MOBICOM CHURN ANALYSIS

By
Nivedita Parab

Churn Prediction at Telecommunication Company

August 2017 – September 2017

- Analyzed telecommunication company's customers data where the company was facing an increase in the
  Churn rate and decreasing ARPU.
- Explored and cleaned data by the imputation of missing values and outliers using R packages and Decision tree.
- Data preparation for Logistic Regression Model.
- Built and validated logistic regression model and interpreted its output to determine important factors driving
  Churn in the company.
- Scored the customers on the basis of the probability to churn.
- Delivered insights to the company in order to help them to build retention strategy, targeting marketing
  budget, and increase in ARPU.

Problem Definition

Recent Industry Survey Report have reported increase in the churn and falling ARPU (Average revenue per unit). Mobicom is already suffering with the relatively high churn rate.

Current retention strategy are not proving much beneficial to Mobicom hence Management wants to roll-off new retention strategy / Churn prevention strategy that would revolve around

- Enhancing marketing programs to increase minutes of usage (MOU)
- Rate plan migration
- Bundling strategy

Top Line Questions of Interest to Senior Management:

- Validation of survey findings.
- What are the top five factors driving likelihood of churn at Mobicom?

  a) Whether "cost and billing" and "network and service quality" are important factors influencing churn behavior

  b) Are data usage connectivity issues turning out to be costly? In other words, is it leading to churn?
- Would you recommend rate plan migration as a proactive retention strategy?
- What would be your recommendation on how to use this churn model for prioritization of customers for a proactive retention campaigns in the future?
- What would be the target segments for proactive retention campaigns? Falling ARPU forecast is also a concern and therefore, Mobicom would like to save their high revenue customers besides managing churn. Given a budget constraint of a contact list of 20% of the subscriber pool, which subscribers should prioritized if "revenue saves" is also a priority besides controlling churn. In other words, controlling churn is the primary objective and revenue saves is the secondary objective.

Outcomes of Market survey

- Customers remain happy and loyal if quality services are provided .
- 40% of customers may switch service provider within 12 months
- 49% customers consider cost and billing as important factor
- 25 % customers consider network and service quality as important factor
- Many customers have problem with data usage
- 20% customers reports slow download speeds
- 17% customers reports throttling
- 16% customers report application don't work
- Recommendation from family and friends and internet are key decision factors

ARPU may continue to fall

Data Exploration

Data provided of 66927 customers consisting of details about the customer demographic data,usage data , billing data etc (81 variables)

50438 customers are still retained with the company and 15859 customers have churned

Current churn rate is 23.9%

Code

```
setwd("C:/Jig12309/capstone")
library(dplyr)
Telecom<-read.csv("telecomfinal.csv")
table(Telecom$churn)
churnrate<-15859/66297
```

