Project - Telco ABC

Abhishek Suragani

2022-05-01

#1. Data Exploration

#set working directory  
#read dataset (dataset in google drive -> workspace)  
telco <- read.csv("IBM\_Telco-Customer-Churn.csv")  
head(telco)

## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## 1 7590-VHVEG Female 0 Yes No 1 No  
## 2 5575-GNVDE Male 0 No No 34 Yes  
## 3 3668-QPYBK Male 0 No No 2 Yes  
## 4 7795-CFOCW Male 0 No No 45 No  
## 5 9237-HQITU Female 0 No No 2 Yes  
## 6 9305-CDSKC Female 0 No No 8 Yes  
## MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection  
## 1 No phone service DSL No Yes No  
## 2 No DSL Yes No Yes  
## 3 No DSL Yes Yes No  
## 4 No phone service DSL Yes No Yes  
## 5 No Fiber optic No No No  
## 6 Yes Fiber optic No No Yes  
## TechSupport StreamingTV StreamingMovies Contract PaperlessBilling  
## 1 No No No Month-to-month Yes  
## 2 No No No One year No  
## 3 No No No Month-to-month Yes  
## 4 Yes No No One year No  
## 5 No No No Month-to-month Yes  
## 6 No Yes Yes Month-to-month Yes  
## PaymentMethod MonthlyCharges TotalCharges Churn  
## 1 Electronic check 29.85 29.85 No  
## 2 Mailed check 56.95 1889.50 No  
## 3 Mailed check 53.85 108.15 Yes  
## 4 Bank transfer (automatic) 42.30 1840.75 No  
## 5 Electronic check 70.70 151.65 Yes  
## 6 Electronic check 99.65 820.50 Yes

#get familiar with the dataset  
str(telco)

## 'data.frame': 7043 obs. of 21 variables:  
## $ customerID : chr "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...  
## $ gender : chr "Female" "Male" "Male" "Male" ...  
## $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : chr "Yes" "No" "No" "No" ...  
## $ Dependents : chr "No" "No" "No" "No" ...  
## $ tenure : int 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : chr "No" "Yes" "Yes" "No" ...  
## $ MultipleLines : chr "No phone service" "No" "No" "No phone service" ...  
## $ InternetService : chr "DSL" "DSL" "DSL" "DSL" ...  
## $ OnlineSecurity : chr "No" "Yes" "Yes" "Yes" ...  
## $ OnlineBackup : chr "Yes" "No" "Yes" "No" ...  
## $ DeviceProtection: chr "No" "Yes" "No" "Yes" ...  
## $ TechSupport : chr "No" "No" "No" "Yes" ...  
## $ StreamingTV : chr "No" "No" "No" "No" ...  
## $ StreamingMovies : chr "No" "No" "No" "No" ...  
## $ Contract : chr "Month-to-month" "One year" "Month-to-month" "One year" ...  
## $ PaperlessBilling: chr "Yes" "No" "Yes" "No" ...  
## $ PaymentMethod : chr "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : chr "No" "No" "Yes" "No" ...

summary(telco)

## customerID gender SeniorCitizen Partner   
## Length:7043 Length:7043 Min. :0.0000 Length:7043   
## Class :character Class :character 1st Qu.:0.0000 Class :character   
## Mode :character Mode :character Median :0.0000 Mode :character   
## Mean :0.1621   
## 3rd Qu.:0.0000   
## Max. :1.0000   
##   
## Dependents tenure PhoneService MultipleLines   
## Length:7043 Min. : 0.00 Length:7043 Length:7043   
## Class :character 1st Qu.: 9.00 Class :character Class :character   
## Mode :character Median :29.00 Mode :character Mode :character   
## Mean :32.37   
## 3rd Qu.:55.00   
## Max. :72.00   
##   
## InternetService OnlineSecurity OnlineBackup DeviceProtection   
## Length:7043 Length:7043 Length:7043 Length:7043   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## TechSupport StreamingTV StreamingMovies Contract   
## Length:7043 Length:7043 Length:7043 Length:7043   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## PaperlessBilling PaymentMethod MonthlyCharges TotalCharges   
## Length:7043 Length:7043 Min. : 18.25 Min. : 18.8   
## Class :character Class :character 1st Qu.: 35.50 1st Qu.: 401.4   
## Mode :character Mode :character Median : 70.35 Median :1397.5   
## Mean : 64.76 Mean :2283.3   
## 3rd Qu.: 89.85 3rd Qu.:3794.7   
## Max. :118.75 Max. :8684.8   
## NA's :11   
## Churn   
## Length:7043   
## Class :character   
## Mode :character   
##   
##   
##   
##

#check missing value  
sum(is.na(telco)) #check counts of NAs

## [1] 11

colSums(is.na(telco)) #check NAs in each column

## customerID gender SeniorCitizen Partner   
## 0 0 0 0   
## Dependents tenure PhoneService MultipleLines   
## 0 0 0 0   
## InternetService OnlineSecurity OnlineBackup DeviceProtection   
## 0 0 0 0   
## TechSupport StreamingTV StreamingMovies Contract   
## 0 0 0 0   
## PaperlessBilling PaymentMethod MonthlyCharges TotalCharges   
## 0 0 0 11   
## Churn   
## 0

which(is.na(telco$TotalCharges)) #check NAs in TotalCharges column

## [1] 489 754 937 1083 1341 3332 3827 4381 5219 6671 6755

telco <- na.omit(telco) #since only 11 NAs in TotalCharges, we will delete the rows

#convert variables  
#variables gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, Churn will be converted   
  
chr <- c("gender", "SeniorCitizen", "Partner", "Dependents", "PhoneService", "MultipleLines", "InternetService", "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "StreamingTV", "StreamingMovies", "Contract", "PaperlessBilling", "PaymentMethod", "Churn")  
telco[, chr] <- lapply(telco[, chr], factor)   
str(telco) #check if they all converted to factor

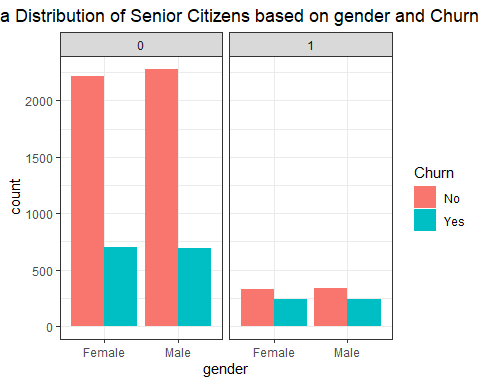
## 'data.frame': 7032 obs. of 21 variables:  
## $ customerID : chr "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ tenure : int 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...  
## $ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...  
## $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...  
## $ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...  
## $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:11] 489 754 937 1083 1341 3332 3827 4381 5219 6671 ...  
## ..- attr(\*, "names")= chr [1:11] "489" "754" "937" "1083" ...

