#Reading the DataSet

setwd("D:")

tele <-read.csv('D:\\Capston project\\telecomfinal.csv',header = TRUE,stringsAsFactors = TRUE)

##---------Creating Data Quality Report(dqr)-----------##

#Extracting Variable names

Variables<-names(tele)

dqr<-as.data.frame(Variables)

rm(Variables)

#Recording Data Type for each Variable

dqr$DataType<-sapply(tele,class)

#No. of Records for each Variable

dqr$No.ofRecords<-nrow(tele)

#Counting No. of Unique Values for each variable

for(i in 1:ncol(tele))

{

dqr$UniqueRecords[i]<-length(unique(tele[,i]))

}

#No.of observations available for each variable and its percentage

dqr$DataAvailable<-colSums(!is.na(tele))

dqr$AvailablePercentage<-round(colMeans(!is.na(tele)),4)

#Total and Percentage of Missing Values for each Variable

dqr$Missing<-colSums(is.na(tele))

dqr$MissingPercentage<-round(colMeans(is.na(tele)),4)

#Minimum, Maximum, Mean, Quantile Values for each Variable

for(i in 1:ncol(tele))

{

dqr$Minimum[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",min(tele[,i],na.rm=T),0),2)

dqr$Maximum[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",max(tele[,i],na.rm=T),0),2)

dqr$Mean[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",mean(tele[,i],na.rm=T),0),2)

dqr$fifthPercentile[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",quantile(tele[,i],p=0.05,na.rm=T),0),2)

dqr$tenthPercentile[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",quantile(tele[,i],p=0.10,na.rm=T),0),2)

dqr$twentyfifthPercentile[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",quantile(tele[,i],p=0.25,na.rm=T),0),2)

dqr$fiftythPercentile[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",quantile(tele[,i],p=0.50,na.rm=T),0),2)

dqr$seventyfifthPercentile[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",quantile(tele[,i],p=0.75,na.rm=T),0),2)

dqr$ninetythPercentile[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",quantile(tele[,i],p=0.90,na.rm=T),0),2)

dqr$ninetyfifthPercentile[i]<-round(ifelse(class(tele[,i])=="integer"|class(tele[,i])=="numeric",quantile(tele[,i],p=0.95,na.rm=T),0),2)

}

str(dqr)

#Exporting Data Quality Report

write.csv(dqr,"Data Quality Report.csv",row.names = T)

#Some Basic Sanity Checks

library(dplyr) #For Data Manipulation

options(scipen = 999) #For Scientific Notation

#Column names with Ascending Order

colnames(tele)

#Churn has been calculated as categorical variable with two outcomes 0 and 1 .

#i.e. '1' = Churned and '0' = Not Churned

#Target or Dependent Variable = "CHURN"

summary(tele$churn)

tele$churn

#Unique Values in Each Variable of Dataset

unique(tele)

#Summary Check

summary(tele)

#Checking for Missing Values in all the Columns

is.na(tele)

#Total Missing Values in Dataset

sum(is.na(tele)) #<-631020 numer of Missing Values

#To Get Datatypes of all Variables Available

str(tele)

# As we get following Variables as Factor

#crclscod, asl\_flag, prizm\_social\_one, area, refurb\_new, hnd\_webcap, marital, ethnic, dwlltype, dwllsize, mailordr, occu1

#wrkwoman, solflag, proptype, mailresp, cartype, car\_buy, children, csa, div\_type .

#retdays contains high missing Values

#Missing Value Treatement for retdays and creating dummy Variables.

#some Values for retdays have numeric number that accounts is to be continuous variable

#Missing Values for this Variable can be assumend to mean there have been no retention calls made by customer

summary(tele$retdays)

hist(tele$retdays)

str(tele$retdays)

is.na(tele$retdays)

sum(is.na(tele$retdays)) #<- 64143 number of missing values

sort(unique(tele$retdays),na.last = FALSE)

tele$retention<-ifelse(is.na(tele$retdays)==TRUE,0,1)

str(tele$retention) #Have Numeric Forms

summary(tele$retention)

#Since New Column Retention has Been Added so Checking for Column Names

names(tele)

telecom <- tele[,colMeans(is.na(tele))<=0.15]

#14 variables rejected based on the percentage of missing values cutoff

#After Looking into Data Dictionary

#Removing blck\_dat\_mean this variable is no longer useful

names(telecom)

telecom<-telecom[,-50]

names(telecom)

------------------------------------

#My Dataset has 67 Variables

#Deciling continuous variables basis target variable churn

str(telecom)

dim(telecom)

#Ratio of 0's and 1's in Complete Dataset

table(tele$churn)/nrow(tele)

table(telecom$churn)/nrow(telecom)

#Cut off Value for customers likely to churn is 0.2392114

#Continuous Variable Profilling

#Denoting all Variables with prefix as data\_(Name of Variable)

# 1.For mou\_mean = data\_mou\_mean / Mean Number of Monthly Mintues of Use

summary(telecom$mou\_Mean)

#There is clear downword trend in event rate (churn) as mou\_Mean increases

telecom%>%mutate(dec=ntile(mou\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(mou\_Mean,na.rm = TRUE),LessThan=max(mou\_Mean,na.rm = TRUE))->data\_mou\_Mean

data\_mou\_Mean

# 2.For Variable totmrc\_mean = data\_totmrc\_Mean / Monthly Recurring Charge is the base cost

#of the Calling Plan Regardless of Actual Minutes Used.

summary(telecom$totmrc\_Mean)

#No Clear Trend

telecom%>%mutate(dec=ntile(totmrc\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(totmrc\_Mean,na.rm = TRUE),LessThan=max(totmrc\_Mean,na.rm = TRUE))->data\_totmrc\_Mean

data\_totmrc\_Mean

# 3.For Variable rev\_range=data\_rev\_Range / Range of Revenue(Charge Amount)

summary(telecom$rev\_Range)

#No clear Trend

telecom%>%mutate(dec=ntile(rev\_Range,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(rev\_Range,na.rm = TRUE),LessThan=max(rev\_Range,na.rm = TRUE))->data\_rev\_Range

data\_rev\_Range

# 4.For Variable mou\_range = data\_mou\_Range / Range of number of Minutes of Use

summary(telecom$mou\_Range)

#No Clear Trend

telecom%>%mutate(dec=ntile(mou\_Range,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(mou\_Range,na.rm = TRUE),LessThan=max(mou\_Range,na.rm = TRUE))->data\_mou\_Range

data\_mou\_Range

# 5.For Variable change\_mou = data\_change\_mou / Percentage change in Monthly Minutes of use vs previous three month average

summary(telecom$change\_mou)

#No Clear Trend

telecom%>%mutate(dec=ntile(change\_mou,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(change\_mou,na.rm = TRUE),LessThan=max(change\_mou,na.rm = TRUE))->data\_change\_mou

data\_change\_mou

# 6.For Variable drop\_blk\_Mean = data\_drop\_blk\_Mean / Mean Number of Dropped or Blocked Calls

summary(telecom$drop\_blk\_Mean)

#No Clear Trend

telecom%>%mutate(dec=ntile(drop\_blk\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(drop\_blk\_Mean,na.rm = TRUE),LessThan=max(drop\_blk\_Mean,na.rm = TRUE))->data\_drop\_blk\_mean

data\_drop\_blk\_mean

# 7.For Variable drop\_vce\_range = data\_drop\_vce\_Range / Range of Number of Dropped(Failed) voice calls

summary(telecom$drop\_vce\_Range)

#No Clear Trend

telecom%>%mutate(dec=ntile(drop\_vce\_Range,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(drop\_vce\_Range,na.rm = TRUE),LessThan=max(drop\_vce\_Range,na.rm = TRUE))->data\_drop\_vce\_Range

data\_drop\_vce\_Range

# 8.For Variable owylis\_vce\_Range = data\_owylis\_vce\_Range / Range of number of Outbound Wireless to Wireless Voice Calls

summary(telecom$owylis\_vce\_Range)

#No Clear Trend

telecom%>%mutate(dec=ntile(owylis\_vce\_Range,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(owylis\_vce\_Range,na.rm = TRUE),LessThan=max(owylis\_vce\_Range,na.rm = TRUE))->data\_owylis\_vce\_Range

data\_owylis\_vce\_Range

# 9.For Variable mou\_opkv\_Range = data\_mou\_opkv\_Range / Range of unrounded minutes of use of off-peak voice calls

summary(telecom$mou\_opkv\_Range)

#Slightly Downwords trends in the Event Rate

telecom%>%mutate(dec=ntile(mou\_opkv\_Range,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(mou\_opkv\_Range,na.rm = TRUE),LessThan=max(mou\_opkv\_Range,na.rm = TRUE))->data\_mou\_opkv\_Range

data\_mou\_opkv\_Range

# 10.For Variable months = data\_months / Total Number of Months in Service

summary(telecom$months)

