**Telecom Customer Churn Prediction Assessment**

**Dataset:**[**Cellphone**](https://olympus.greatlearning.in/courses/4742/files/411877/download?wrap=1)**[View in a new window](https://olympus.greatlearning.in/courses/4742/files/411877/download?wrap=1)**

Customer Churn is a burning problem for Telecom companies. In this project, we simulate one such case of customer churn where we work on a data of postpaid customers with a contract. The data has information about customer usage behavior, contract details, and payment details. The data also indicates which the customers who canceled their service were. Based on this past data, we need to build a model which can predict whether a customer will cancel their service in the future or not.

You are expected to do the following:

* Detailed Exploratory Data Analysis report of the dataset along with the missing value treatment
* Multicollinearity check and summarization of problem statement for business stakeholders
* Logistic Regression Model: creation and interpretation of the results
* Comparing the model performances using confusion matrix, GINI coefficient and  Kolmogorov Smirnov(KS-chart) along with the remarks on the best model
* Actionable Insights for the business stakeholders

**Group Members**

1. Aditya Godse
2. Akshay Surve
3. Amay Saxena

**Answer**

**1] Exploratory Data Analysis**

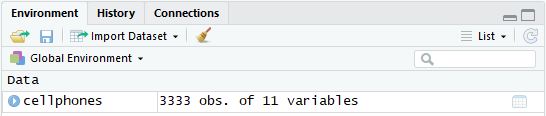
setwd("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment 5 PM")

getwd()

cellphones<-read.csv("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment 5 PM/Cellphone csv.csv")

checking for structure of the data

str(cellphones)



'data.frame': 3333 obs. of 11 variables:

$ Churn : int 0 0 0 0 0 0 0 0 0 0 ...

$ AccountWeeks : int 128 107 137 84 75 118 121 147 117 141 ...

$ ContractRenewal: int 1 1 1 0 0 0 1 0 1 0 ...

$ DataPlan : int 1 1 0 0 0 0 1 0 0 1 ...

$ DataUsage : num 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...

$ CustServCalls : int 1 1 0 2 3 0 3 0 1 0 ...

$ DayMins : num 265 162 243 299 167 ...

$ DayCalls : int 110 123 114 71 113 98 88 79 97 84 ...

$ MonthlyCharge : num 89 82 52 57 41 57 87.3 36 63.9 93.2 ...

$ OverageFee : num 9.87 9.78 6.06 3.1 7.42 ...

$ RoamMins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...

Converting variables contract renewal and data plan in to factor variables since they are of categorical type.

cellphones$Churn<-as.factor(cellphones$Churn)

cellphones$ContractRenewal<-as.factor(cellphones$ContractRenewal)

cellphones$DataPlan<-as.factor(cellphones$DataPlan)

checking structure of dataset again

str(cellphones)

'data.frame': 3333 obs. of 11 variables:

$ Churn : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ AccountWeeks : int 128 107 137 84 75 118 121 147 117 141 ...

$ ContractRenewal: Factor w/ 2 levels "0","1": 2 2 2 1 1 1 2 1 2 1 ...

$ DataPlan : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 2 1 1 2 ...

$ DataUsage : num 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...

$ CustServCalls : int 1 1 0 2 3 0 3 0 1 0 ...

$ DayMins : num 265 162 243 299 167 ...

$ DayCalls : int 110 123 114 71 113 98 88 79 97 84 ...

$ MonthlyCharge : num 89 82 52 57 41 57 87.3 36 63.9 93.2 ...

$ OverageFee : num 9.87 9.78 6.06 3.1 7.42 ...