summary(telco) #check out quantities in variables

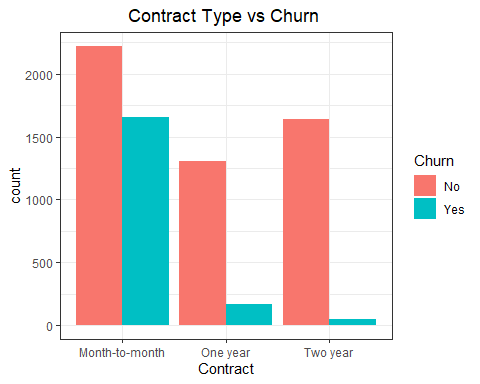
## customerID gender SeniorCitizen Partner Dependents  
## Length:7032 Female:3483 0:5890 No :3639 No :4933   
## Class :character Male :3549 1:1142 Yes:3393 Yes:2099   
## Mode :character   
##   
##   
##   
## tenure PhoneService MultipleLines InternetService  
## Min. : 1.00 No : 680 No :3385 DSL :2416   
## 1st Qu.: 9.00 Yes:6352 No phone service: 680 Fiber optic:3096   
## Median :29.00 Yes :2967 No :1520   
## Mean :32.42   
## 3rd Qu.:55.00   
## Max. :72.00   
## 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 Contract   
## No :2809 No :2781 Month-to-month:3875   
## No internet service:1520 No internet service:1520 One year :1472   
## Yes :2703 Yes :2731 Two year :1685   
##   
##   
##   
## PaperlessBilling PaymentMethod MonthlyCharges   
## No :2864 Bank transfer (automatic):1542 Min. : 18.25   
## Yes:4168 Credit card (automatic) :1521 1st Qu.: 35.59   
## Electronic check :2365 Median : 70.35   
## Mailed check :1604 Mean : 64.80   
## 3rd Qu.: 89.86   
## Max. :118.75   
## TotalCharges Churn   
## Min. : 18.8 No :5163   
## 1st Qu.: 401.4 Yes:1869   
## Median :1397.5   
## Mean :2283.3   
## 3rd Qu.:3794.7   
## Max. :8684.8

#2. Data Inspection(Visualizations)

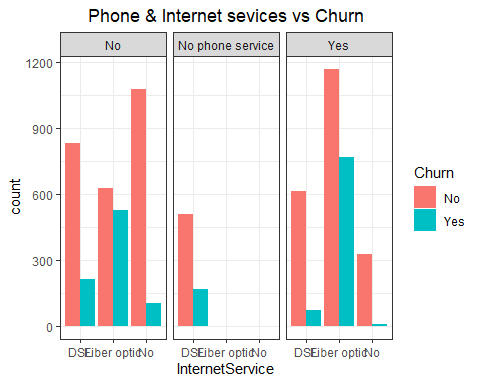
#gender and SeniorCitizen against churn   
telco %>%  
 group\_by(gender,SeniorCitizen, Churn) %>%  
 summarise(count = n()) %>%   
 ggplot(aes(x = gender, y = count, fill = Churn)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 facet\_wrap(~SeniorCitizen) + theme\_bw() +   
 ggtitle("Data Distribution of Senior Citizens based on gender and Churn") +  
 theme(plot.title = element\_text(hjust=0.5))



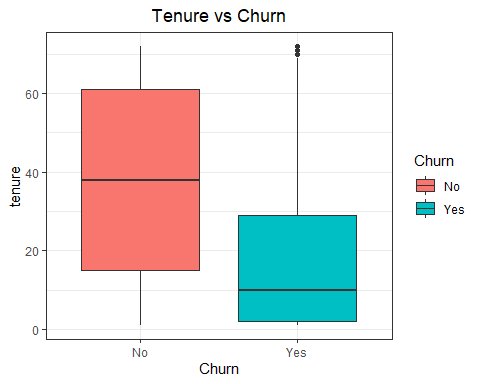
#contract against churn   
telco %>%  
 group\_by(Contract, Churn) %>%  
 summarise(count = n()) %>%   
 ggplot(aes(x = Contract, y = count, fill = Churn)) +  
 geom\_bar(stat = "identity", position = "dodge") + theme\_bw() +   
 ggtitle("Contract Type vs Churn ") +  
 theme(plot.title = element\_text(hjust=0.5))



#phone and internet services against churn   
telco %>%  
 group\_by(MultipleLines, InternetService, Churn) %>%  
 summarise(count = n()) %>%   
 ggplot(aes(x = InternetService, y = count, fill = Churn)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 facet\_wrap(~MultipleLines) + theme\_bw() +   
 ggtitle("Phone & Internet sevices vs Churn") +  
 theme(plot.title = element\_text(hjust=0.5))



#tenure against churn  
ggplot(data = telco) +  
 geom\_boxplot(aes(x=Churn, y=tenure, fill = Churn)) + theme\_bw() +   
 ggtitle("Tenure vs Churn") +  
 theme(plot.title = element\_text(hjust=0.5))



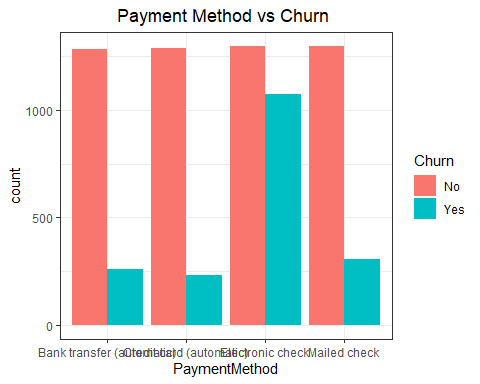
mean(telco$tenure)

## [1] 32.42179

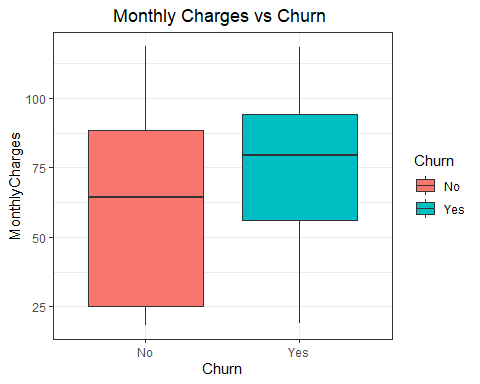
median(telco$tenure)

## [1] 29

#paymenym method vs. churn  
telco %>%  
 group\_by(PaymentMethod, Churn) %>%  
 summarise(count = n()) %>%   
 ggplot(aes(x = PaymentMethod, y = count, fill = Churn)) +  
 geom\_bar(stat = "identity", position = "dodge") + theme\_bw() +   
 ggtitle("Payment Method vs Churn ") +  
 theme(plot.title = element\_text(hjust=0.5))