#No Clear Trend

#One Well-Known Observation is that ,Many people (42%) tend to leave company between 10-12 months

telecom%>%mutate(dec=ntile(months,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(months,na.rm = TRUE),LessThan=max(months,na.rm = TRUE))->data\_months

data\_months

# 11.For Variable totcalls = data\_totcalls / Total Number of calls over the life of the Customer

summary(telecom$totcalls)

#Slightly Increase Upward Trend in the event rate as totcalls increase

telecom%>%mutate(dec=ntile(totcalls,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(totcalls,na.rm = TRUE),LessThan=max(totcalls,na.rm = TRUE))->data\_totcalls

data\_totcalls

# 12.For Variable eqpdays = data\_eqpdays / Number of Days (Age) of current Equipment

summary(telecom$eqpdays)

#Slightly Increase Upward Trend in the event rate as eqpdays

telecom%>%mutate(dec=ntile(eqpdays,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(eqpdays,na.rm = TRUE),LessThan=max(eqpdays,na.rm = TRUE))->data\_eqpdays

data\_eqpdays

# 13.For Variable custcare\_Mean = data\_custcare\_Mean / Mean number of Customer care calls

#Less than 4 Deciles

summary(telecom$custcare\_Mean)

#Many Values are Zero and May not be Useful for the model

telecom%>%mutate(dec=ntile(custcare\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(custcare\_Mean,na.rm = TRUE),LessThan=max(custcare\_Mean,na.rm = TRUE))->data\_custcare\_Mean

hist(telecom$custcare\_Mean) #Positively Skewded

data\_custcare\_Mean

# 14.For Variable callwait\_mean = data\_callwait\_mean / Mean Number of call waiting calls

summary(telecom$callwait\_Mean)

#Many Values are Zero and May not be Useful for the model

telecom%>%mutate(dec=ntile(callwait\_Mean,4))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(callwait\_Mean,na.rm = TRUE),LessThan=max(callwait\_Mean,na.rm = TRUE))->data\_callwait\_Mean

data\_callwait\_Mean

# 15.For Variable iwylis\_vce\_mean = data\_iwylis\_vce\_mean / Mean Number of inbound wireless voice calls

summary(telecom$iwylis\_vce\_Mean)

telecom%>%mutate(dec=ntile(iwylis\_vce\_Mean,6))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(iwylis\_vce\_Mean,na.rm = TRUE),LessThan=max(iwylis\_vce\_Mean,na.rm = TRUE))->data\_iwylis\_vce\_Mean

data\_iwylis\_vce\_Mean

# 16.For Variable callwait\_range = data\_callwait\_Range/ Range of Number of call Waiting Calls

#Medain is Zero which shows 50% is data is Zero

summary(telecom$callwait\_Range)

#Many Values are Zero and May not be Useful for the model

telecom%>%mutate(dec=ntile(callwait\_Range,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(callwait\_Range,na.rm = TRUE),LessThan=max(callwait\_Range,na.rm = TRUE))->data\_callwait\_Range

data\_callwait\_Range

# 17.For Variable ccrndmou\_Range = data\_ccrndmou\_Range / Mean rounded minutes of use of customer care calls

#Medain is Zero

#Need to Omitted

summary(telecom$ccrndmou\_Range)

#Slightly Increase Upward Trend in the event rate as ccrndmou\_Range

telecom%>%mutate(dec=ntile(ccrndmou\_Range,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(ccrndmou\_Range,na.rm = TRUE),LessThan=max(ccrndmou\_Range,na.rm = TRUE))->data\_ccrndmou\_Range

data\_ccrndmou\_Range

# 18.For Variable adjqty = data\_adjqty / Billing Adjusted total number of calls over life of customer

summary(telecom$adjqty)

#Slightly Increase Upward Trend in the event rate

telecom%>%mutate(dec=ntile(adjqty,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(adjqty,na.rm = TRUE),LessThan=max(adjqty,na.rm = TRUE))->data\_adjqty

data\_adjqty

# 19.For Variable overrev\_mean = data\_overrev\_mean /Mean overage revenue

summary(telecom$ovrrev\_Mean)

#No Clear Trend

telecom%>%mutate(dec=ntile(ovrrev\_Mean,4))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(ovrrev\_Mean,na.rm = TRUE),LessThan=max(ovrrev\_Mean,na.rm = TRUE))->data\_overrev\_mean

data\_overrev\_mean

# 20.For Variable rev\_mean = data\_rev\_Mean / Men Monthly Revenue (Charge amount)

summary(telecom$rev\_Mean)

#No clear Trend

telecom%>%mutate(dec=ntile(rev\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(rev\_Mean,na.rm = TRUE),LessThan=max(rev\_Mean,na.rm = TRUE))->data\_rev\_Mean

data\_rev\_Mean

# 21.For Variable ovrmou\_Mean = data\_ovrmou\_Mean / Mean Overage Minutes of Use

summary(telecom$ovrmou\_Mean)

#Many Values are Zero ,and Maynot be usedful for the model

telecom%>%mutate(dec=ntile(ovrmou\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(ovrmou\_Mean,na.rm = TRUE),LessThan=max(ovrmou\_Mean,na.rm = TRUE))->data\_ovrmou\_Mean

data\_ovrmou\_Mean

# 22.For Variable comp\_vce\_mean = data\_comp\_vce\_Mean / Mean Number of Completed voice calls

summary(telecom$comp\_vce\_Mean)

#Slightly Downwards trend in the event rate as comp\_vce\_mean

telecom%>%mutate(dec=ntile(comp\_vce\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(comp\_vce\_Mean,na.rm = TRUE),LessThan=max(comp\_vce\_Mean,na.rm = TRUE))->data\_comp\_vce\_Mean

data\_comp\_vce\_Mean

# 23.For Variable plcd\_vce\_mean = data\_plcd\_vce\_Mean / Mean Number of attempted voice calls placed

summary(telecom$plcd\_vce\_Mean)

#Slightly Downwards trend in the event rate as plcd\_vce\_mean

telecom%>%mutate(dec=ntile(plcd\_vce\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(plcd\_vce\_Mean,na.rm = TRUE),LessThan=max(plcd\_vce\_Mean,na.rm = TRUE))->data\_plcd\_vce\_Mean

data\_plcd\_vce\_Mean

# 24.For Variable avg3mou = data\_avg3mou / Average Monthly Minutes of Use Over the previous three Months

summary(telecom$avg3mou)

#Slightly Downwards trend in the event rate as plcd\_vce\_mean

telecom%>%mutate(dec=ntile(avg3mou,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(avg3mou,na.rm = TRUE),LessThan=max(avg3mou,na.rm = TRUE))->data\_avg3mou

data\_avg3mou

# 25.For Variable avgmou = data\_avgmou / Average Monthly Minutes of Use Over the Life of the Customer

summary(telecom$avgmou)

#No Clear Trend

telecom%>%mutate(dec=ntile(avgmou,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(avgmou,na.rm = TRUE),LessThan=max(avgmou,na.rm = TRUE))->data\_avgmou

data\_avgmou

# 26.For Variable avg3qty = data\_av3qty / Average Monthly Number of calls over previous three months

summary(telecom$avg3qty)

#Clear Downward trend in the event rate as avg3qty

telecom%>%mutate(dec=ntile(avg3qty,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(avg3qty,na.rm = TRUE),LessThan=max(avg3qty,na.rm = TRUE))->data\_avg3qty

data\_avg3qty

# 27.For Variable avgqty = data\_avgqty / Average Monthly Number of calls over the life of the customer

summary(telecom$avgqty)

telecom%>%mutate(dec=ntile(avgqty,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(avgqty,na.rm = TRUE),LessThan=max(avgqty,na.rm = TRUE))->data\_avgqty

data\_avgqty

# 28.For Variable avg6mou = data\_avg6mou / Average Monthly Minutes of use over the previous six months

summary(telecom$avg6mou)

#Clear Downward trend in the event rate as avg6mou increased

telecom%>%mutate(dec=ntile(avg6mou,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(avg6mou,na.rm = TRUE),LessThan=max(avg6mou,na.rm = TRUE))->data\_avg6mou

data\_avg6mou

# 29.For Variable avg6qty = data\_avg6qty / Average Monthly number of calls over the previous six months

summary(telecom$avg6qty)

#Clear Downward trend in the event rate as avg6qty increased

telecom%>%mutate(dec=ntile(avg6qty,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(avg6qty,na.rm = TRUE),LessThan=max(avg6qty,na.rm = TRUE))->data\_avg6qty

data\_avg6qty

--------------------------------------------

# 39.For Variable age1=data\_age1 / Age of First Householder Member

#00=default

#Other values signify a valid age

summary(telecom$age1)