$ RoamMins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...

summary(cellphones)

Churn AccountWeeks ContractRenewal DataPlan DataUsage

Min. :0.0000 Min. : 1.0 Min. :0.0000 Min. :0.0000 Min. :0.0000

1st Qu.:0.0000 1st Qu.: 74.0 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.0000

Median :0.0000 Median :101.0 Median :1.0000 Median :0.0000 Median :0.0000

Mean :0.1449 Mean :101.1 Mean :0.9031 Mean :0.2766 Mean :0.8165

3rd Qu.:0.0000 3rd Qu.:127.0 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.7800

Max. :1.0000 Max. :243.0 Max. :1.0000 Max. :1.0000 Max. :5.4000

CustServCalls DayMins DayCalls MonthlyCharge OverageFee

Min. :0.000 Min. : 0.0 Min. : 0.0 Min. : 14.00 Min. : 0.00

1st Qu.:1.000 1st Qu.:143.7 1st Qu.: 87.0 1st Qu.: 45.00 1st Qu.: 8.33

Median :1.000 Median :179.4 Median :101.0 Median : 53.50 Median :10.07

Mean :1.563 Mean :179.8 Mean :100.4 Mean : 56.31 Mean :10.05

3rd Qu.:2.000 3rd Qu.:216.4 3rd Qu.:114.0 3rd Qu.: 66.20 3rd Qu.:11.77

Max. :9.000 Max. :350.8 Max. :165.0 Max. :111.30 Max. :18.19

RoamMins

Min. : 0.00

1st Qu.: 8.50

Median :10.30

Mean :10.24

3rd Qu.:12.10

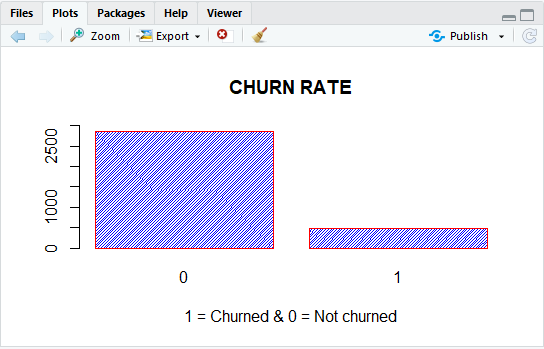
Max. :20.00

From summary it can be seen that there are no missing values in the dataset. (i.e NA’s)

**a) Basic data summary, Univariate, Bivariate analysis, graphs**

attach(cellphones)

barplot(table(cellphones$Churn),main = "CHURN RATE",col="Blue",border="Red",density=50,ylim = c(0,3000),xlab="1 = Churned & 0 = Not churned")



summary(Churn)

0 1

2850 483

From summary of churn rate of customers, it can be seen that

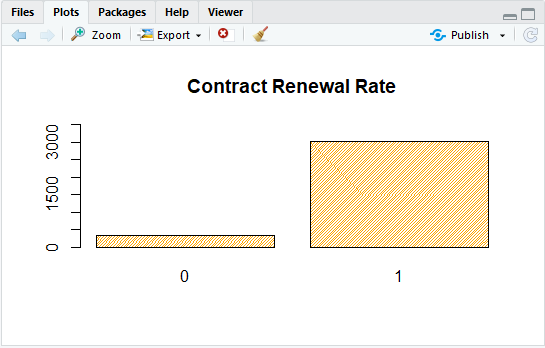
* Total number of churned customers = 2850
* Total number of customers who did not churn = 483

Percentage of customer churned as per the data = (483/(483+2850))\*100 = 14.49145 %

Now, considering only factor variables,

**1) Contract Renewal**

barplot(table(cellphones$ContractRenewal),main = "Contract Renewal Rate",density = 50,col="Orange",ylim = c(0,3500))



summary(ContractRenewal)

0 1

323 3010

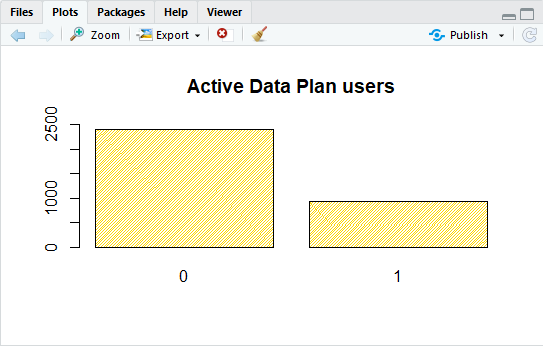
From summary of churn rate of customers, it can be seen that

* Total number of customers who renewed their contract = 323
* Total number of customers who did not renew their contract = 3010

Percentage of customer who renew their contract as per the data = (323/(3010+323))\*100 = 9.690969%

**2) Data plan**

barplot(table(cellphones$DataPlan),main = "Active Data Plan users",density = 50,col="gold",ylim = c(0,2500))



summary(DataPlan)

0 1

2411 922

From summary of churn rate of customers, it can be seen that

* Total number of customers who own data plan = 922
* Total number of customers who do not own data plan = 2411

Percentage of customer who own data plan as per the data = (922/(922+2411))\*100 = 27.66277%

**b)Check for Multicollinearity - Plot the graph based on Multicollinearity**

For checking multi-collinearity of the dataset, considering only numeric variables.