Code used for accessing data quality

#preparing data quality report

names(Telecom) #to know variables present in dataset

#exploring variable

summary(Telecom$mou_Mean) # knowing values min max mean / levels

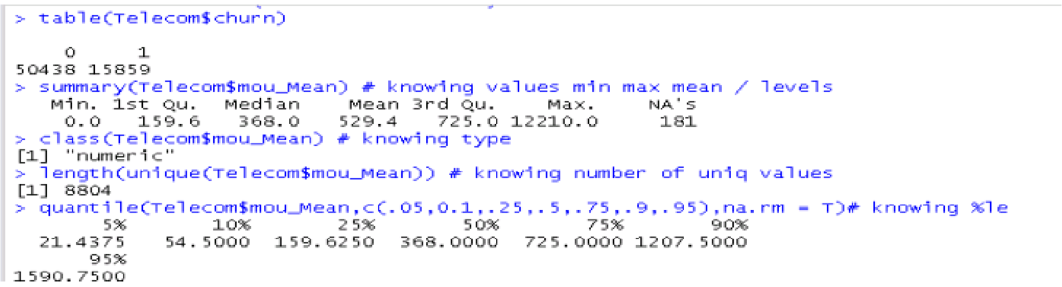
class(Telecom$mou_Mean) # knowing type

length(unique(Telecom$mou_Mean)) # knowing number of unique values

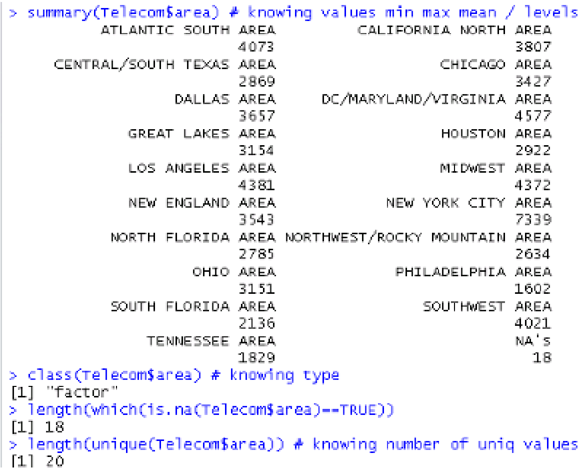
quantile(Telecom$mou_Mean,c(.05,0.1,.25,.5,.75,.9,.95),na.rm = T)# knowing %le

same code used for each variable in dataset to know mean,max,min / levels, NA, class, uniq values , percentiles

Quality report created is provided in C:\Jig12309\capstone\data_quality.xls



```
> table(Telecom$churn)

    0     1
50438 15859
> summary(Telecom$mou_Mean) # knowing values min max mean / levels
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0   159.6   368.0   529.4   725.0 12210.0     181
> class(Telecom$mou_Mean) # knowing type
[1] "numeric"
> length(unique(Telecom$mou_Mean)) # knowing number of uniq values
[1] 8804
> quantile(Telecom$mou_Mean,c(.05,0.1,.25,.5,.75,.9,.95),na.rm = T)# knowing %le
       5%       10%       25%       50%       75%       90%
  21.4375   54.5000  159.6250  368.0000  725.0000 1207.5000
      95%
1590.7500
```



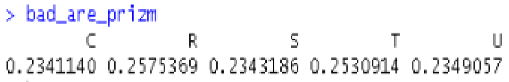
```
> summary(Telecom$area) # knowing values min max mean / levels
          ATLANTIC SOUTH AREA         CALIFORNIA NORTH AREA
                         4073                          3807
     CENTRAL/SOUTH TEXAS AREA                  CHICAGO AREA
                         2869                          3427
                  DALLAS AREA     DC/MARYLAND/VIRGINIA AREA
                         3657                          4577
             GREAT LAKES AREA                  HOUSTON AREA
                         3154                          2922
             LOS ANGELES AREA                  MIDWEST AREA
                         4381                          4372
             NEW ENGLAND AREA            NEW YORK CITY AREA
                         3543                          7339
           NORTH FLORIDA AREA NORTHWEST/ROCKY MOUNTAIN AREA
                         2785                          2634
                    OHIO AREA              PHILADELPHIA AREA
                         3151                          1602
           SOUTH FLORIDA AREA                SOUTHWEST AREA
                         2136                          4021
               TENNESSEE AREA                          NA's
                         1829                            18
> class(Telecom$area) # knowing type
[1] "factor"
> length(which(is.na(Telecom$area)==TRUE))
[1] 18
> length(unique(Telecom$area)) # knowing number of uniq values
[1] 20
```

Data preparation For Categorical variable

- We created churn-rate for each level of categorical variable .
- levels with similar churn rate are merged together to form dummy if needed.
- NA if any is replaced with the Level with nearest matching churn-rates.