#Many values are Zero but not useful for this model

telecom%>%mutate(dec=ntile(age1,6))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(age1,na.rm = TRUE),LessThan=max(age1,na.rm = TRUE))->data\_age1

data\_age1

# 40.For Variable age2 = data\_age / Age of Secong Householder Member

#00 = default

#Other Values Signify a valid age

#Conatins less than 4 deciles , needs to be omited

summary(telecom$age2)

telecom%>%mutate(dec=ntile(age2,4))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(age2,na.rm = TRUE),LessThan=max(age2,na.rm = TRUE))->data\_age2

data\_age2

# 41.For Variable modles = data\_models / Number of Models issued

#contains less than 4 deciles ,need to be omited but Categorical Variable deal with it later

summary(telecom$models)

telecom%>%mutate(dec=ntile(models,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(models,na.rm = TRUE),LessThan=max(models,na.rm = TRUE))->data\_models

data\_models

# 42,For Variable hnd\_price = data\_hnd\_price / Current handset price ##

summary(telecom$hnd\_price)

#It turns out to be categorical need to be deal with it later.

telecom%>%mutate(dec=ntile(hnd\_price,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(hnd\_price,na.rm = TRUE),LessThan=max(hnd\_price,na.rm = TRUE))->data\_hnd\_price

data\_hnd\_price

# 43.For Variable actvsubs = data\_actvsubs / Number of Models issued

#contains less than 4 deciles , but cannot declined its a categorical variable

summary(telecom$actvsubs)

telecom%>%mutate(dec=ntile(actvsubs,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(actvsubs,na.rm = TRUE),LessThan=max(actvsubs,na.rm = TRUE))->data\_actvsubs

data\_actvsubs

# 44.For Variable uniqsubs = data\_uniqsubs / Number of Unique Subscribers in the Household

#getting less than 4 deciles , needs to be omit

summary(telecom$uniqsubs)

telecom%>%mutate(dec=ntile(uniqsubs,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(uniqsubs,na.rm = TRUE),LessThan=max(uniqsubs,na.rm = TRUE))->data\_uniqsubs

data\_uniqsubs

# 45.For Variable forgntvl = data\_forgntvl / Foreign travel dummy variable

#0 = No

#1 = Yes

#Contains less than 4 deciles ,needs to be omited but deal with it later as it is categorical Variable

summary(telecom$forgntvl)

telecom%>%mutate(dec=ntile(forgntvl,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(forgntvl,na.rm = TRUE),LessThan=max(forgntvl,na.rm = TRUE))->data\_forgntvl#Median is Zero

data\_forgntvl

# 46.For Variable opk\_dat\_Mean =data\_opk\_dat\_Mean / MEan number of off-peak data calls

#Medain is Zero

#Contains less than 4 deciles

summary(telecom$opk\_dat\_Mean)

telecom%>%mutate(dec=ntile(opk\_dat\_Mean,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(opk\_dat\_Mean,na.rm = TRUE),LessThan=max(opk\_dat\_Mean,na.rm = TRUE))->data\_opk\_dat\_Mean

data\_opk\_dat\_Mean

# 47.For Variable mtrcycle and 48. For Variable truck both are Categorical but stored as Numeric value

# 49.For Variable roam\_Mean = data\_roam\_Mean / Mean Number of roaming calls

#contains less than 4 deciles , need to be omit. as many values are Zero

summary(telecom$roam\_Mean)

telecom%>%mutate(dec=ntile(roam\_Mean,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(roam\_Mean,na.rm = TRUE),LessThan=max(roam\_Mean,na.rm = TRUE))->data\_roam\_mean

data\_roam\_mean

# 50.For Variable recv\_sms\_Mean = data\_recv\_sms\_Mean / Mean Number of Received sms calls

#Contains less than 4 Deciles ,need to be omit as many values are zero also

summary(telecom$recv\_sms\_Mean)

#Median is Zero

telecom%>%mutate(dec=ntile(recv\_sms\_Mean,4))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(recv\_sms\_Mean,na.rm = TRUE),LessThan=max(recv\_sms\_Mean,na.rm = TRUE))->data\_recv\_sms\_Mean

data\_recv\_sms\_Mean

# 51.For Variable mou\_pead\_Mean = data\_mou\_pead\_Mean

summary(telecom$mou\_pead\_Mean)

#Median is Zero

telecom%>%mutate(dec=ntile(mou\_pead\_Mean,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(mou\_pead\_Mean,na.rm = TRUE),LessThan=max(mou\_pead\_Mean,na.rm = TRUE))->data\_mou\_pead\_Mean

data\_mou\_pead\_Mean

# 52.For Variable da\_mean = data\_da\_Mean / Mean number of Directory assisted calls

summary(telecom$da\_Mean)

#Many values are zero not useful for model

telecom%>%mutate(dec=ntile(da\_Mean,4))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(da\_Mean,na.rm = TRUE),LessThan=max(da\_Mean,na.rm = TRUE))->data\_da\_mean

data\_da\_mean

# 53.For Variable da\_Range = data\_da\_Range / RAnge of Number of Directory assisted calls

summary(telecom$da\_Range)

#Many values are zero cannot be useful for model

telecom%>%mutate(dec=ntile(da\_Range,4))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(da\_Range,na.rm = TRUE),LessThan=max(da\_Range,na.rm = TRUE))->data\_da\_Range

data\_da\_Range

# 54.For Variable datovr\_mean =data\_datovr\_mean / Mean revenue of Data Overage

#Contains less than 4 deciles , need to be omit and many values are zero also

summary(telecom$datovr\_Mean)

telecom%>%mutate(dec=ntile(datovr\_Mean,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(datovr\_Mean,na.rm = TRUE),LessThan=max(datovr\_Mean,na.rm = TRUE))->data\_datovr\_Mean

data\_datovr\_Mean

# 55.For Variable datovr\_Range = data\_datovr\_Range / RAnge of revenue of data overage

#Median is Zero

#Contains less than 4 deciles ,need to be omit

summary(telecom$datovr\_Range)

telecom%>%mutate(dec=ntile(datovr\_Range,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(datovr\_Range,na.rm = TRUE),LessThan=max(datovr\_Range,na.rm = TRUE))->data\_datovr\_Range

data\_datovr\_Range

# 56.For variable drop\_dat\_mean = data\_drop\_dat\_Mean / mean number of dropped (failed) data calls

#Contains less than 4 deciles ,need to be omit and many values are zero also .

summary(telecom$drop\_dat\_Mean)

telecom%>%mutate(dec=ntile(drop\_dat\_Mean,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(drop\_dat\_Mean,na.rm = TRUE),LessThan=max(drop\_dat\_Mean,na.rm = TRUE))->data\_drop\_dat\_Mean

data\_drop\_dat\_Mean

# 57.For Variable drop\_vce\_Mean = data\_drop\_vce\_Mean / mean number of dropped (Failed) voice calls

summary(telecom$drop\_vce\_Mean)

telecom%>%mutate(dec=ntile(drop\_vce\_Mean,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(drop\_vce\_Mean,na.rm = TRUE),LessThan=max(drop\_vce\_Mean,na.rm = TRUE))->data\_drop\_vce\_Mean

data\_drop\_vce\_Mean

# 58.For Variable adjmou = data\_adjmou / Billing adjusted total minutes of use over the life of the customer

summary(telecom$adjmou)

telecom%>%mutate(dec=ntile(adjmou,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(adjmou,na.rm = TRUE),LessThan=max(adjmou,na.rm = TRUE))->data\_adjmou

data\_adjmou

# 59.For Variable totrev=data\_totrev / Total Revenue

summary(telecom$totrev)

telecom%>%mutate(dec=ntile(totrev,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(totrev,na.rm = TRUE),LessThan=max(totrev,na.rm = TRUE))->data\_totrev

data\_totrev

# 60.For variable adjrev=data\_adjrev / Billing adjusted total revenue over the life of the customer

summary(telecom$adjrev)

#Slight upward trends

telecom%>%mutate(dec=ntile(adjrev,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(adjrev,na.rm = TRUE),LessThan=max(adjrev,na.rm = TRUE))->data\_adj\_rev

data\_adj\_rev

# 61.For Var5iable avgrev = data\_avgrev / Average monthly revenue over the life of the customer

summary(telecom$avgrev)

#No clear trend

telecom%>%mutate(dec=ntile(avgrev,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(avgrev,na.rm = TRUE),LessThan=max(avgrev,na.rm = TRUE))->data\_avgrev

data\_avgrev

# 62.For Variable comp\_dat\_mean = data\_comp\_dat\_mean / Mean number of completed data calls

#has to relted with complete mean then omit

summary(telecom$comp\_dat\_Mean)

telecom%>%mutate(dec=ntile(comp\_dat\_Mean,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(comp\_dat\_Mean,na.rm = TRUE),LessThan=max(comp\_dat\_Mean,na.rm = TRUE))->data\_comp\_dat\_Mean

#Medain is Zero

data\_comp\_dat\_Mean

# 63.For Variable plcd\_dat\_mean = data\_plcd\_dat\_Mean / Mean number of attempted data calls placed

#has to relted with plcd attempt\_mean then omit.