Therefore, discarding the factor variable columns.

numericdata<-cellphones[,c(-1,-3,-4)]

print(cor(numericdata),digits = 3)

AccountWeeks DataUsage CustServCalls DayMins DayCalls MonthlyCharge

AccountWeeks 1.00000 0.01439 -0.00380 0.00622 0.03847 0.01258

DataUsage 0.01439 1.00000 -0.02172 0.00318 -0.00796 0.78166

CustServCalls -0.00380 -0.02172 1.00000 -0.01342 -0.01894 -0.02802

DayMins 0.00622 0.00318 -0.01342 1.00000 0.00675 0.56797

DayCalls 0.03847 -0.00796 -0.01894 0.00675 1.00000 -0.00796

MonthlyCharge 0.01258 0.78166 -0.02802 0.56797 -0.00796 1.00000

OverageFee -0.00675 0.01964 -0.01296 0.00704 -0.02145 0.28177

RoamMins 0.00951 0.16275 -0.00964 -0.01015 0.02156 0.11743

OverageFee RoamMins

AccountWeeks -0.00675 0.00951

DataUsage 0.01964 0.16275

CustServCalls -0.01296 -0.00964

DayMins 0.00704 -0.01015

DayCalls -0.02145 0.02156

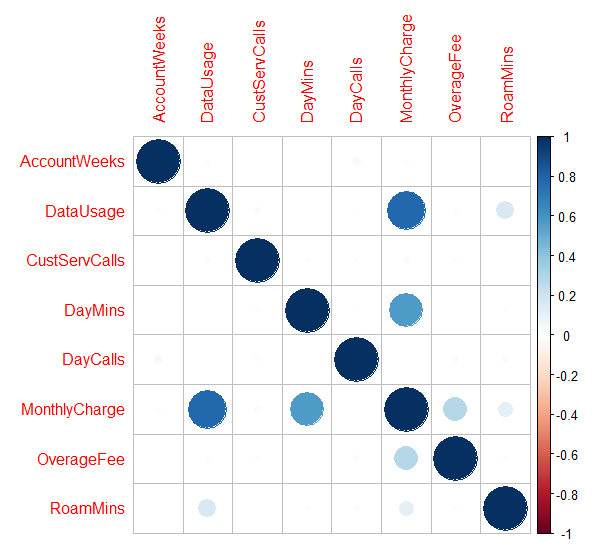
MonthlyCharge 0.28177 0.11743

OverageFee 1.00000 -0.01102

RoamMins -0.01102 1.00000

library(corrplot)

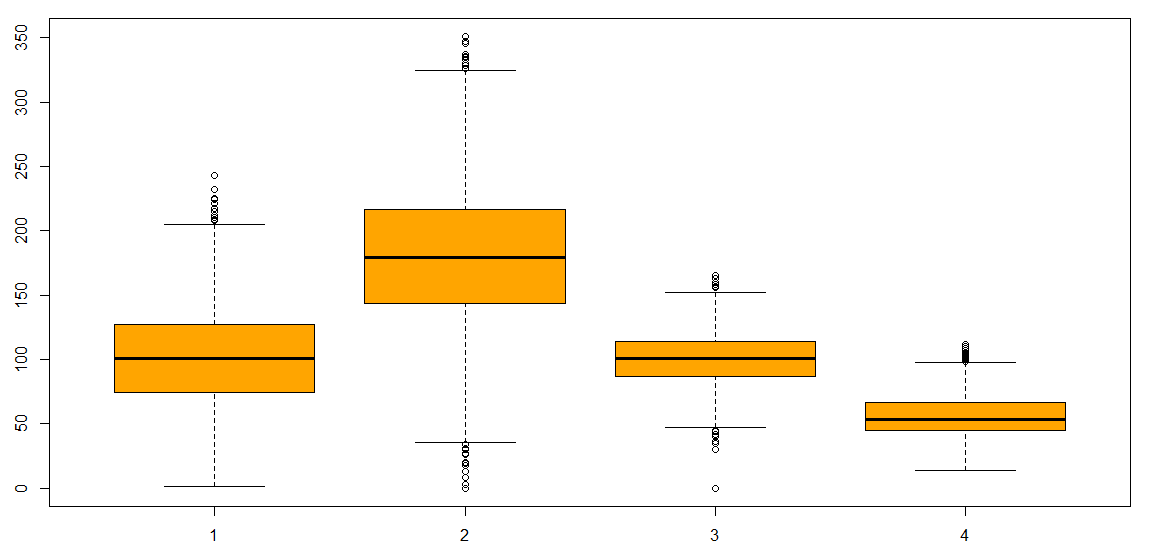
corrplot(cor(numericdata))