Code used to determine churn rate per level of factor variable

```
Tel_crcls<-table(Telecom$crclscod,Telecom$churn)#creating 1 - 0 table for churn / level

bad_are_crcls<-Tel_crcls[,2]/rowSums(Tel_crcls) #calculating rate where churn=1

summary(bad_are_crcls)
```

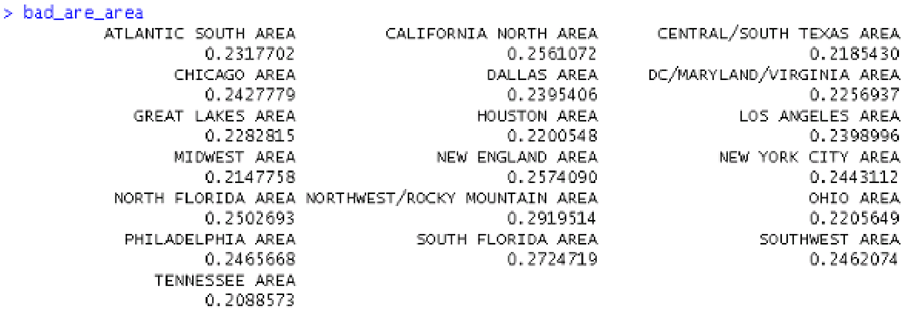


```
> bad_are_prizm
        C         R         S         T         U 
0.2341140 0.2575369 0.2343186 0.2530914 0.2349057 
```

Similar code used for all factor variables

Note:-some of integer variables with few uniq values are treated in similar fashion

This table shows badrate of each level of area

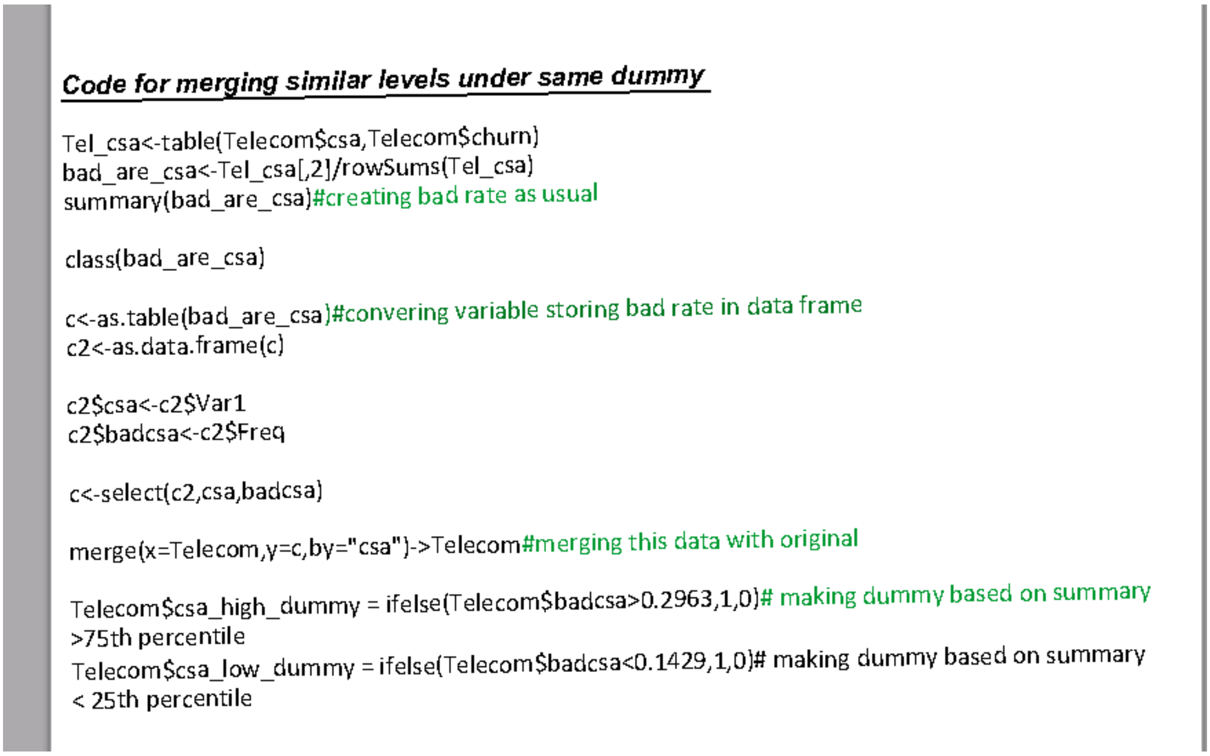


```
> bad_are_area
         ATLANTIC SOUTH AREA          CALIFORNIA NORTH AREA      CENTRAL/SOUTH TEXAS AREA 
                   0.2317702                      0.2561072                     0.2185430 
                CHICAGO AREA                    DALLAS AREA   DC/MARYLAND/VIRGINIA AREA 
                   0.2427779                      0.2395406                     0.2256937 
            GREAT LAKES AREA                   HOUSTON AREA             LOS ANGELES AREA 
                   0.2282815                      0.2200548                     0.2398996 
                MIDWEST AREA              NEW ENGLAND AREA           NEW YORK CITY AREA 
                   0.2147758                      0.2574090                     0.2443112 
          NORTH FLORIDA AREA NORTHWEST/ROCKY MOUNTAIN AREA                    OHIO AREA 
                   0.2502693                      0.2919514                     0.2205649 
           PHILADELPHIA AREA             SOUTH FLORIDA AREA              SOUTHWEST AREA 
                   0.2465668                      0.2724719                     0.2462074 
              TENNESSEE AREA 
                   0.2088573 
```