summary(telecom$plcd\_dat\_Mean)

telecom%>%mutate(dec=ntile(plcd\_dat\_Mean,2))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(plcd\_dat\_Mean,na.rm = TRUE),LessThan=max(plcd\_dat\_Mean,na.rm = TRUE))->data\_plcd\_dat\_Mean

data\_plcd\_dat\_Mean

#64. for Variable retention = data\_retention

summary(telecom$retention)

telecom%>%mutate(dec=ntile(retention,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(retention,na.rm = TRUE),LessThan=max(retention,na.rm = TRUE))->data\_retention

data\_retention

#One of the area to understand is that the effect of network and service quality influencing churn rate

#To get the top idea about network and service quality we can drive the following variable

telecom$compl\_vce\_percentage<-(telecom$comp\_vce\_Mean/telecom$plcd\_vce\_Mean)

telecom$compl\_dat\_percentage<-(telecom$comp\_dat\_Mean/telecom$plcd\_dat\_Mean)

# 65.For variable compl\_vce\_percentage

summary(telecom$compl\_vce\_percentage)

telecom%>%mutate(dec=ntile(compl\_vce\_percentage,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(compl\_vce\_percentage,na.rm = TRUE),LessThan=max(compl\_vce\_percentage,na.rm = TRUE))->data\_compl\_vce\_percentage

data\_compl\_vce\_percentage

# 66.For variable compl\_dat\_percentage

summary(telecom$compl\_dat\_percentage)

telecom%>%mutate(dec=ntile(compl\_dat\_percentage,10))%>%group\_by(dec)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,2),GreaterThan=min(compl\_dat\_percentage,na.rm = TRUE),LessThan=max(compl\_dat\_percentage,na.rm = TRUE))->data\_compl\_dat\_percentage

data\_compl\_dat\_percentage

#Removing compl\_dat\_percentage from the dataset

names(telecom)

telecom<-telecom[,-69]

telecom\_Declined\_continuous<-rbind(data\_mou\_Mean,data\_totmrc\_Mean,data\_rev\_Range,data\_mou\_Range,data\_change\_mou,data\_drop\_blk\_mean,data\_drop\_vce\_Mean,

data\_owylis\_vce\_Range,data\_mou\_opkv\_Range,data\_months,data\_totcalls,data\_eqpdays,data\_iwylis\_vce\_Mean,data\_adjqty,

data\_rev\_Mean,data\_comp\_vce\_Mean,data\_plcd\_vce\_Mean,data\_avg3mou,data\_avgmou,data\_avg3qty,data\_avgqty,

data\_avg6mou,data\_avg6qty,data\_hnd\_price,data\_adjmou,data\_adj\_rev,data\_avgrev,data\_compl\_vce\_percentage)

str(telecom\_Declined\_continuous)

write.csv(telecom\_Declined\_continuous,'Declined Output Continuous Variables.csv',row.names = F)

#We can remove the variables based on declined binning outputs

#Lets Remove those variable which either have many zero or no Variablity these variables do not add any significance to the model

#custcare\_Mean,callwait\_Mean,call,wait\_Range,ccrndmou\_Range,ovrmou\_Mean,opk\_dat\_Mean,roam\_Mean,recv\_sms\_Mean,mou\_pead\_Mean,

#da\_Mean,da\_Range,datovr\_Mean,datovr\_Range,drop\_dat\_Mean,comp\_dat\_Mean,plcd\_dat\_Mean

names(telecom)

telecom<-telecom[,-c(13,14,16,17,21,45,48:50,54:58,65,66)]

dim(telecom)

names(telecom)

#Before checking event rates for the Categorical variables, lets convert the variable which are supposed to be categorical

#But stored as numeric into Categorical Variable

telecom$models <- as.factor(telecom$models)

telecom$hnd\_price <- as.factor(telecom$hnd\_price)

telecom$actvsubs <- as.factor(telecom$actvsubs)

telecom$uniqsubs <- as.factor(telecom$uniqsubs)

telecom$forgntvl <- as.factor(telecom$forgntvl)

telecom$truck <- as.factor(telecom$truck)

telecom$mtrcycle <- as.factor(telecom$mtrcycle)

str(telecom)

names(telecom)

##########################################################

#CATEGORICAL VARIABLES PROFILLING AND DECILING

##Event Rate For Each Level in a Categorical Variable

# 1. For the Variable crclscod=data\_crclscod / Credit class code

#A Represtents best rating ,Z represents worst rating

#Due to low Percent of churn which are less than 5%

#Which can show insignificant Behaviour

library(dplyr)

summary(telecom$crclscod)

telecom%>%group\_by(crclscod)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_crclscod

class(telecom$crclscod)

data\_crclscod

#2. For Variable asl\_flag=data\_asl\_flag / Account Spending Limit

#N = NO

#Y = Yes

#There is somes differrence in the event rate across different levels

summary(telecom$asl\_flag)

telecom%>%group\_by(asl\_flag)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_asl\_flag

class(telecom$asl\_flag) #Trend in Churn Perc

data\_asl\_flag

#3. For Variable prizm\_social\_one = data\_prizm\_social\_one / Social Group letter only

#Based on degree of population density of area

#C = City

#R = Rural

#S = Suburban

#T = Town

#U = Urban

library(tidyverse) # for using fct\_explicit\_na function

summary(telecom$prizm\_social\_one)

telecom%>%group\_by(fct\_explicit\_na(prizm\_social\_one))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_prizm\_social\_one

class(telecom$prizm\_social\_one)

data\_prizm\_social\_one

#4. For Variable area = data\_area / Geographic Area

#There are many levels we will try to find the levels with similar event rate

#Event Rate Matches with few other levels ,we can use this info to treat missing values

summary(telecom$area)

telecom%>%group\_by(fct\_explicit\_na(area))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_area

class(telecom$area)

data\_area

#5. For Variable refurb\_new = data\_refurb\_new # Handset : Refurbished or new

#N = New

#R = Refurbished

#There is some difference in the event rate across different levels

summary(telecom$refurb\_new)

telecom%>%group\_by(fct\_explicit\_na(refurb\_new))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_refurb\_new

class(telecom$refurb\_new)

data\_refurb\_new

#6. For Variable hnd\_webcap = data\_hnd\_webcap / Handset Web Capability

#WC = Web Capable

#WC = Web Capable Mini-Browser

#NA = Not Applicable

#Unkw = Unable to collect these data

#There is some difference in the event rate across different level

summary(telecom$hnd\_webcap)

telecom%>%group\_by(fct\_explicit\_na(hnd\_webcap))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_hnd\_webcap

class(telecom$hnd\_webcap)

data\_hnd\_webcap

#7. For Variable Marital = data\_marital / Marital Status

#Indicates if Anyone in the household is Married

#U = Unknown

#M = Married

#S = Single

#B = Inferred Single

#A = Inferred Married

summary(telecom$marital)

telecom%>%group\_by(fct\_explicit\_na(marital))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_marital

class(telecom$marital)

data\_marital

#8. For Variable ethnic = data\_ethnic / Ethnicity roll-up code

#B = Asian (Non-Oriental)

#D = Southern European

#F = French

#G = German

#H = Hispanic

#I = Italian

#J = Jewish

#M = Miscellaneous

#N = Northern European

#O = Asian

#P = Polynesian

#R = Arab

#S = Scottish / Irish

#U = Unknown

#Z = African-American

summary(telecom$ethnic)

telecom%>%group\_by(fct\_explicit\_na(ethnic))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_ethnic

class(telecom$ethnic)

data\_ethnic

#9. For Variable car\_buy = data\_car\_buy / New or Used car buyer

#Indicates a History of new car buying in the household.

summary(telecom$car\_buy)

telecom%>%group\_by(fct\_explicit\_na(car\_buy))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_car\_buy

class(telecom$car\_buy)

data\_car\_buy

#10. For Variable models = data\_models / Number of models issued

summary(telecom$models)

telecom%>%group\_by(fct\_explicit\_na(models))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_models

class(telecom$models)

data\_models

#11. For Variable hnd\_price = data\_hnd\_price / current handset price

summary(telecom$hnd\_price)

telecom%>%group\_by(fct\_explicit\_na(hnd\_price))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_hnd\_price

class(telecom$hnd\_price)

data\_hnd\_price

#12.For Variable actvsubs = data\_actvsubs / Number of active subscribers in household

summary(telecom$actvsubs)

telecom%>%group\_by(fct\_explicit\_na(actvsubs))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_actvsubs

class(telecom$actvsubs)

data\_actvsubs

#13.For Variable uniqsubs = data\_uniqsubs / Number of Unique Subscribers in the household

summary(telecom$uniqsubs)

telecom%>%group\_by(fct\_explicit\_na(uniqsubs))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_uniqsubs

class(telecom$uniqsubs)

data\_uniqsubs

#14. For variable forgntvl

summary(telecom$forgntvl)

telecom%>%group\_by(fct\_explicit\_na(forgntvl))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_forgntvl

class(telecom$forgntvl)

data\_forgntvl

#15. For Variable mtrcycle = data\_mtrcycle / Motorcycle indicator

#Indicates motocycle owner in household

summary(telecom$mtrcycle)

telecom%>%group\_by(fct\_explicit\_na(mtrcycle))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_mtrcycle

class(telecom$mtrcycle)

data\_mtrcycle

#16. For Variable truck = data\_truck / Truck indicator

#Indicator a truck owner in a household

summary(telecom$truck)

telecom%>%group\_by(fct\_explicit\_na(truck))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_truck

class(telecom$truck)

data\_truck

#For Variable csa = q10 / Communications local service area

#Refer to Specific Location of the customer , usually indication city

summary(telecom$csa)

telecom%>%group\_by(fct\_explicit\_na(csa))%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_csa

class(telecom$csa) #Many Levels Showing less than 50% churn rate

#and having high levels may show insignificant nature during model running

data\_csa

#now those categorical variable which do not have factor nature

#So also Converting those into factor datatype

#For Variable retention = data\_retention

#Retention calls include any calls from the customer regarding loyalty or retention , e.g. contract renewal , relating competitor's offer ,etc.