#MonthlyCharges against churn  
ggplot(data = telco) +  
 geom\_boxplot(aes(x=Churn, y=MonthlyCharges, fill = Churn)) + theme\_bw() +   
 ggtitle("Monthly Charges vs Churn ") +  
 theme(plot.title = element\_text(hjust=0.5))



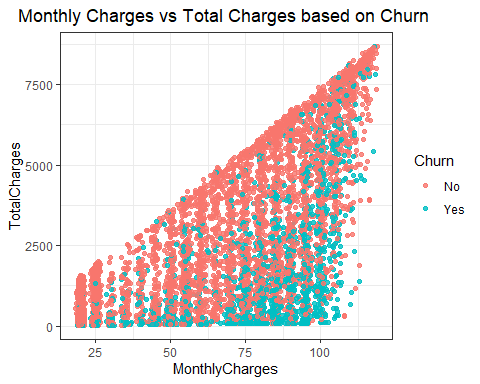
mean(telco$MonthlyCharges)

## [1] 64.79821

median(telco$MonthlyCharges)

## [1] 70.35

#charges vs. churn  
ggplot(aes(x = MonthlyCharges, y = TotalCharges, color = Churn), data = telco) +  
 geom\_point(alpha = 0.8) + theme\_bw() +   
 ggtitle("Monthly Charges vs Total Charges based on Churn ") +  
 theme(plot.title = element\_text(hjust=0.5))



#IBM Attrition will be a good example for us to analyze the dataset

#3. Model Specification

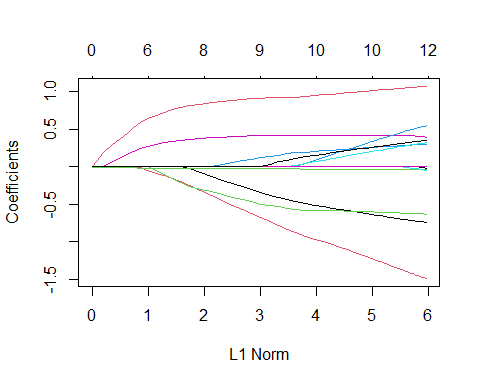
##3.1 Selecting Initial Set of Predictors based on Business Domain Rationale

Initial Set of Predictors: 1. Contract (Month to Month, One Year, Two Years) - There will always be unhappy customers who are likely to change the telecom service if the contract is month to month. 2. Tenure - If the Tenure is long, customer will usually be happy to understand and stick with the same operator even if there are some Issues whereas if Tenure is short customer will easily churn. 3. Monthly Charges - Amount charged to customer monthly has a direct Impact as the comparison will be made with other advertisers. 4. Senior Citizen - Senior Citizens will not be ready for the change. 5. Tech Support - Support team has a huge Impact as most of the companies lose customers if the support is not up-to the mark. 6. Payment Method (Bank Transfer, Card, Electronic Check, Mailed check) - cancellations will be less if the payment method is automatic. For example, any premium service that customers use as a part of auto- renewal process. 7. Internet Service - Type of Internet Service(DSL vs Fiber) will effect the speed of Internet and can result in churn rate because of the experience. 8. Multiple Lines - Having Multiple lines will have an Impact on decision making 9. StreamingTV -

Model Specification 1 will contain the following 8 predictors - Contract, tenure, MonthlyCharges, SeniorCitizen, PaymentMethod, InternetService, MultipleLines, StreamingTV.

##3.2 Model Specification 2 ( Based on Variable Selection Method - Lasso)

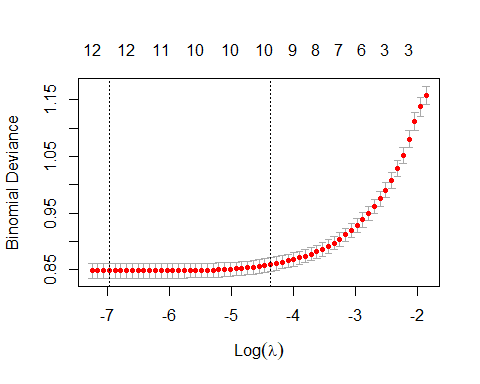
#Lasso   
library(glmnet)  
  
x <- model.matrix(Churn ~ Contract + tenure + MonthlyCharges + SeniorCitizen + PaymentMethod + InternetService + MultipleLines + StreamingTV,  
 data = telco)[,-1]  
  
y <- telco$Churn  
  
set.seed(1)  
  
lasso.mod <- glmnet(x, y, family = "binomial", alpha = 1)  
plot(lasso.mod)



lasso.cv <- cv.glmnet(x, y, family = "binomial", alpha = 1)   
  
cbind("Lambda" = lasso.cv$lambda,   
 "10FCV Deviance" = lasso.cv$cvm)[1:20, ]

## Lambda 10FCV Deviance  
## [1,] 0.15640141 1.1573919  
## [2,] 0.14250714 1.1374411  
## [3,] 0.12984720 1.1111332  
## [4,] 0.11831193 1.0802124  
## [5,] 0.10780143 1.0521178  
## [6,] 0.09822464 1.0280451  
## [7,] 0.08949864 1.0076032  
## [8,] 0.08154782 0.9901863  
## [9,] 0.07430334 0.9751731  
## [10,] 0.06770244 0.9617901  
## [11,] 0.06168794 0.9492564  
## [12,] 0.05620775 0.9379620  
## [13,] 0.05121441 0.9281311  
## [14,] 0.04666466 0.9195532  
## [15,] 0.04251910 0.9115730  
## [16,] 0.03874182 0.9040011  
## [17,] 0.03530011 0.8973110  
## [18,] 0.03216414 0.8914304  
## [19,] 0.02930677 0.8862430  
## [20,] 0.02670324 0.8817160

plot(lasso.cv)



bestlamda.lasso <- lasso.cv$lambda.min  
min.dev.lasso <- min(lasso.cv$cvm)  
  
cbind("Best LASSO Lambda" = bestlamda.lasso,   
 "Best LASSO Deviance" = min.dev.lasso)

## Best LASSO Lambda Best LASSO Deviance  
## [1,] 0.0009376018 0.8480045

lasso.coef.best <- coef(lasso.mod, s=bestlamda.lasso)  
  
lasso.coef.best

## 15 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) -0.75898161  
## ContractOne year -0.73683931  
## ContractTwo year -1.47754041  
## tenure -0.03590114  
## MonthlyCharges .   
## SeniorCitizen1 0.29898753  
## PaymentMethodCredit card (automatic) -0.04259280  
## PaymentMethodElectronic check 0.40143446  
## PaymentMethodMailed check -0.01224905  
## InternetServiceFiber optic 1.07154180  
## InternetServiceNo -0.63017326  
## MultipleLinesNo phone service 0.53705006  
## MultipleLinesYes 0.31556810  
## StreamingTVNo internet service .   
## StreamingTVYes 0.34718202

#LASSO dropped MonthlyCharges and StreamingTV

Final Model Specification 2 - Will contain the following 6 predictors : Contract , tenure , SeniorCitizen , PaymentMethod , InternetService , MultipleLines.

