   
## $sensitivity  
## [1] 0.5055249

# Hyperparameter Tuning  
svm.tune <- tune(svm,  
 Churn~.,  
 data = df\_clean2[training==1,],  
 kernel = "linear",  
 ranges = list(cost = 10^(-5:0)))  
   
print(svm.tune)

##   
## Parameter tuning of 'svm':  
##   
## - sampling method: 10-fold cross validation   
##   
## - best parameters:  
## cost  
## 1  
##   
## - best performance: 0.1993228

svm.tune$best.model

##   
## Call:  
## best.tune(method = svm, train.x = Churn ~ ., data = df\_clean2[training ==   
## 1, ], ranges = list(cost = 10^(-5:0)), kernel = "linear")  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 1   
## gamma: 0.03225806   
##   
## Number of Support Vectors: 2589

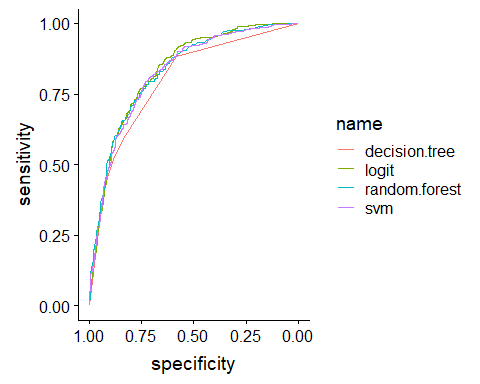
df\_clean.bestsvm <- svm(Churn~.,  
 data = df\_clean2[training==1,],  
 kernel = "linear",  
 cost = 0.1,  
 probaility = TRUE)  
summary(df\_clean.bestsvm)

##   
## Call:  
## svm(formula = Churn ~ ., data = df\_clean2[training == 1, ], kernel = "linear",   
## cost = 0.1, probaility = TRUE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 0.1   
## gamma: 0.03225806   
##   
## Number of Support Vectors: 2608  
##   
## ( 1308 1300 )  
##   
##   
## Number of Classes: 2   
##   
## Levels:   
## No Yes

getinfo(df\_clean.bestsvm,df\_clean2)[c("confusion\_matrix", "accuracy", "sensitivity")]

## $confusion\_matrix  
## predicted  
## observed No Yes  
## No 915 106  
## Yes 178 184  
##   
## $accuracy  
## [1] 0.7946493  
##   
## $sensitivity  
## [1] 0.5082873

# Performance evaluation - Learning Curves and Fitted Graphs  
# AUC Curve  
# First assemble the probability matrix  
prob\_matrix <- data.frame(  
 "logit" = predict(df\_clean.bestlogit2,df\_clean2[training==2,],type = "response"),  
 "d\_tree" = predict(df\_clean.besttree, df\_clean2[training==2,],type="prob")[,2],  
 "r\_forest" = predict(df\_clean.rforest, df\_clean2[training==2,], type = "prob")[,2],  
 "svm" = as.numeric(attr(predict(df\_clean.bestsvm, df\_clean2[training==2,], decision.values = TRUE),"decision.values"))  
 )  
# Create the ROC Varible  
  
logit.roc <- roc(df\_clean2$Churn[training==2],prob\_matrix$logit)  
d\_tree.roc <- roc(df\_clean2$Churn[training==2],prob\_matrix$d\_tree)  
r\_forest.roc <- roc(df\_clean2$Churn[training==2],prob\_matrix$r\_forest)  
svm.roc <- roc(df\_clean2$Churn[training==2],prob\_matrix$svm)  
  
ggroc(list(logit=logit.roc,decision.tree=d\_tree.roc,random.forest=r\_forest.roc,svm=svm.roc),legacy.axes = FALSE)



tmp4 <- c(logit.roc$auc,d\_tree.roc$auc,r\_forest.roc$auc,svm.roc$auc)  
  
tmp5 <- data.frame(  
 "AUC" = tmp4  
)  
row.names(tmp5)<- c("Logit","Decision Tree", "Random Forest","SVM")  
  
tmp5

## AUC  
## Logit 0.8358193  
## Decision Tree 0.7981680  
## Random Forest 0.8314863  
## SVM 0.8270775

A small discussion about cutoff point:

As we are attempting to identify customer that are going to churn, we thus need to focus on sensitivity metric compared to accuracy. As it is comparitively more expensive to acquire customer than retain customer, thus we are not as concern with false positive, but rather concerned with false negative. We would idealy like a model that is able to successful target all customer that are going to churn, and it should matter less if we have a higher number of false postive to us a telcommunication company. Thus we should have a lower threshold value than 0.5, though the actual value often require domain knowledge which we lack, thus we are going to use a more objective method to set out threshold value.

# Lets use logistic regression as it has the largest AUC out of all three method  
  
# output <- matrix(0,100,3)  
# x\_axis <- seq(0.01,0.8,length=100)  
#   
# for (i in 1:100)  
# {  
# output[i,]=threshold(x\_axis[i])  
# }  
#   
# plot(x\_axis,output[,1], type = "l", col = "darkgreen", xlab = "Threshold Value", ylab = "Values")  
# lines(x\_axis,output[,2],col = "red")  
# lines(x\_axis,output[,3], col = "blue")  
# legend("bottom",col=c(2,"darkgreen",4,"darkred"),text.font =3,inset = 0.02,  
# box.lty=0,cex = 0.8,  
# lwd=c(2,2,2,2),c("Specificity","Senitivity","Accuracy"))  
#   
#   
# x\_axis[which(abs(output[,1]-output[,2])<0.01)]