#Missing Values for this variable can be assumed to mean there have been no retention calls made by the customer

summary(telecom$retention)

telecom%>%group\_by(retention)%>%summarise(n=sum(churn),N=n(),churn\_perc=round(n/N,4))->data\_retention

class(telecom$retention)

names(telecom)

data\_retention

#Omit where there is not enough change in the event rate at different levels

#Variables having churn rate less than 5% No need

telecom<-telecom[,-c(39:41,43)] #car\_buy,truck,mtrcycle,forgntvl

names(telecom)

dim(telecom)

#DATA PREPARATION

# (1) OUTLIER TREATMENT

#First lets treat continuous variable

#To detect the outlier

#Graphical method we are using is boxlot .

colnames(telecom)

names(telecom)

list<-names(telecom)

#Removing categorical Variable

list<-list[-c(25:42,50,51)]

list

#boxplot(telecom$churn,telecom$prizm\_social\_one)

#Plotting the outlier

dev.off()

par("mar")

par(mar=c(1,1,1,1))

par(mfrow=c(3,11))

for (i in 1:length(list))

{

boxplot(telecom[,list[i]],main=list[i])

} #Warnings

#Outlier Treatment

for (i in 1:length(list))

{

a<-boxplot(telecom[,list[i]],main=list[i])

out<-a$out

index<-which(telecom[,list[i]]%in%a$out)

telecom[index,list[i]]<-mean(telecom[,list[i]],na.rm = TRUE)

rm(a)

rm(out)

}

#Checking after treatment

for (i in 1:length(list))

{

boxplot(telecom[,list[i]],main=list[i])

}

for (i in 1:length(list))

{

plot(telecom[,list[i]],main=list[i])

}

#lets treat Categorical Variable

colSums(is.na(telecom))

#For Variable crclsod

#There are no missing values

sum(is.na(telecom$crclscod))

#For variable asl\_flag

#There are no missing values

sum(is.na(telecom$asl\_flag))

#For variable prize\_social\_one

sum(is.na(telecom$prizm\_social\_one)) #<- 4751 missing values

missing\_p<-which(is.na(telecom$prizm\_social\_one))

telecom$prizm\_social\_one[missing\_p]<-'T'

#Replacing with level T as there are similarity in the event rate

#As Observed above ,levels C,S and U have similar rate event we can combine them together

index\_p <-which(telecom$prizm\_social\_one=='S' | telecom$prizm\_social\_one=='U')

telecom$prizm\_social\_one[index\_p]<-'C'

unique(telecom$prizm\_social\_one)

summary(telecom$prizm\_social\_one)

#For Variable area

sum(is.na(telecom$area)) #<- 18 missing values

summary(telecom$area)#<- Event rate for the missing value is same as the CENTRAL/SOUTH TEXAS AREA

missing\_a<-which(is.na(telecom$area))

missing\_a

telecom$area[missing\_a]<-'CENTRAL/SOUTH TEXAS AREA'

#For variable refurb\_new

sum(is.na(telecom$refurb\_new)) #Only one missing observation

missing\_r<-which(is.na(telecom$refurb\_new))

telecom[missing\_r,] #Dataset has either 0 or missing values for many variables hence we need to omit it from dataset

telecom[-missing\_r,]->telecom

dim(telecom)

#For variable hnd\_webcap

sum(is.na(telecom$hnd\_webcap)) #<-6062 missing values

summary(telecom$hnd\_webcap) #We can replace missing values with new 'WC' level based on event rate Similarity

missing\_hnd<-which(is.na(telecom$hnd\_webcap))

telecom$hnd\_webcap[missing\_hnd]<- 'WC'

#For Variable Marital

sum(is.na(telecom$marital)) #<-1152 Missing Values

summary(telecom$marital) #We can add missing values to level 'S' based on event rate similarity

missing\_m<-which(is.na(telecom$marital))

telecom$marital[missing\_m]<-'S'

#For Variable ethnic

sum(is.na(telecom$ethnic)) #<-1152 missing values

summary(telecom$ethnic) #We can add missing values to level 'M' based on event rate similarity

missing\_e<-which(is.na(telecom$ethnic))

telecom$ethnic[missing\_e]<-'M'

#For Variable hnd\_price

sum(is.na(telecom$hnd\_price)) #<-635 missing values

missing\_hnd\_price<-which(is.na(telecom$hnd\_price))

telecom$hnd\_price[missing\_hnd\_price]<-'299.9899902'

summary(telecom$hnd\_price)

#For Variable csa

sum(is.na(telecom$csa)) #<-18 missing values

missing\_c<-which(is.na(telecom$csa))

summary(telecom$csa) #we can replace missing values with level 'DALFTW817'based on event rate similarity

telecom$csa[missing\_c]<-'DALFTW817'

#Now Lets Deal with Continuous Variables

colSums(is.na(telecom))

#For Variable mou\_Mean

summary(telecom$mou\_Mean)# 181 missing values Omitting them

missing\_mou\_Mean<-which(is.na(telecom$mou\_Mean))

telecom<-telecom[-missing\_mou\_Mean,]

summary(telecom$mou\_Mean)

dim(telecom)

#For Variable change\_mou

summary(telecom$change\_mou) #233 missing values Omitting them and have a clean dataset

missing\_change\_mou<-which(is.na(telecom$change\_mou))

telecom<-telecom[-missing\_change\_mou,]

summary(telecom$change\_mou)

dim(telecom)

library(dplyr)

#for variable avg6mou

summary(telecom$avg6mou) #<-2029 missing values

#Lets impute the missing values with the average of 10th Decile Obervation

telecom%>%mutate(dec=ntile(avg6mou,10))->telecom

telecom%>%filter(dec==10)%>%summarise(avg=sum(avg6mou/n()))

missing\_avg6mou<-which(is.na(telecom$avg6mou))

telecom$avg6mou[missing\_avg6mou]<-1688.485

#For Variable avg6qty

summary(telecom$avg6qty) #<-2029 missing Values

#Lets impute the missing values with average of 10th Decile Observation

telecom%>%mutate(dec=ntile(avg6qty,10))->telecom

telecom%>%filter(dec==10)%>%summarise(avg=sum(avg6qty/n()))

missing\_avg6qty<-which(is.na(telecom$avg6qty))

telecom$avg6qty[missing\_avg6mou]<-610.5087

#For variable age1

summary(telecom$age1)#<-1144 missing values lets impute them with median as data is skewed

missing\_age1<-which(is.na(telecom$age1))

hist(telecom$age1) #Data is skewed imputing with median

telecom$age1[missing\_age1]<-36.00

telecom%>%filter(age1==0)%>%summarise(n=n()) #17888 values are zero and we know that can't be Zero

#but we can convert age1 variable to categorical and named missing or 0 is equals None

telecom$age1\_dummy<-ifelse(telecom$age1=='0','None',

ifelse(telecom$age1<=30,'Young',

ifelse(telecom$age1>30 & telecom$age1<=55,'Mid-Age','Senior')))

unique(telecom$age1\_dummy)

str(telecom$age1\_dummy)

#Converting to Factor

telecom$age1\_dummy <- as.factor(telecom$age1\_dummy)

#For Variable age2

summary(telecom$age2)

telecom%>%filter(age2==0)%>%summarise(n=n()) #Half of the 'age2' values seems to be zero may not be useful

#Lets omit age2 from the data and age1 also as i created dummy variable for age1 data

names(telecom)

telecom<-telecom[,-c(33,34)]

View(telecom)

#We have Missing Values in compl\_vce\_percentage due to the fact tht corrospondes to values of 'plcd\_vce\_Mean' and 'plcd\_dat\_mean'are Zeros

#While creating this variable due to zeros in numerator and denomiator ,the values become NaN

#We need to Omit this observation

summary(telecom$compl\_vce\_percentage)

missing\_compl<-which(is.na(telecom$compl\_vce\_percentage))

View(missing\_compl)

telecom<-telecom[-missing\_compl,]

dim(telecom)

#Let's just create another column optimum for checking the ratio of overage revenue / total revenue