#4. Bionomial Logistic Regression

##4.1 Logistic Model with Model Specification 1 (Businesss Rationale)

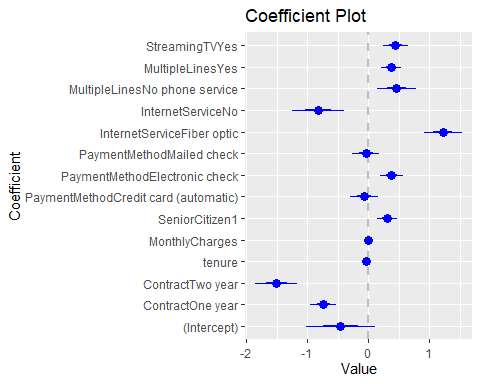
#your analytics question is a classification, run a Logistic regression. In either case you must include the predictors identified above. Later in the project you will refine this initial set of predictors through variable selection, best subsets, or other methods.  
  
#Initial Logistic Model with 8 Predictors  
  
telco.logit <- glm(Churn ~ Contract + tenure + MonthlyCharges + SeniorCitizen +  
 PaymentMethod + InternetService + MultipleLines + StreamingTV,  
 family = "binomial"(link = "logit"), data = telco)  
summary(telco.logit)

##   
## Call:  
## glm(formula = Churn ~ Contract + tenure + MonthlyCharges + SeniorCitizen +   
## PaymentMethod + InternetService + MultipleLines + StreamingTV,   
## family = binomial(link = "logit"), data = telco)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9076 -0.6703 -0.3015 0.7374 3.1803   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.458650 0.282496 -1.624 0.104469   
## ContractOne year -0.740343 0.105221 -7.036 1.98e-12 \*\*\*  
## ContractTwo year -1.507829 0.172174 -8.758 < 2e-16 \*\*\*  
## tenure -0.035772 0.002267 -15.782 < 2e-16 \*\*\*  
## MonthlyCharges -0.005740 0.005187 -1.107 0.268451   
## SeniorCitizen1 0.304991 0.082357 3.703 0.000213 \*\*\*  
## PaymentMethodCredit card (automatic) -0.074058 0.113038 -0.655 0.512366   
## PaymentMethodElectronic check 0.379707 0.093524 4.060 4.91e-05 \*\*\*  
## PaymentMethodMailed check -0.042179 0.112467 -0.375 0.707633   
## InternetServiceFiber optic 1.223545 0.156095 7.838 4.56e-15 \*\*\*  
## InternetServiceNo -0.820517 0.211687 -3.876 0.000106 \*\*\*  
## MultipleLinesNo phone service 0.461708 0.161612 2.857 0.004278 \*\*   
## MultipleLinesYes 0.367647 0.082465 4.458 8.26e-06 \*\*\*  
## StreamingTVNo internet service NA NA NA NA   
## StreamingTVYes 0.446309 0.104747 4.261 2.04e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8143.4 on 7031 degrees of freedom  
## Residual deviance: 5929.3 on 7018 degrees of freedom  
## AIC: 5957.3  
##   
## Number of Fisher Scoring iterations: 6

log.odds <- coef(telco.logit) # Extract the log-odds coefficients  
odds <- exp(log.odds) # Convert the log-odds to odds  
print(cbind("Log-Odds" = log.odds,   
 "Odds" = odds),   
 digits = 2)

## Log-Odds Odds  
## (Intercept) -0.4587 0.63  
## ContractOne year -0.7403 0.48  
## ContractTwo year -1.5078 0.22  
## tenure -0.0358 0.96  
## MonthlyCharges -0.0057 0.99  
## SeniorCitizen1 0.3050 1.36  
## PaymentMethodCredit card (automatic) -0.0741 0.93  
## PaymentMethodElectronic check 0.3797 1.46  
## PaymentMethodMailed check -0.0422 0.96  
## InternetServiceFiber optic 1.2235 3.40  
## InternetServiceNo -0.8205 0.44  
## MultipleLinesNo phone service 0.4617 1.59  
## MultipleLinesYes 0.3676 1.44  
## StreamingTVNo internet service NA NA  
## StreamingTVYes 0.4463 1.56

coefplot(telco.logit)



##4.2 Initial Insights from Logistic Model:

Looking into the deviance statistics, The selected 10 predictors in the model helped to reduce the deviance of null model by (8143.4 - 5881.2) / 8143.4 = 0.277 or 27.7%. This Percentage reduction is not very large but this appears as a good model. After the Variable selection method, we’re positive that the deviance will get further reduced.

##4.3 Assumption Tests - Multicollinearity Test

library(klaR)  
library(car)  
CIvalue <- cond.index(telco.logit, data = telco)  
CIvalue

## [1] 1.000000e+00 1.585463e+00 2.262366e+00 2.354945e+00 2.543281e+00  
## [6] 2.788187e+00 2.965535e+00 3.395387e+00 3.971226e+00 4.415542e+00  
## [11] 5.130653e+00 7.309216e+00 8.836453e+00 3.713878e+01 1.029284e+09

As Condition Index is < 30, collinearity is not a concern. So there is no concern of Multicollinearity

##4.2 Logistic Model with Model Specification 2 (Lasso)

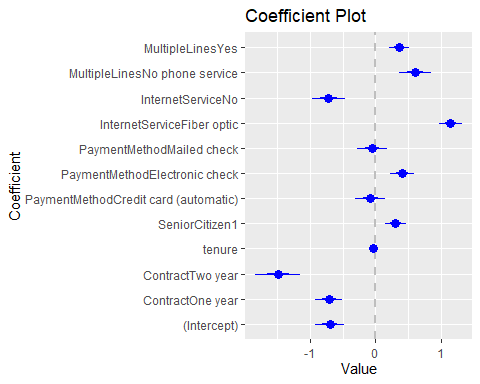
telco.logit.6 <- glm(Churn ~ Contract + tenure +  
 SeniorCitizen + PaymentMethod +  
 InternetService + MultipleLines,  
 family = "binomial"(link = "logit"), data = telco)  
summary(telco.logit.6)