From the above corrplot we can see that the variables monthly charge have a high positive correlation with data usage and day\_mins.

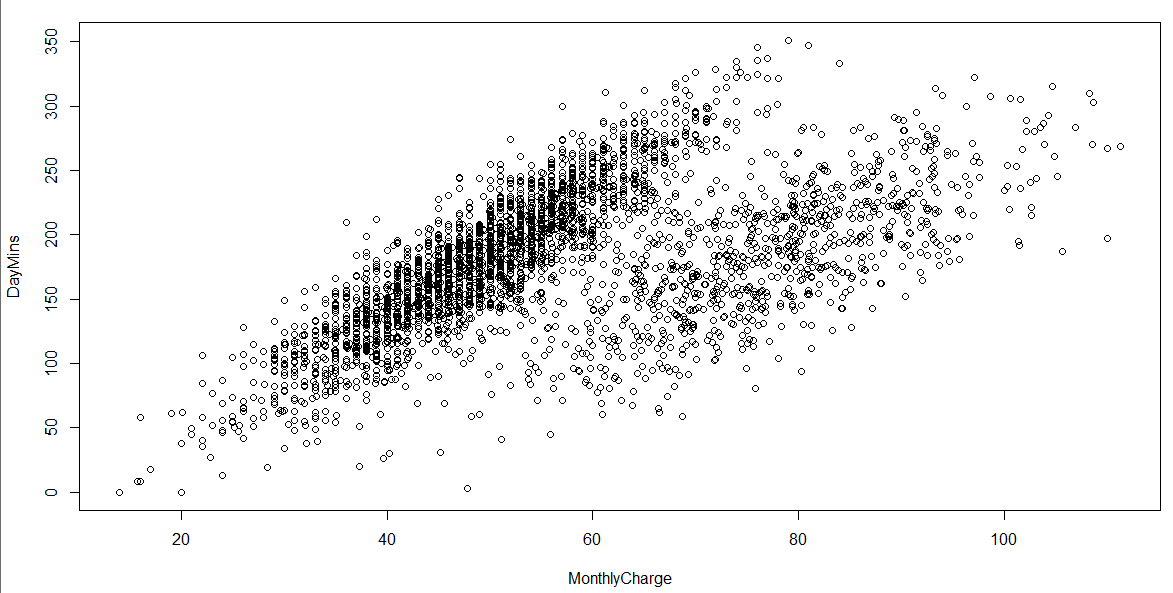
**c) Check for Outliers and missing values and check the summary of the dataset**

boxplot(AccountWeeks,DayMins,DayCalls,MonthlyCharge,col = "orange")



From the boxplot of Account Weeks, Day\_Mins, Day\_Calls, Monthly\_Charge it can be seen that day\_mins have significant number of outliers as compared to other variables.

plot(MonthlyCharge,DayMins)



From the graphs it is evident that the monthly charge increases with increase calling minutes.

**d) Interpreting the business problem and sharing the observations**

To understand which variables are responsible for the churning of customers in telecom industry.