Code for replacing missing values

```
######replcaling missing value
ind<-which(is.na(Telecom$area))
table(Telecom$churn[ind])/length(ind)
Telecom$area[ind]<-"OHIO AREA"
```

```
> table(Telecom$churn[ind])/length(ind)

        0         1 
0.7777778 0.2222222 
```



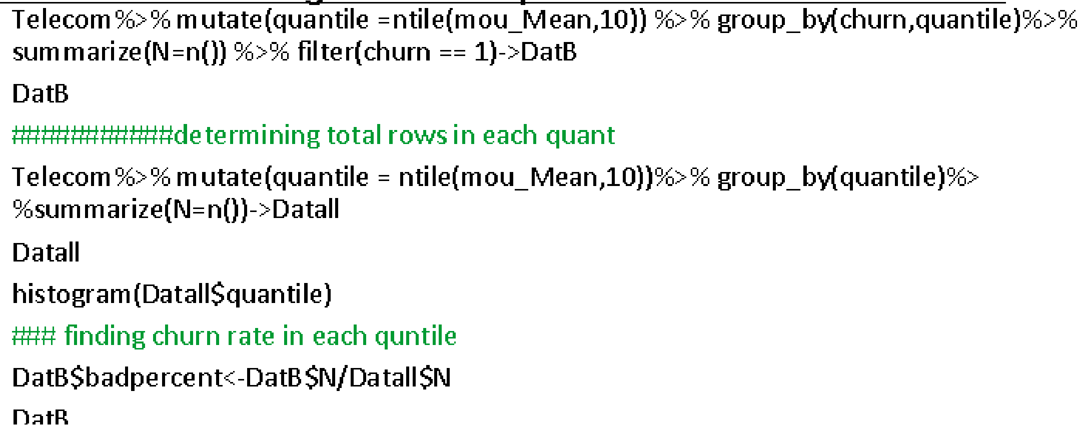
Code for merging similar levels under same dummy

```
Tel_csa<-table(Telecom$csa,Telecom$churn)
bad_are_csa<-Tel_csa[,2]/rowSums(Tel_csa)
summary(bad_are_csa)#creating bad rate as usual

class(bad_are_csa)

c<-as.table(bad_are_csa)#convering variable storing bad rate in data frame
c2<-as.data.frame(c)

c2$csa<-c2$Var1
c2$badcsa<-c2$Freq

c<-select(c2,csa,badcsa)

merge(x=Telecom,y=c,by="csa")->Telecom#merging this data with original

Telecom$csa_high_dummy = ifelse(Telecom$badcsa>0.2963,1,0)# making dummy based on summary
>75th percentile
Telecom$csa_low_dummy = ifelse(Telecom$badcsa<0.1429,1,0)# making dummy based on summary
< 25th percentile
```