#i.e overrev\_mean / totrev

telecom$optimum <- telecom$ovrrev\_Mean/telecom$totrev

summary(telecom$optimum)

dim(telecom)

#dec variable is longer useful

names(telecom)

telecom<-telecom[,-47]

dim(telecom)

#Finally there is no NA's or missing values found in dataset

colSums(is.na(telecom))

View(telecom)

dim(telecom)

#checking overall data cleaning impacted on churn or not

sum(telecom$churn)/nrow(telecom) #Churn rate is 23.35 we havent lost lot of information

#Making one copy of dataset in case of data loss

final\_telecom<-telecom

dim(final\_telecom)

names(final\_telecom)

#################################################################################################

##BUILDING LOGISTIC REGRESSION

#Splitting the data set into training an testing dataset

set.seed(200)

partition<-sample(nrow(telecom),0.70\*nrow(telecom),replace = FALSE)

training<-telecom[partition,]

testing<-telecom[-partition,]

#Calculating Churn Rate for both training and testing

table(training$churn)/nrow(training)

table(testing$churn)/nrow(testing)

#Almost Near values of both Training and Testing

colnames(telecom)

#Building a Model

mod1<-glm(churn~.,data = training,family = 'binomial') #Error: cannot allocate vector of size 295.8 Mb

summary(mod1)

#Since step takes more time for individual step,doing it manually

#Lets create dummy variable for the representive of significant levels for Categorical Variable

summary(training$age1\_dummy)

training$age1\_midage <- ifelse(training$age1\_dummy=='mid age',1,0)

testing$age1\_midage <- ifelse(testing$age1\_dummy=='mid age',1,0)

training$age1\_old <- ifelse(training$age1\_dummy=='old',1,0)

testing$age1\_old <- ifelse(testing$age1\_dummy=='old',1,0)

training$age1\_young <- ifelse(training$age1\_dummy=='young',1,0)

testing$age1\_young <- ifelse(testing$age1\_dummy=='young',1,0)

#for variable prizm\_social\_one\_a

summary(training$prizm\_social\_one)

training$prizm\_social\_one\_R <- ifelse(training$prizm\_social\_one=='R',1,0)

testing$prizm\_social\_one\_R <- ifelse(testing$prizm\_social\_one=='R',1,0)

training$prizm\_social\_one\_T <- ifelse(training$prizm\_social\_one=='T',1,0)

testing$prizm\_social\_one\_T <- ifelse(testing$prizm\_social\_one=='T',1,0)

#For Variable uniqsubs

summary(telecom$uniqsubs)

training$uniqsubs\_2 <- ifelse(training$uniqsubs=='2',1,0)

testing$uniqsubs\_2 <- ifelse(testing$uniqsubs=='2',1,0)

training$uniqsubs\_3 <- ifelse(training$uniqsubs=='3',1,0)

testing$uniqsubs\_3 <- ifelse(testing$uniqsubs=='3',1,0)

training$uniqsubs\_4 <- ifelse(training$uniqsubs=='4',1,0)

testing$uniqsubs\_4 <- ifelse(testing$uniqsubs=='4',1,0)

training$uniqsubs\_5 <- ifelse(training$uniqsubs=='5',1,0)

testing$uniqsubs\_5 <- ifelse(testing$uniqsubs=='5',1,0)

training$uniqsubs\_6 <- ifelse(training$uniqsubs=='6',1,0)

testing$uniqsubs\_6 <- ifelse(testing$uniqsubs=='6',1,0)

training$uniqsubs\_7 <- ifelse(training$uniqsubs=='7',1,0)

testing$uniqsubs\_7 <- ifelse(testing$uniqsubs=='7',1,0)

training$uniqsubs\_9 <- ifelse(training$uniqsubs=='9',1,0)

testing$uniqsubs\_9 <- ifelse(testing$uniqsubs=='9',1,0)

#For Variable hnd\_price

summary(training$hnd\_price)

training$hnd\_price\_129.98 <- ifelse(training$hnd\_price =='129.9899902',1,0)

testing$hnd\_price\_129.98 <- ifelse(testing$hnd\_price == '129.9899902',1,0)

training$hnd\_price\_199.98 <- ifelse(training$hnd\_price =='199.9899902',1,0)

testing$hnd\_price\_199.98 <- ifelse(testing$hnd\_price == '199.9899902',1,0)

training$hnd\_price\_249.98 <- ifelse(training$hnd\_price =='249.9899902',1,0)

testing$hnd\_price\_249.98 <- ifelse(testing$hnd\_price == '249.9899902',1,0)

training$hnd\_price\_299.98 <- ifelse(training$hnd\_price =='299.9899902',1,0)

testing$hnd\_price\_299.98 <- ifelse(testing$hnd\_price == '299.9899902',1,0)

#For Variable ethnic

summary(training$ethnic)

training$ethnic\_C <- ifelse(training$ethnic == "C",1,0)

testing$ethnic\_C <- ifelse(testing$ethnic == "C",1,0)

training$ethnic\_N <- ifelse(training$ethnic == "N",1,0)

testing$ethnic\_N <- ifelse(testing$ethnic == "N",1,0)

training$ethnic\_O <- ifelse(training$ethnic == "O",1,0)

testing$ethnic\_O <- ifelse(testing$ethnic == "O",1,0)

training$ethnic\_S <- ifelse(training$ethnic == "S",1,0)

testing$ethnic\_S <- ifelse(testing$ethnic == "S",1,0)

training$ethnic\_U <- ifelse(training$ethnic == "U",1,0)

testing$ethnic\_U <- ifelse(testing$ethnic == "U",1,0)

training$ethnic\_Z <- ifelse(training$ethnic == "Z",1,0)

testing$ethnic\_Z <- ifelse(testing$ethnic == "Z",1,0)

#For Variable Area

summary(training$area)

training$area\_calnrth <- ifelse(training$area == 'CALIFORNIA NORTH AREA',1,0)

testing$area\_calnrth <- ifelse(testing$area == 'CALIFORNIA NORTH AREA',1,0)

training$area\_texas <- ifelse(training$area == 'CENTRAL/SOUTH TEXAS AREA',1,0)

testing$area\_texas <- ifelse(testing$area == 'CENTRAL/SOUTH TEXAS AREA',1,0)

training$area\_nrthflorida <- ifelse(training$area == 'NORTH FLORIDA AREA',1,0)

testing$area\_nrthflorida <- ifelse(testing$area == 'NORTH FLORIDA AREA',1,0)

training$area\_nrthwst <- ifelse(training$area == 'NORTHWEST/ROCK MOUNTAIN AREA',1,0)

testing$area\_nrthwst <- ifelse(testing$area == 'NORTHWEST/ROCK MOUNTAIN AREA',1,0)

training$area\_southflorida <- ifelse(training$area == 'SOUTH FLORIDA AREA',1,0)

testing$area\_southflorida <- ifelse(testing$area == 'SOUTH FLORIDA AREA',1,0)

training$area\_southwest <- ifelse(training$area == 'SOUTHWEST AREA',1,0)

testing$area\_southwest <- ifelse(testing$area == 'SOUTHWEST AREA',1,0)

training$area\_tenese <- ifelse(training$area == 'TENNESSEE AREA',1,0)

testing$area\_tenese <- ifelse(testing$area == 'TENNESSEE AREA',1,0)

#for variable asl\_flag

summary(training$asl\_flag)

training$asl\_flag\_y <- ifelse(training$asl\_flag == 'Y',1,0)

testing$asl\_flag\_y <- ifelse(testing$asl\_flag == 'Y',1,0)

#For Variable refurb\_new

summary(training$refurb\_new)

training$refurb\_new\_r <- ifelse(training$refurb\_new == 'R',1,0)

testing$refurb\_new\_r <- ifelse(testing$refurb\_new == 'R',1,0)

#For Variable Marital

summary(training$marital)

training$marital\_s<- ifelse(training$marital=='S',1,0)

testing$marital\_s<-ifelse(testing$marital=='S',1,0)

training$marital\_a<- ifelse(training$marital=='A',1,0)

testing$marital\_a<-ifelse(testing$marital=='A',1,0)

training$marital\_b<- ifelse(training$marital=='B',1,0)

testing$marital\_b<-ifelse(testing$marital=='B',1,0)

training$marital\_m<- ifelse(training$marital=='M',1,0)

testing$marital\_m<-ifelse(testing$marital=='M',1,0)

training$marital\_u<- ifelse(training$marital=='U',1,0)

testing$marital\_u<-ifelse(testing$marital=='U',1,0)

#For Variable Models

summary(training$models)

summary(testing$models)

training$model\_2<-ifelse(training$models == '2',1,0)

testing$model\_2<-ifelse(testing$models == '2',1,0)

training$model\_3<-ifelse(training$models == '3',1,0)

testing$model\_3<-ifelse(testing$models == '3',1,0)

training$model\_4<-ifelse(training$models == '4',1,0)

testing$model\_4<-ifelse(testing$models == '4',1,0)

training$model\_5<-ifelse(training$models == '5',1,0)

testing$model\_5<-ifelse(testing$models == '5',1,0)

training$model\_8<-ifelse(training$models == '8',1,0)

testing$model\_8<-ifelse(testing$models == '8',1,0)