##   
## Call:  
## glm(formula = Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod +   
## InternetService + MultipleLines, family = binomial(link = "logit"),   
## data = telco)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8348 -0.6777 -0.3084 0.7520 3.1639   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.696095 0.109652 -6.348 2.18e-10 \*\*\*  
## ContractOne year -0.709785 0.103525 -6.856 7.07e-12 \*\*\*  
## ContractTwo year -1.485498 0.170212 -8.727 < 2e-16 \*\*\*  
## tenure -0.034674 0.002123 -16.331 < 2e-16 \*\*\*  
## SeniorCitizen1 0.304466 0.082093 3.709 0.000208 \*\*\*  
## PaymentMethodCredit card (automatic) -0.079223 0.112748 -0.703 0.482271   
## PaymentMethodElectronic check 0.405725 0.093119 4.357 1.32e-05 \*\*\*  
## PaymentMethodMailed check -0.053788 0.112136 -0.480 0.631467   
## InternetServiceFiber optic 1.138926 0.087575 13.005 < 2e-16 \*\*\*  
## InternetServiceNo -0.717191 0.126798 -5.656 1.55e-08 \*\*\*  
## MultipleLinesNo phone service 0.601073 0.124131 4.842 1.28e-06 \*\*\*  
## MultipleLinesYes 0.359457 0.077426 4.643 3.44e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8143.4 on 7031 degrees of freedom  
## Residual deviance: 5955.1 on 7020 degrees of freedom  
## AIC: 5979.1  
##   
## Number of Fisher Scoring iterations: 6

log.odds.6 <- coef(telco.logit.6) # Extract the log-odds coefficients  
odds.6 <- exp(log.odds.6) # Convert the log-odds to odds  
print(cbind("Log-Odds" = log.odds.6,   
 "Odds" = odds.6),   
 digits = 2)

## Log-Odds Odds  
## (Intercept) -0.696 0.50  
## ContractOne year -0.710 0.49  
## ContractTwo year -1.485 0.23  
## tenure -0.035 0.97  
## SeniorCitizen1 0.304 1.36  
## PaymentMethodCredit card (automatic) -0.079 0.92  
## PaymentMethodElectronic check 0.406 1.50  
## PaymentMethodMailed check -0.054 0.95  
## InternetServiceFiber optic 1.139 3.12  
## InternetServiceNo -0.717 0.49  
## MultipleLinesNo phone service 0.601 1.82  
## MultipleLinesYes 0.359 1.43

coefplot(telco.logit.6)

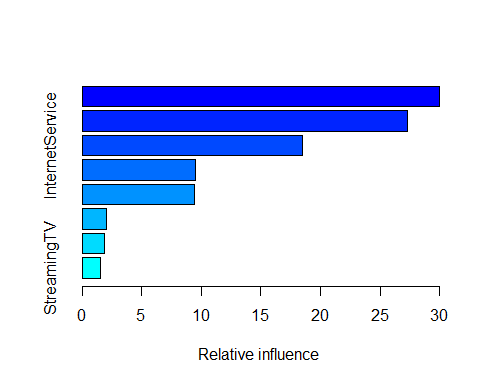


#5. Bootstrap Aggregation Regression Tree ##5.1 Bootstrap Aggregation with Model Specification 1

#first set of predictors, business logic  
library(gbm)  
set.seed(1)  
#fit the boosted tree model and present Relative Influence graph  
telco.boost <- gbm(Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod + InternetService + MultipleLines + MonthlyCharges + StreamingTV, data = telco,   
 distribution = "gaussian",  
 shrinkage = 0.01,  
 cv.folds = 10,  
 n.trees = 5000,  
 interaction.depth = 1)  
telco.boost

## gbm(formula = Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod +   
## InternetService + MultipleLines + MonthlyCharges + StreamingTV,   
## distribution = "gaussian", data = telco, n.trees = 5000,   
## interaction.depth = 1, shrinkage = 0.01, cv.folds = 10)  
## A gradient boosted model with gaussian loss function.  
## 5000 iterations were performed.  
## The best cross-validation iteration was 2481.  
## There were 8 predictors of which 8 had non-zero influence.

summary(telco.boost)

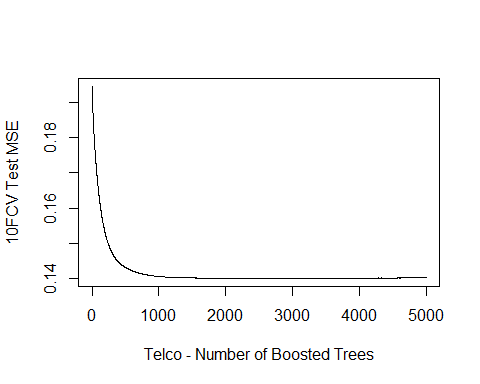


## var rel.inf  
## Contract Contract 30.002660  
## tenure tenure 27.273353  
## InternetService InternetService 18.447329  
## PaymentMethod PaymentMethod 9.462828  
## MonthlyCharges MonthlyCharges 9.421066  
## MultipleLines MultipleLines 2.050049  
## SeniorCitizen SeniorCitizen 1.855937  
## StreamingTV StreamingTV 1.486779

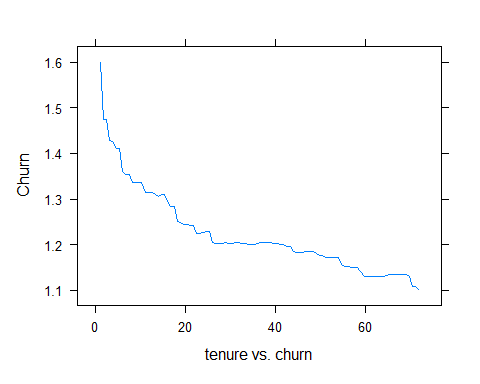
#indicates the best tree  
best.num.trees <- which.min(telco.boost$cv.error)  
min.mse.boost <- round(min(telco.boost$cv.error), digits = 3)  
min.rmse.boost <- round(sqrt(min.mse.boost), digits = 3)  
paste("Min 10FCV Test MSE =", min.mse.boost,  
 "Test RMSE = ", min.rmse.boost,  
 "at", best.num.trees, "trees")

## [1] "Min 10FCV Test MSE = 0.14 Test RMSE = 0.374 at 2481 trees"

#Plot CV Test Error against Number of Trees  
plot(telco.boost$cv.error,  
 type = "l",  
 xlab = "Telco - Number of Boosted Trees",  
 ylab = "10FCV Test MSE")



#Plot Partial Dependency Graph  
plot(telco.boost,  
 i = "tenure",  
 ylab = "Churn",  
 xlab = "tenure vs. churn")