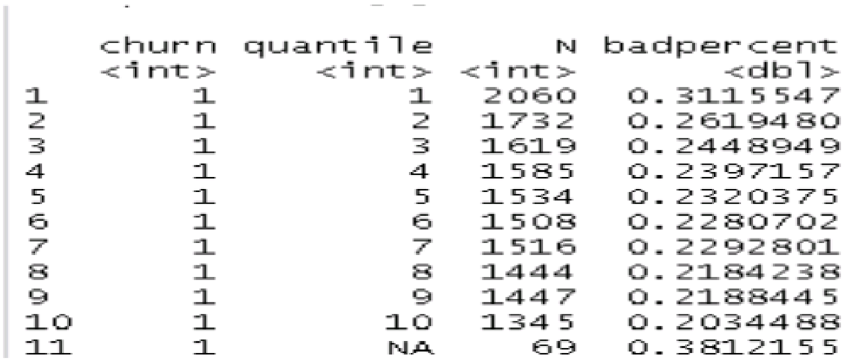
Data preparation for Continues Variables

- Each continues variable is divided in to decile.
- Churn rate across that deciles is determined
- Na if any are replaced with the deciles having nearest churn rates
- Any outliers are removed / replaced with the help of outlier library , box plots.

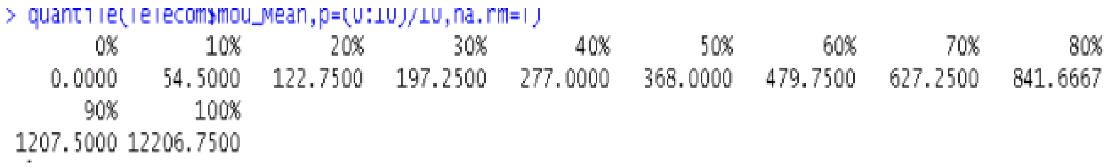
Code for determining Churn Rate per decile of continues variable



```
Telecom%>% mutate(quantile =ntile(mou_Mean,10)) %>% group_by(churn,quantile)%>%
summarize(N=n()) %>% filter(churn == 1)->DatB

DatB

############determining total rows in each quant

Telecom%>% mutate(quantile = ntile(mou_Mean,10))%>% group_by(quantile)%>%
%summarize(N=n())->Datall

Datall

histogram(Datall$quantile)

### finding churn rate in each quntile

DatB$badpercent<-DatB$N/Datall$N

DatB
```



```
   churn quantile     N badpercent
   <int>    <int> <int>      <dbl>
1      1        1  2060  0.3115547
2      1        2  1732  0.2619480
3      1        3  1619  0.2448949
4      1        4  1585  0.2397157
5      1        5  1534  0.2320375
6      1        6  1508  0.2280702
7      1        7  1516  0.2292801
8      1        8  1444  0.2184238
9      1        9  1447  0.2188445
10     1       10  1345  0.2034488
11     1       NA    69  0.3812155
```



```
> quantile(Telecom$mou_Mean,p=(0:10)/10,na.rm=T)
       0%       10%       20%       30%       40%       50%       60%       70%       80% 
   0.0000   54.5000  122.7500  197.2500  277.0000  368.0000  479.7500  627.2500  841.6667 
      90%      100% 
1207.5000 12206.7500 
```

quantile(Telecom$mou_Mean,p=(0:10)/10,na.rm=T)