#Create for Event rate (churn rate) in trainting and testing datasets,to make sure the samples are not baised

sum(telecom$churn)/nrow(telecom)

sum(training$churn)/nrow(training)

colnames(training)

summary(training)

step(model1,direction = 'both')

model2 <- glm(churn ~ mou\_Mean + totmrc\_Mean + rev\_Range + mou\_Range + change\_mou + drop\_blk\_Mean + drop\_vce\_Range + owylis\_vce\_Range

+ mou\_opkv\_Range + months + totcalls + eqpdays + iwylis\_vce\_Mean + adjqty + ovrrev\_Mean + rev\_Mean + comp\_vce\_Mean

+ plcd\_vce\_Mean + avg3mou + avgmou + avg3qty + avgqty + avg6mou + avg6qty + crclscod + asl\_flag\_y +area\_tenese + area\_southwest

+ area\_southflorida +area\_nrthwst +area\_nrthflorida +area\_texas +area\_calnrth + refurb\_new\_r + hnd\_price\_129.98 + hnd\_price\_199.98

+ hnd\_price\_249.98 + hnd\_price\_299.98 + marital\_a + marital\_b + marital\_m +marital\_s +marital\_u + model\_2 +model\_3 +model\_4

+ model\_5+ model\_8 + ethnic\_N + ethnic\_S + ethnic\_U + ethnic\_Z +ethnic\_C +ethnic\_O +actvsubs + uniqsubs\_2 +uniqsubs\_3 +uniqsubs\_4

+ uniqsubs\_5 + uniqsubs\_6 +uniqsubs\_7 + uniqsubs\_9 + drop\_vce\_Mean +adjmou +totrev + adjrev+avgrev + retention + compl\_vce\_percentage

+ prizm\_social\_one\_R+prizm\_social\_one\_T+optimum,data = training,family = binomial(link = 'logit'))

summary(model2)

#Aic is Decreased 44827

#memory.limit(size = 40000)

#Lets Remove the insignificant variables from the above model i.e p>0.05 are drop\_blk\_Mean,drop\_vce\_Range,mou\_opkv\_Range,totcalls,ovrrev\_Mean

#comp\_vce\_Mean,plcd\_vce\_Mean,avg3qty,avgqty,area\_nrthwst,area\_nrthflorida,marital\_a,marital\_b,marital\_u,model\_8,actvsubs,uniqsubs\_9

#totrev,adjrev,avgrev,uniqsubs\_6,area\_calnrth,area\_southwest ,adjqty,avg6qty

#For variable crclscod

summary(training$crclscod)

training$crclscodE4<- ifelse(training$crclscod == 'E4',1,0)

testing$crclscodE4<-ifelse(testing$crclscod == 'E4',1,0)

training$crclscodEA<- ifelse(training$crclscod == 'EA',1,0)

testing$crclscodEA<-ifelse(testing$crclscod == 'EA',1,0)

model3 <- glm(churn ~ mou\_Mean + totmrc\_Mean + rev\_Range + mou\_Range + change\_mou + owylis\_vce\_Range

+ months + eqpdays + iwylis\_vce\_Mean + rev\_Mean

+ avg3mou + avgmou + avg6mou + crclscodE4+crclscodEA + asl\_flag\_y +area\_tenese

+ area\_southflorida +area\_texas + refurb\_new\_r + hnd\_price\_129.98 + hnd\_price\_199.98

+ hnd\_price\_249.98 + hnd\_price\_299.98 + marital\_m +marital\_s + model\_2 +model\_3 +model\_4

+ model\_5 + ethnic\_N + ethnic\_S + ethnic\_U + ethnic\_Z +ethnic\_C +ethnic\_O + uniqsubs\_2 +uniqsubs\_3 +uniqsubs\_4

+ uniqsubs\_5 + drop\_vce\_Mean +adjmou + retention + compl\_vce\_percentage

+ prizm\_social\_one\_R+prizm\_social\_one\_T+optimum,data = training,family = binomial(link = 'logit'))

summary(model3)

#AIC is Decreased <-44811

#Lets Re-Run the model removing some insignificant variable which is model\_5

model4 <- glm(churn ~ mou\_Mean + totmrc\_Mean + rev\_Range + mou\_Range + change\_mou + owylis\_vce\_Range

+ months + eqpdays + iwylis\_vce\_Mean + rev\_Mean

+ avg3mou + avgmou + avg6mou + crclscodE4+crclscodEA + asl\_flag\_y +area\_tenese

+ area\_southflorida +area\_texas + refurb\_new\_r + hnd\_price\_129.98 + hnd\_price\_199.98

+ hnd\_price\_249.98 + hnd\_price\_299.98 + marital\_m +marital\_s + model\_2 +model\_3 +model\_4

+ ethnic\_N + ethnic\_S + ethnic\_U + ethnic\_Z +ethnic\_C +ethnic\_O + uniqsubs\_2 +uniqsubs\_3 +uniqsubs\_4

+ uniqsubs\_5 + drop\_vce\_Mean +adjmou + retention + compl\_vce\_percentage

+ prizm\_social\_one\_R+prizm\_social\_one\_T+optimum,data = training,family = binomial(link = 'logit'))

summary(model4)

#AIC is 44812

#Lets remove some insignificants variables

model5 <- glm(churn ~ mou\_Mean + totmrc\_Mean + rev\_Range + mou\_Range + change\_mou + owylis\_vce\_Range

+ months + eqpdays + iwylis\_vce\_Mean + rev\_Mean

+ avg3mou + avgmou + avg6mou + crclscodE4+crclscodEA + asl\_flag\_y +area\_tenese

+ area\_southflorida +area\_texas + refurb\_new\_r + hnd\_price\_129.98 + hnd\_price\_199.98

+ hnd\_price\_249.98 + hnd\_price\_299.98 + marital\_m +marital\_s + model\_2 +model\_3 +model\_4

+ ethnic\_N + ethnic\_S + ethnic\_U + ethnic\_Z +ethnic\_C +ethnic\_O + uniqsubs\_2 +uniqsubs\_3 +uniqsubs\_4

+ uniqsubs\_5 + drop\_vce\_Mean +adjmou + retention + compl\_vce\_percentage

+ prizm\_social\_one\_R+prizm\_social\_one\_T+optimum,data = training,family = binomial(link = 'logit'))

summary(model5)

#No change in AIC i.e 44812 lets finalise this model

#VIF::vif(model5) #multicollinearity check for >5 values

plot(model5) #Plotting the model

######################################################################################3

#MODEL VALIDATION

#Model 5 is the best model we have got till now with aic 44812

sum(testing$churn) #<- 4270

sum(training$churn) #<-9992

sum(testing$churn)/nrow(testing) #23%

predicted\_values<-predict(model5,type = 'response',newdata = testing)

#type = 'response' link gives us probabilities it converts log of odd ratios values of dependent variables to probabilities of the event occurance

head(predicted\_values)

confint(model5)

# we need to choose a cut off value for predicted probabilities to define a churn

#There is no strict rule to decide on a cutoff value ,completely depends upon the study we are doing

#If correctly identifying positives is important for us then, we should choose a model with higher Sensitivity

#However ,if correctly identifying negatives is more important ,then we should choose Specificity

#RORC curve

library(ROCR)

length(predicted\_values)

pred<-predict(model5,type='response',newdata = testing)

head(pred)

#Setting the cutoff

#Assuming cutoff probabilities as per the churn rate in the dataset

table(telecom$churn)/nrow(telecom)

#Proportion is 0.2335391

pred1<-ifelse(pred>=0.2335391,1,0)

#getting zero will be churn

#Getting non-zero i.e 1 will not churn

library(lpSolve)

library(irr)

library(lattice)

library(caret)

predauc<-prediction(pred1,testing$churn)

predicted\_values\_cutoff<-ifelse(predicted\_values>0.2335391,1,0)

head(predicted\_values\_cutoff)

kappa2(data.frame(testing$churn,predicted\_values\_cutoff)) #Kappa is 0.134 Level of Agrement None

testing$probability<-predict(model5,type = 'response',newdata = testing)

testing$result<-ifelse(predicted\_values>0.2335391,1,0)

head(testing)

confusionMatrix(as.factor(testing$result),as.factor(testing$churn),positive = '1')

#57.47 % Accuracy

#AUC

auc<-performance(predauc,measure = 'auc')

auc

auc<-auc@y.values[[1]]

auc

#0.5918171 which is greater than 0.50

#According to solution of this project auc 0.5918171 is very good

#In this model validation ,model4 is turning out to be a good model . hence ,we can finalize it

#Let's built the gains chart

gain\_chart<-gains::gains(testing$churn,predict(model5,type = 'response',newdata = testing),groups = 10)

#The gains chart showsthat by targetting top 30% customers by probabilities ,on an average we will target 43.1 % of customers who will churn

class(gain\_chart)

#plot.gains(gain\_chart)

a=gain\_chart[1]

b=gain\_chart[6]

library(ggplot2)

gains=as.data.frame(c(a,b))

ggplot(data = gains)+geom\_line(aes(x=gains$depth,y=gains$cume.pct.of.total))

###################################################################################################################################3

#Question and answers

#1] What are the top five factors driving likelyhood of churn at mobicom ?

head(sort(abs(model5$coefficients),decreasing = TRUE),10)

model5$fitted.values

#Ans->#1.optimum(3.6377705) <- This variable is the ratio of overage revenue / total revenue #i.e overrev\_mean / totrev

#2.ethnic\_C (1.4554729)

#3.hnd\_price\_249.98(1.0732195)

#4.retention(0.7593517)<-Retention calls include any calls from the customer regarding loyalty or retention , e.g. contract renewal , relating competitor's offer ,etc.