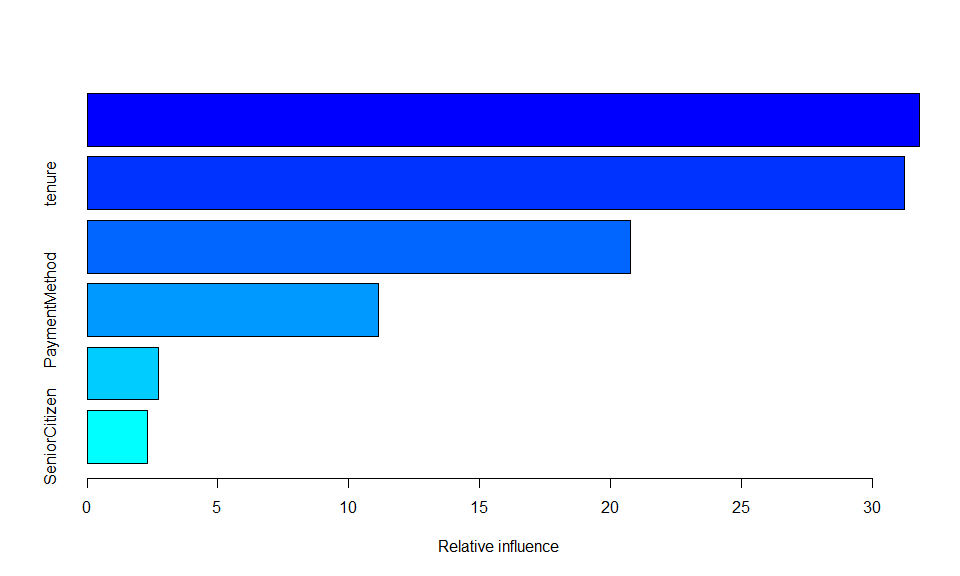


##5.2 Bootstrap Aggregation with Model Specification 2

#second set of predictors, selected by LASSO  
set.seed(1)  
#fit the boosted tree model and present Relative Influence graph  
telco.boost <- gbm(Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod + InternetService + MultipleLines, data = telco,   
 distribution = "gaussian",  
 shrinkage = 0.01,  
 cv.folds = 10,  
 n.trees = 5000,  
 interaction.depth = 1)  
telco.boost

## gbm(formula = Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod +   
## InternetService + MultipleLines, distribution = "gaussian",   
## data = telco, n.trees = 5000, interaction.depth = 1, shrinkage = 0.01,   
## cv.folds = 10)  
## A gradient boosted model with gaussian loss function.  
## 5000 iterations were performed.  
## The best cross-validation iteration was 2515.  
## There were 6 predictors of which 6 had non-zero influence.

summary(telco.boost)

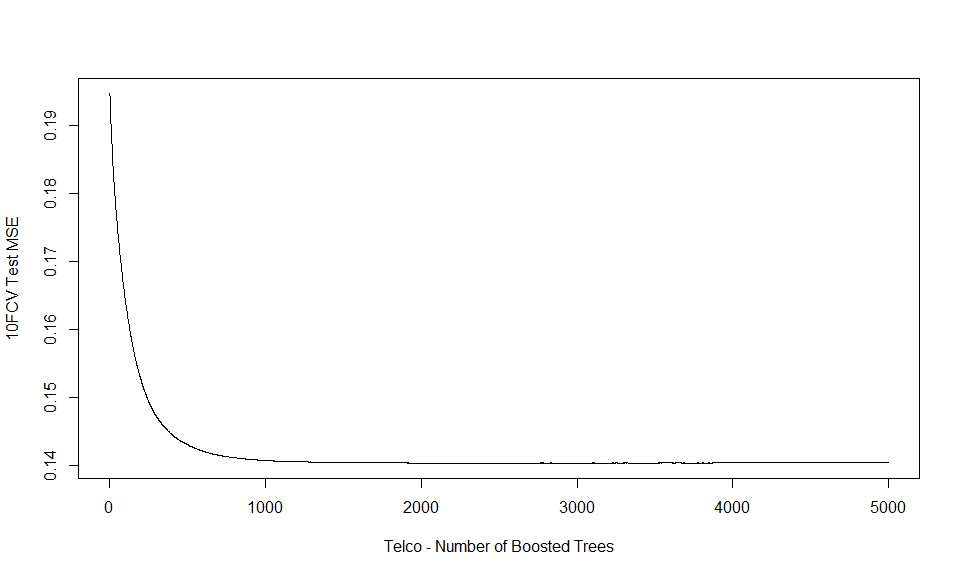


## var rel.inf  
## Contract Contract 31.812558  
## tenure tenure 31.232384  
## InternetService InternetService 20.775840  
## PaymentMethod PaymentMethod 11.126063  
## MultipleLines MultipleLines 2.738179  
## SeniorCitizen SeniorCitizen 2.314976

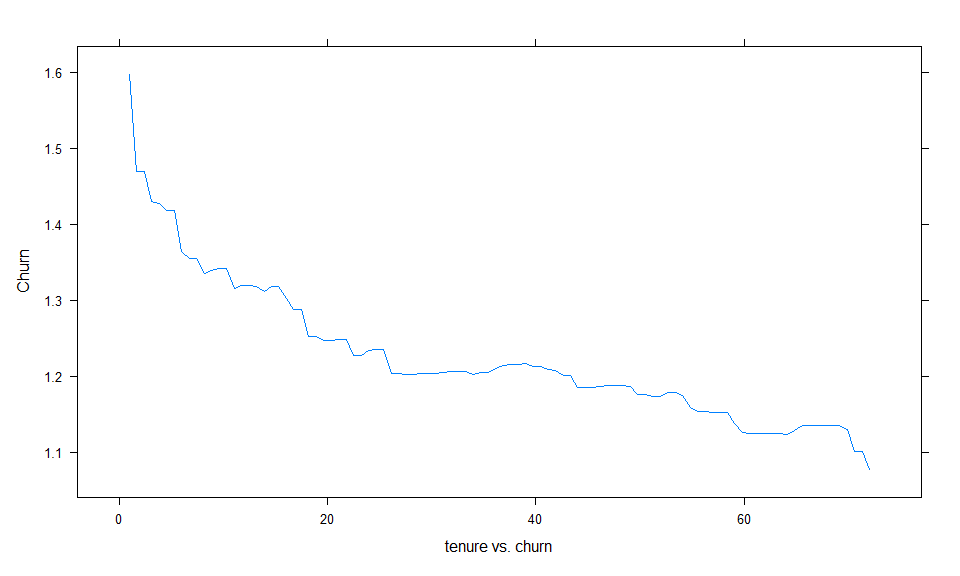
#indicates the best tree  
best.num.trees <- which.min(telco.boost$cv.error)  
min.mse.boost <- round(min(telco.boost$cv.error), digits = 3)  
min.rmse.boost <- round(sqrt(min.mse.boost), digits = 3)  
paste("Min 10FCV Test MSE =", min.mse.boost,  
 "Test RMSE = ", min.rmse.boost,  
 "at", best.num.trees, "trees")

## [1] "Min 10FCV Test MSE = 0.14 Test RMSE = 0.374 at 2515 trees"

#Plot CV Test Error against Number of Trees  
plot(telco.boost$cv.error,  
 type = "l",  
 xlab = "Telco - Number of Boosted Trees",  
 ylab = "10FCV Test MSE")