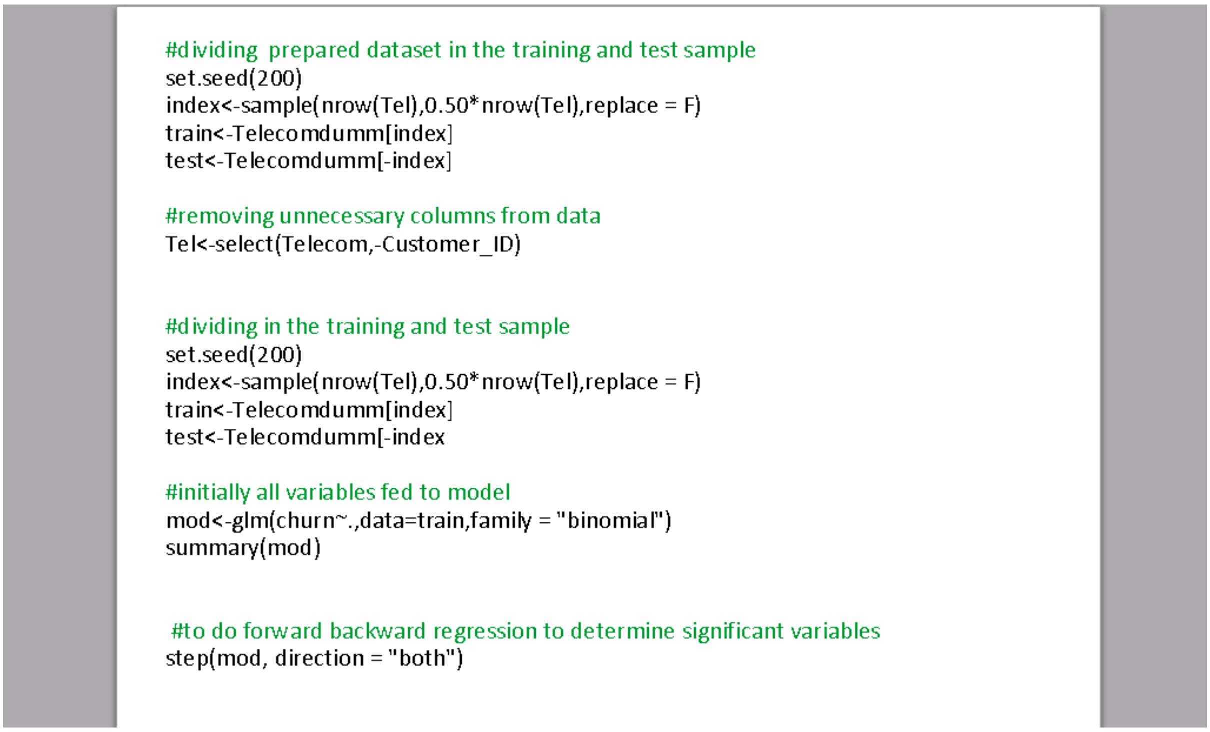
- All continues variables are profiled with the similar code.
- NA are replaced with the mean / nearest churn rate matching decile.
- Outliers are treated by help of outliers library and boxplot.
- Whenever needed deciles having similar churn rates are combined in one dummy

Creating Model

- Logistic Regression Model is created to determining probability of customer to churn.
- Our target variable is Churn which is binary variable taking value 1 if customer left and 0 if did not.
- Probability modeled was for churn = 1