**2] Applying Supervised Machine Learning Techniques (Test & Train)**

* Dividing the data into train & test dataset samples in 70:30 ratio

set.seed(1104)

library(caTools)

split=sample.split(cellphones$Churn,SplitRatio = 0.70)

train.logit<-subset(cellphones,split==TRUE)

test.logit<-subset(cellphones,split==FALSE)

* Checking the data division

dim(train.logit)

[1] 2333 11

Train sample has 2333 observations

dim(test.logit)

[1] 1000 11

Test sample has 1000 observations

Therefore, we can say that data is uniformly divided.

summary(train.logit$Churn)

0 1

1995 338

Checking whether the train sample has respective number of churned customers.

summary(test.logit$Churn)

0 1

855 145

Checking whether the test sample has respective number of churned customers.

**a) Applying Logistic Regression and Interpret the Regression model output**

* Regression model considering all variables

logit<-glm(Churn~.,data=train.logit,family = binomial)

summary(logit)

Call:

glm(formula = Churn ~ ., family = binomial, data = train.logit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0411 -0.5107 -0.3437 -0.2023 2.9947

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.138461 0.652405 -9.409 < 2e-16 \*\*\*

AccountWeeks 0.001732 0.001670 1.037 0.29981

ContractRenewal1 -1.984749 0.173442 -11.443 < 2e-16 \*\*\*

DataPlan1 -1.274565 0.659709 -1.932 0.05336 .

DataUsage -2.635363 2.293838 -1.149 0.25060

CustServCalls 0.534950 0.046490 11.507 < 2e-16 \*\*\*

DayMins -0.033278 0.038769 -0.858 0.39069

DayCalls 0.002290 0.003315 0.691 0.48977

MonthlyCharge 0.271397 0.227890 1.191 0.23369

OverageFee -0.312841 0.388466 -0.805 0.42063

RoamMins 0.073538 0.026301 2.796 0.00517 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1930.4 on 2332 degrees of freedom

Residual deviance: 1525.7 on 2322 degrees of freedom

AIC: 1547.7

Number of Fisher Scoring iterations: 6

From summary it can be seen that contract\_renewal, cust\_serv\_calls and Roam\_Mins are significant variables considering 95% confidence level.

* Checking for Variation Inflation Factor

library(car)

x<-vif(logit)

x

AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls

1.002514 1.070528 14.709675 1514.469611 1.097137

DayMins DayCalls MonthlyCharge OverageFee RoamMins

981.515896 1.003901 2793.552366 211.591753 1.169852

Since VIF value for Data\_Usage, Day\_mins, Monthly\_charge have VIF value greater than 10, hence it can be said that model has multi-collinearity problem.

* Likelihood Ratio Test (LR test)

lrtest(logit)

Model 1: Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +

CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee +

RoamMins

Model 2: Churn ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

1 11 -762.87

2 1 -965.21 -10 404.69 < 2.2e-16 \*\*\*

* Looking for McFadden value

Higher value will indicate that model will give predictive information about the outcome.

pR2(logit)

llh llhNull G2 McFadden r2ML r2CU

-762.8654855 -965.2094900 404.6880089 0.2096374 0.1592513 0.2829458

In our case the McFadden value is around 20% which indicate that model has low probability of giving prediction about outcome.

**b) Applying Logistic Regression with relevant variables and Interpret the Regression model output.**

* Regression model without considering variables with high VIF

Running regression model once again on the model but without considering variables with high VIF. (i.e Data\_usage, Monthly\_charge, Day\_mins)

logit1<-glm(Churn~AccountWeeks+ContractRenewal+DataPlan+CustServCalls+DayCalls+OverageFee+RoamMins,data = train.logit,family = binomial())

summary(logit1)

Call:

glm(formula = Churn ~ AccountWeeks + ContractRenewal + DataPlan +

CustServCalls + DayCalls + OverageFee + RoamMins, family = binomial(),

data = train.logit)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7886 -0.5434 -0.3977 -0.2723 2.7976

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.380485 0.551405 -6.131 8.75e-10 \*\*\*