#Plot Partial Dependency Graph  
plot(telco.boost,  
 i = "tenure",  
 ylab = "Churn",  
 xlab = "tenure vs. churn")



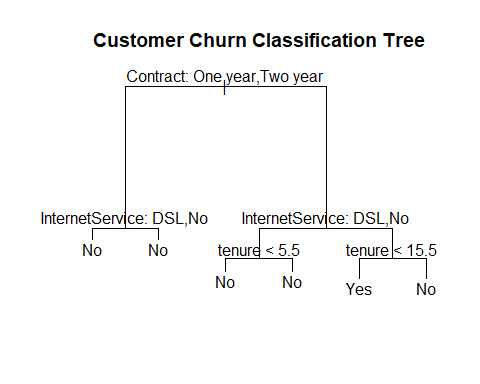
#6. Binomial Classification Tree

##6.1 Binomial Classification Tree with Model Specification 1 (Business Rationale)

telco$Churn <- as.factor(telco$Churn)  
  
class(telco$Churn)

## [1] "factor"

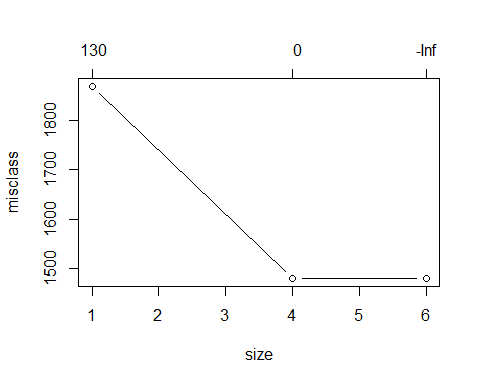
library(tree)  
  
churn.tree <- tree(Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod +  
 InternetService + MultipleLines + MonthlyCharges +   
 StreamingTV,telco)  
  
plot(churn.tree)  
  
text(churn.tree, pretty = 0)  
  
title("Customer Churn Classification Tree")



set.seed(1)  
  
churn.tree.cv <- cv.tree(churn.tree, FUN = prune.misclass)  
  
cbind("Tree Size" = churn.tree.cv$size, "Misclass" = churn.tree.cv$dev)

## Tree Size Misclass  
## [1,] 6 1480  
## [2,] 4 1480  
## [3,] 1 1869

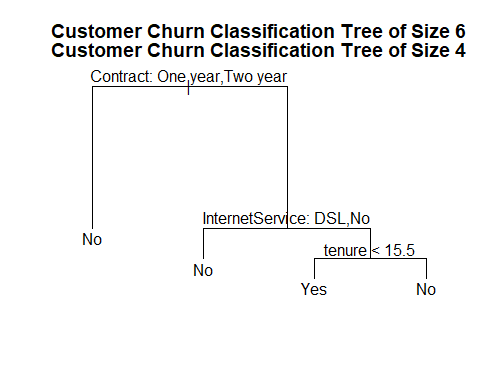
plot(churn.tree.cv, type = "b")



min.dev <- min(churn.tree.cv$dev)  
  
best.ind <- which(churn.tree.cv$dev == min.dev)  
  
best.size <- churn.tree.cv$size[best.ind]  
  
cbind("Smallest Misclass" = min.dev,"Which Tree" = best.ind,"Best Tree Size" = best.size)

## Smallest Misclass Which Tree Best Tree Size  
## [1,] 1480 1 6  
## [2,] 1480 2 4

churn.prune <- prune.misclass(churn.tree, best = best.size)  
  
plot(churn.prune)  
  
text(churn.prune, pretty = 0)  
  
title(paste("Customer Churn Classification Tree of Size", best.size))



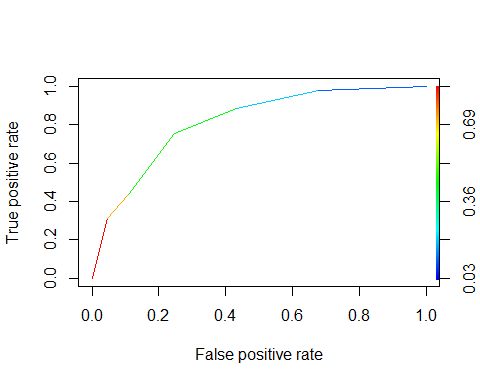
library(ROCR)  
set.seed(1)  
  
tr.size <- 0.7  
  
train <- sample(1:nrow(telco), tr.size \* nrow(telco))  
  
telco.train <- telco[train,]  
  
nrow(telco.train)

## [1] 4922

telco.test <- telco[-train,]  
nrow(telco.test)

## [1] 2110

telco.tree.tr <- tree(Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod + InternetService + MultipleLines,telco.train)  
  
telco.tree.prob <- predict(telco.tree.tr, telco.test)[ , 2]  
  
  
pred <- prediction(telco.tree.prob, telco.test$Churn)  
perf <- performance(pred,"tpr","fpr")  
plot(perf, colorize=T)

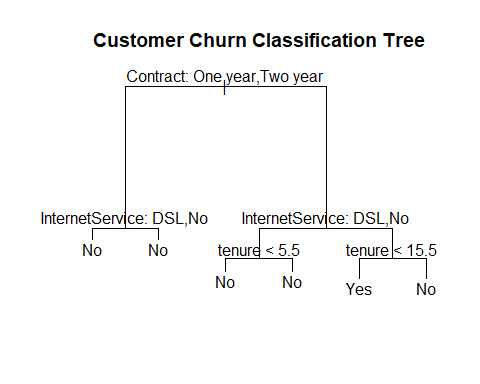


auc <- performance(pred, "auc")  
auc.name <- auc@y.name[[1]]  
auc.val <- round(auc@y.values[[1]], digits = 3)  
paste(auc.name, auc.val)

## [1] "Area under the ROC curve 0.813"

##6.2 Binomial Classification Tree with Model Specification 2

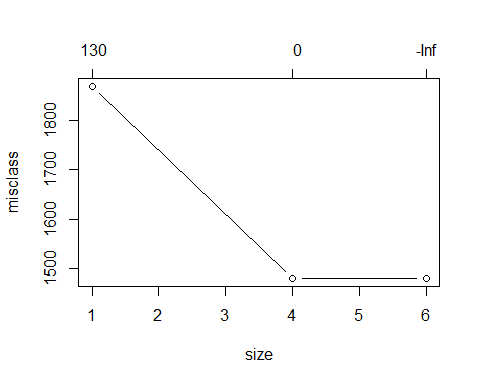
library(tree)  
  
churn.tree <- tree(Churn ~ Contract + tenure +  
 SeniorCitizen + PaymentMethod +  
 InternetService + MultipleLines ,telco)  
  
plot(churn.tree)  
  
text(churn.tree, pretty = 0)  
  
title("Customer Churn Classification Tree")



set.seed(1)  
  
churn.tree.cv <- cv.tree(churn.tree, FUN = prune.misclass)  
  
cbind("Tree Size" = churn.tree.cv$size, "Misclass" = churn.tree.cv$dev)