#5.crclscodE4(0.6852267)

#2] Validation and Survey

#2A] Whether "Cost and Billing " and "Network and Serive Quality" are important Factors Influencing the Churn Behaviour ?

#ANs-> #Analysis Shows the cost and billing is causing churn as mou overage is Constructing to churn

#Network and service Quality is also found to be impacting churn as call drop is an issue as shown in analysis

#drop voice call and customer care call variable shows impact

summary(model5)

#2B] Are the data usage connectivity issue turning out to be costly ? In other words ,is it leading to churn ?

#ANs-> #Data Usage connectivity is not causing Significance impact on churn .

#Variables assiocated with data usage and connectivity are

#comp\_dat\_mean,plcd\_dat\_mean,opk\_dat\_Mean,blck\_dat\_Mean,datovr\_Mean,datovr\_range,drop\_dat\_Mean.

#3] Would you recommend rate plan migration as a proactive retention strategy ?

#ANS-> Yes, I would recommend rate plan migration as a proactive renteation strategy as there is direct relationship between

#Mean overrage revenue (overev\_Mean) and churn rate . Also, as mentioned above subscribers with the non-optimal rate

#tend to ovrrage and have significantly higher odds of churn as compared to subscriber with optimal rate.

#4] What would be your Recommendation on how to use this churn model for prioritization of customers for proactive retention campaigns in the future ?

#ANS-> #For Used Based Promotions

View(data\_adjmou)

#WE can Define Billing Adjusted total minutes of use over the life of the customer (adjmou) lessthan 5053 as low usage

#5053 is the average of 5th percentile lower and upper bounds

Customer\_low\_usage<- testing[testing$probability>0.2335391 & testing$adjmou>5053,]

nrow(testing)

nrow(Customer\_low\_usage)

head(Customer\_low\_usage)

Customer\_low\_usage<-as.data.frame(Customer\_low\_usage)

write.csv(Customer\_low\_usage,'Target Customer for Usage Campaigns.csv')

#We can also Target Customers with less than 5053 adjmou and probability churn more than 0.2335391 (Cutoff Value) for proactive usage increase plan

#Rate plan Migration : It is a strategy to move customers from non-optimal to optimal plans as it has been observed that subscribers on non-optimal rate plans

#Have Significantly higher odds of churn relative to subscriber on optimal rate

quantile(testing$optimum,c(p=1:10)/10)

#we can choose the customer for whom the 'Optimum and probability' is more than 0.2 (20%) ,which means their overage revenue contibutes more than 20% to the total Revenue

customer\_plan\_migration<-testing[testing$probability>0.2335391 & testing$adjmou<5053 & testing$optimum>0.022644329,]

nrow(customer\_plan\_migration)

customer\_plan\_migration<-as.data.frame(customer\_plan\_migration)

write.csv(customer\_plan\_migration,'Target Customer for Plan Migration.csv')

#Family Bunding Offer

#We can offer bundle offer to families to reduce churn

#For this the target Cutomer can those whose family a minimum 4 unique subscriber in the household and high churn rate probability

str(testing$actvsubs)

testing$actvsubs\_int<-as.numeric(testing$actvsubs)

customer\_FamilyBundel<- testing[testing$probability>0.2335391 & testing$actvsubs\_int>=4,]

nrow(customer\_FamilyBundel)

customer\_FamilyBundel<-as.data.frame(customer\_FamilyBundel)

write.csv(customer\_FamilyBundel,'Target Customer for Family Bundel.csv')

#Annual Retention Offers: one observation from the analysis is that ,on an average 43% who are with Mobicom between 10-12 months tends to leave .

#Mobicom can target customers who are having high churn Probability and are with the carrier for 10-12 months, to an annual offer to reduce the churn

customer\_annualRetention<-testing[testing$probability>0.2335391 & testing$months>=10 & testing$months<=12,]

nrow(customer\_annualRetention)

summary(customer\_annualRetention$months)

customer\_annualRetention<-as.data.frame(customer\_annualRetention)

write.csv(customer\_annualRetention,'Target Customer for Annual Retention Offer.csv')

#5] What will be the target segment for proactive customer Campaigns ? Falling ARPU forecast is also a concern and therefore Mobicom would you like to save their high revenue customers besides managing churn

#Giving a budget constraint of a contact list of 20% of the subscriber pool ,which Subscriber sholud prioritised if "Revenue Saves" is also a priority besides controlling Churn.

#In other words , Controlling Churn is the Primary Objective and Revenue saves it from secondary object .

#ANS-> Assuming that the customer in the test is our Subscriber pool

gain\_chart

#By looking at the gain chart , if we target top 20% of the subscriber pool (test file) by churn probability ,we will able to reach 3664 Customers

#Lets find out who are the customer

quantile(testing$probability,c(p=(1:10)/10))

#Top 20 Customer have probability between 0.2984393 and 0.7993235

#Lets Extract their data into new vector

customer\_proactive\_retention<-testing[testing$probability>0.2984393,]

nrow(customer\_proactive\_retention)

customer\_proactive\_retention<-as.data.frame(customer\_proactive\_retention)

write.csv(customer\_proactive\_retention,'Targeted Customer for Proactive Retention.csv')

#Lets see whom to target Based on Revenue as Priority

#The idea is to focus more on customers who give high revenue to Mobicom and retain them on priority

quantile(testing$probability,c(p=(1:100)/100))

quantile(customer\_proactive\_retention$probability,c(p=(1:10)/10))

customer\_proactive\_retention$Churn\_Level<-ifelse(customer\_proactive\_retention$probability<=0.3200020,'Low(0.30-0.3200020)',ifelse(customer\_proactive\_retention$probability>0.3200020 & customer\_proactive\_retention$probability<=0.3711780,'Medium(0.3200020-0.3711780)','High(0.3711780-0.8437166)'))

customer\_proactive\_retention$Churn\_Level<-as.factor(customer\_proactive\_retention$Churn\_Level)

summary(customer\_proactive\_retention$Churn\_Level)

quantile(customer\_proactive\_retention$totrev,c(p=(1:10)/10))

customer\_proactive\_retention$Revenue\_Level<-ifelse(customer\_proactive\_retention$totrev<=719.226,'Low(420-719.226)',ifelse(customer\_proactive\_retention$totrev>719.226 & customer\_proactive\_retention$totrev<=1337.972,'Medium(719.226-1337.972)','High(1337.972-27321.500)'))

customer\_proactive\_retention$Revenue\_Level<-as.factor(customer\_proactive\_retention$Revenue\_Level)

summary(customer\_proactive\_retention$Revenue\_Level)

table(customer\_proactive\_retention$Churn\_Level,customer\_proactive\_retention$Revenue\_Level)

customer\_proactive\_retention$Customer\_ID

library(dplyr)

#Lets Extract the Customer Id's of those who give high and medium revenue as well as have high and medium Probability of churn

customer\_proactive\_retention%>%filter(Revenue\_Level=='High(1337.972-27321.500)'& Churn\_Level=='High(0.3711780-0.8437166)')->Customer\_prio\_Target1

nrow(Customer\_prio\_Target1)

customer\_proactive\_retention%>%filter(Revenue\_Level=='High(1337.972-27321.500)' & Churn\_Level=='Medium(0.3200020-0.3711780)')->Customer\_prio\_Target2

nrow(Customer\_prio\_Target2)

customer\_proactive\_retention%>%filter(Revenue\_Level=='Medium(719.226-1337.972)' & Churn\_Level=='High(0.3711780-0.8437166)')->Customer\_prio\_Target3

nrow(Customer\_prio\_Target3)

customer\_prio\_Target<-rbind(Customer\_prio\_Target1,Customer\_prio\_Target2,Customer\_prio\_Target3)

nrow(customer\_prio\_Target)

names(customer\_prio\_Target)

names(customer\_proactive\_retention)

#Lets Extract Customer Id of Priority target segments

customer\_prio\_Target<-customer\_prio\_Target[,c(44,95,96)]

dim(customer\_prio\_Target)

write.csv(customer\_prio\_Target,'Target Customer for Priority ProActive Retention.csv')

getwd()