AccountWeeks 0.001617 0.001609 1.005 0.31487

ContractRenewal1 -1.911588 0.165611 -11.543 < 2e-16 \*\*\*

DataPlan1 -0.965630 0.168303 -5.737 9.61e-09 \*\*\*

CustServCalls 0.477813 0.044208 10.808 < 2e-16 \*\*\*

DayCalls 0.002632 0.003190 0.825 0.40936

OverageFee 0.136718 0.025852 5.288 1.23e-07 \*\*\*

RoamMins 0.070256 0.023615 2.975 0.00293 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1930.4 on 2332 degrees of freedom

Residual deviance: 1641.9 on 2325 degrees of freedom

AIC: 1657.9

Number of Fisher Scoring iterations: 5

vif(logit1)

AccountWeeks ContractRenewal DataPlan CustServCalls DayCalls

1.001169 1.049227 1.025596 1.052851 1.003430

OverageFee RoamMins

1.022242 1.004756

lrtest(logit1)

Likelihood ratio test

Model 1: Churn ~ AccountWeeks + ContractRenewal + DataPlan + CustServCalls +

DayCalls + OverageFee + RoamMins

Model 2: Churn ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

1 8 -820.93

2 1 -965.21 -7 288.57 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

pR2(logit1)

llh llhNull G2 McFadden r2ML r2CU

-820.9255222 -965.2094900 288.5679355 0.1494846 0.1163460 0.2067148

Now we can see that the VIF score for all variables are less than 4 hence the multicollinearity is well handled.

However though the VIF is passed in pseudo R test the Mcfadden score has drooped.

Hence now check with adding Day\_mins and cross checking the VIF score & pR2 test for validation.

* Regression model considering variable Day\_mins

logit2<-glm(Churn~AccountWeeks+ContractRenewal+DayMins+DataPlan+CustServCalls+DayCalls+OverageFee+RoamMins,data = train.logit,family = binomial())

summary(logit2)

Call:

glm(formula = Churn ~ AccountWeeks + ContractRenewal + DayMins +

DataPlan + CustServCalls + DayCalls + OverageFee + RoamMins,

family = binomial(), data = train.logit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0395 -0.5070 -0.3431 -0.2028 3.0426

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.042816 0.639237 -9.453 < 2e-16 \*\*\*

AccountWeeks 0.001687 0.001668 1.012 0.31174

ContractRenewal1 -1.974566 0.172908 -11.420 < 2e-16 \*\*\*

DayMins 0.012872 0.001267 10.161 < 2e-16 \*\*\*

DataPlan1 -1.055169 0.174860 -6.034 1.60e-09 \*\*\*

CustServCalls 0.533687 0.046482 11.482 < 2e-16 \*\*\*

DayCalls 0.002341 0.003316 0.706 0.48016

OverageFee 0.148651 0.027089 5.488 4.08e-08 \*\*\*

RoamMins 0.076768 0.024391 3.147 0.00165 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1930.4 on 2332 degrees of freedom

Residual deviance: 1527.3 on 2324 degrees of freedom

AIC: 1545.3

Number of Fisher Scoring iterations: 5

Now, from summary it can be seen that most of the variables are significant.

lrtest(logit2)

Likelihood ratio test

Model 1: Churn ~ AccountWeeks + ContractRenewal + DayMins + DataPlan +

CustServCalls + DayCalls + OverageFee + RoamMins

Model 2: Churn ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

1 9 -763.64

2 1 -965.21 -8 403.13 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

pR2(logit2)

llh llhNull G2 McFadden r2ML r2CU

-763.6426864 -965.20949 403.13360 0.208832 0.1586910 0.2819502

vif(logit2)

AccountWeeks ContractRenewal DayMins DataPlan CustServCalls

1.001615 1.066210 1.049890 1.036772 1.094684 DayCalls OverageFee RoamMins

1.003406 1.030660 1.006976

odds<-exp(coef(logit2))

(Intercept) AccountWeeks ContractRenewal1 DayMins DataPlan1

0.002374862 1.001688630 0.138821494 1.01295498 0.348133

CustServCalls DayCalls OverageFee RoamMins

1.705207218 1.002343686 1.160268294 1.079791209

From the above test we can now see that the Mcfaden test score has improved and now the VIF is in control too.

It is also passing the log likelihood test since P-value is significant.

For the variables with odds value greater than 1 indicate that that their probability of being significant is more than 50%.