## Tree Size Misclass  
## [1,] 6 1480  
## [2,] 4 1480  
## [3,] 1 1869

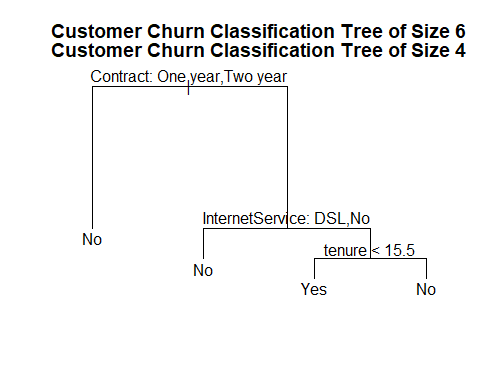
plot(churn.tree.cv, type = "b")



min.dev <- min(churn.tree.cv$dev)  
  
best.ind <- which(churn.tree.cv$dev == min.dev)  
  
best.size <- churn.tree.cv$size[best.ind]  
  
cbind("Smallest Misclass" = min.dev,"Which Tree" = best.ind,"Best Tree Size" = best.size)

## Smallest Misclass Which Tree Best Tree Size  
## [1,] 1480 1 6  
## [2,] 1480 2 4

churn.prune <- prune.misclass(churn.tree, best = best.size)  
  
plot(churn.prune)  
  
text(churn.prune, pretty = 0)  
  
title(paste("Customer Churn Classification Tree of Size", best.size))



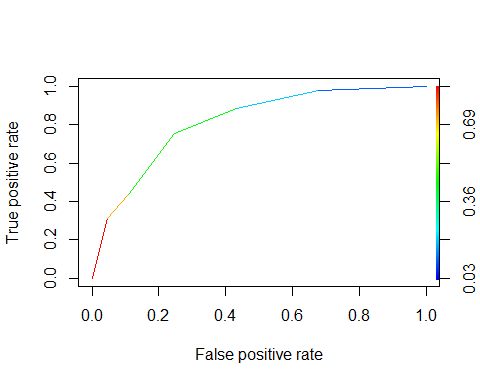
library(ROCR)  
set.seed(1)  
  
tr.size <- 0.7  
  
train <- sample(1:nrow(telco), tr.size \* nrow(telco))  
  
telco.train <- telco[train,]  
  
nrow(telco.train)

## [1] 4922

telco.test <- telco[-train,]  
nrow(telco.test)

## [1] 2110

telco.tree.tr <- tree(Churn ~ Contract + tenure + SeniorCitizen + PaymentMethod + InternetService + MultipleLines,telco.train)  
  
telco.tree.prob <- predict(telco.tree.tr, telco.test)[ , 2]  
  
  
pred <- prediction(telco.tree.prob, telco.test$Churn)  
perf <- performance(pred,"tpr","fpr")  
plot(perf, colorize=T)



auc <- performance(pred, "auc")  
auc.name <- auc@y.name[[1]]  
auc.val <- round(auc@y.values[[1]], digits = 3)  
paste(auc.name, auc.val)

## [1] "Area under the ROC curve 0.813"

#7. Cross Validation and Final Model Selection   
  
  
#8. Other Plots and Anovve Tests   
  
```r  
#-------- Kensey  
#Analyze the predictors  
library("fastDummies")  
telco<-dummy\_cols(telco, select\_columns="Churn")#Convert churn into binary  
  
summary(aov(telco$Churn\_Yes ~ telco$tenure))

## Df Sum Sq Mean Sq F value Pr(>F)   
## telco$tenure 1 172 172.01 1008 <2e-16 \*\*\*  
## Residuals 7030 1200 0.17   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov(telco$Churn\_Yes ~ telco$MonthlyCharges))

## Df Sum Sq Mean Sq F value Pr(>F)   
## telco$MonthlyCharges 1 51 51.04 271.6 <2e-16 \*\*\*  
## Residuals 7030 1321 0.19   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov(telco$Churn\_Yes ~ telco$MultipleLines))

## Df Sum Sq Mean Sq F value Pr(>F)   
## telco$MultipleLines 2 2.2 1.0998 5.642 0.00356 \*\*  
## Residuals 7029 1370.0 0.1949   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

lm.fit.tenure<-lm(Churn\_Yes ~tenure, data=telco)  
  
summary(lm.fit.tenure)

##   
## Call:  
## lm(formula = Churn\_Yes ~ tenure, data = telco)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.4660 -0.3258 -0.1219 0.5340 0.9864   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4723900 0.0081637 57.86 <2e-16 \*\*\*  
## tenure -0.0063724 0.0002008 -31.74 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4132 on 7030 degrees of freedom  
## Multiple R-squared: 0.1254, Adjusted R-squared: 0.1252   
## F-statistic: 1008 on 1 and 7030 DF, p-value: < 2.2e-16

lm.fit.selectVars<-lm(Churn\_Yes ~tenure+MonthlyCharges+MultipleLines, data=telco)  
  
summary(lm.fit.selectVars)

##   
## Call:  
## lm(formula = Churn\_Yes ~ tenure + MonthlyCharges + MultipleLines,   
## data = telco)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7098 -0.2913 -0.1238 0.3867 1.1933   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.2156713 0.0124296 17.351 < 2e-16 \*\*\*  
## tenure -0.0080515 0.0002039 -39.496 < 2e-16 \*\*\*  
## MonthlyCharges 0.0042368 0.0001808 23.428 < 2e-16 \*\*\*  
## MultipleLinesNo phone service 0.1127031 0.0166900 6.753 1.57e-11 \*\*\*  
## MultipleLinesYes 0.0609696 0.0114525 5.324 1.05e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3913 on 7027 degrees of freedom  
## Multiple R-squared: 0.2159, Adjusted R-squared: 0.2154   
## F-statistic: 483.6 on 4 and 7027 DF, p-value: < 2.2e-16

anova(lm.fit.tenure,lm.fit.selectVars)

## Analysis of Variance Table  
##   
## Model 1: Churn\_Yes ~ tenure  
## Model 2: Churn\_Yes ~ tenure + MonthlyCharges + MultipleLines  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7030 1200.2   
## 2 7027 1076.0 3 124.22 270.41 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

library(corrplot)  
telcoCorr <-cor(telco[, c("tenure", "MonthlyCharges", "Churn\_Yes")])  
corrplot(telcoCorr, method = "number", order="hclust")