**3] Model Performance Measures (Test & Train)**

**a) Confusion matrix interpretation for all models**

* Checking the prediction accuracy on the test data

predict.logit<-predict.glm(logit2, newdata=test.logit, type="response")

summary(predict.logit)

Min. 1st Qu. Median Mean 3rd Qu. Max.

* 1. 0.04012 0.08716 0.15143 0.19229 0.98625
* Preparing Confusion matrix

library(caret)

library(SDMTools)

library(pROC)

library(Hmisc)

table.logit<-confusion.matrix(test.logit$Churn,predict.logit,threshold = 0.5)

table.logit

obs

pred 0 1

0 829 112

1 26 33

attr(,"class")

[1] "confusion.matrix"

Accuracy of model

ModelAccuracy<-(829+33)/(829+112+33+26)

ModelAccuracy

[1] 0.862

The accuracy of model is 86.2% which is good.

|  |  |
| --- | --- |
| **Accuracy=(TP+TN)/total** | **0.8620** |
| **Misclassification Rate or error rate** | **0.138** |
| **Sensitivity or recall or True Positive Rate** | **0.2275** |
| **Specificity = TN/actual no** | **0.9695** |

**b) Interpretation of other Model Performance Measures for logistic (KS, AUC, GINI)**

* ROC curve for test data

accuracy.logit<-

roc.logit<-roc(test.logit$Churn,predict.logit)

Setting levels: control = 0, case = 1

Setting direction: controls < cases

roc.logit

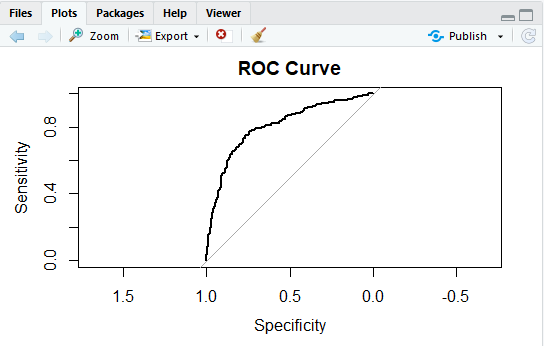
Call:

roc.default(response = test.logit$Churn, predictor = predict.logit)

Data: predict.logit in 855 controls (test.logit$Churn 0) < 145 cases (test.logit$Churn 1).

Area under the curve: 0.8067

plot(roc.logit,main="ROC Curve")



From the output it is seen that the area under the curve (AUC) is 80.67% percent which is >70% and hence is good fit indicator.

* KS score

library(ROCR)

ks1<-prediction(predict.logit,test.logit$Churn)

perf\_m3\_train<-performance(ks1,"tpr","fpr")

ks\_stats<-round(max(attr(perf\_m3\_train,'y.values')[[1]]-attr(perf\_m3\_train,'x.values')[[1]]),4)\*100

ks\_stats

[1] 52.01

Since the KS stats is 52.01% which is greater than 20% indicates good separation between the cumulative good rate & bad rate.

* Gini values for logistics regression

install.packages("ineq")

library(ineq)

gini\_stats=round(ineq(predict.logit,type="Gini"),4)\*100

gini\_stats

[1] 53.59

The gini value for the model is 53.59 which is a fair fit. Hence by comparing all the validation models the output is as below and the model is fit to predict.

**c) Remarks on Model validation exercise (Which model performed the best)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **AUC** | **KS Score** | **Gini Value** | **Confusion matrix** |
| **Accuracy in %** | **80.67** | **52.01** | **53.59** | **86.20** |

After comparing the accuracy of the AUC, KS Score and Gini value it can be seen that confusion matrix gives better performance in terms of accuracy in our Logistic Regression Model.

**4] Actionable Insights and Recommendations**

* From the above validation of the model we can state that logistic model developed using variables (AccountWeeks, ContractRenewal, DayMins, DataPlan, CustServCalls, DayCalls, OverageFee, RoamMins) are fit to Predict the churn rate of customers in telecom sector.
* Telecom Company should focus more on Data Plan, Customer Service Calls, Account Weeks, Contract Renewal, Daily Minutes, Daily Calls, Overage Fee and Roaming Minutes in order to reduce the churn rate in customers.