Code

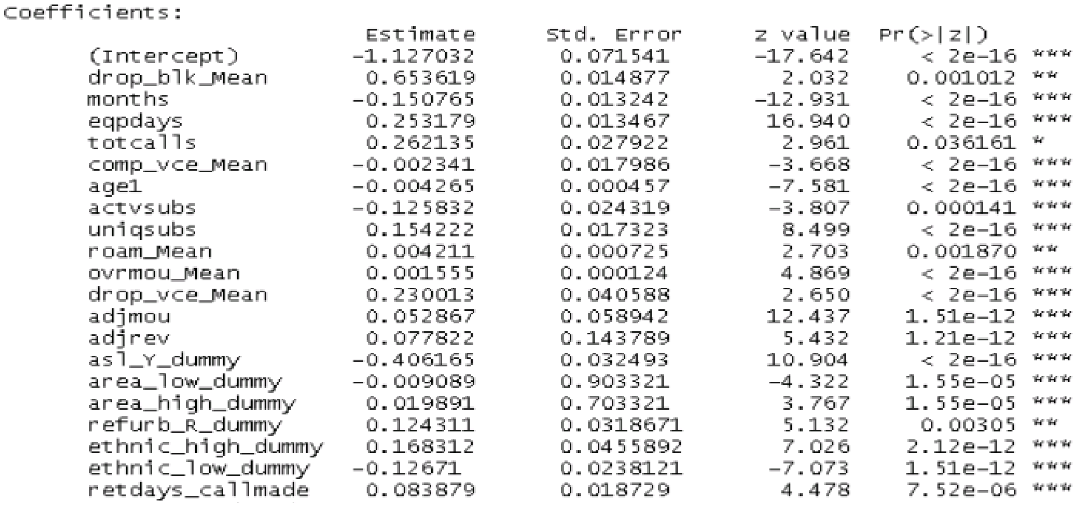
```
#importing necessary  packages
library(gains)
library(dplyr)
library(irr)
library(caret)
```



```
#dividing  prepared dataset in the training and test sample
set.seed(200)
index<-sample(nrow(Tel),0.50*nrow(Tel),replace = F)
train<-Telecomdumm[index]
test<-Telecomdumm[-index]

#removing unnecessary columns from data
Tel<-select(Telecom,-Customer_ID)


#dividing in the training and test sample
set.seed(200)
index<-sample(nrow(Tel),0.50*nrow(Tel),replace = F)
train<-Telecomdumm[index]
test<-Telecomdumm[-index

#initially all variables fed to model
mod<-glm(churn~.,data=train,family = "binomial")
summary(mod)


 #to do forward backward regression to determine significant variables
step(mod, direction = "both")
```

After several iteration we finalized this model

```
#final model having all significant variable and correct signs
mod2<-glm(formula=churn ~  drop_blk_Mean +  months + totcalls + eqpdays + comp_vce_Mean
    + age1 + + actvsubs + uniqsubs +   roam_Mean + ovrmou_Mean +
    drop_vce_Mean+adjmou+adjrev+ asl_Y_dummy + area_low_dummy + area_high_dummy +
    refurb_R_dummy   + ethnic_high_dummy + ethnic_low_dummy + retdays_callmade+
    avg6mou_reduced,data=Test,family="binomial")
summary(mod)
```

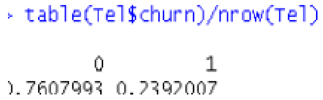


```
Coefficients:
                      Estimate      Std. Error      z value   Pr(>|z|)
     (Intercept)      -1.127032      0.071541      -17.642     < 2e-16 ***
     drop_blk_Mean     0.653619      0.014877        2.032    0.001012 **
     months           -0.150765      0.013242      -12.931     < 2e-16 ***
     eqpdays           0.253179      0.013467       16.940     < 2e-16 ***
     totcalls          0.262135      0.027922        2.961    0.036161 *
     comp_vce_Mean    -0.002341      0.017986       -3.668     < 2e-16 ***
     age1             -0.004265      0.000457       -7.581     < 2e-16 ***
     actvsubs         -0.125832      0.024319       -3.807    0.000141 ***
     uniqsubs          0.154222      0.017323        8.499     < 2e-16 ***
     roam_Mean         0.004211      0.000725        2.703    0.001870 **
     ovrmou_Mean       0.001555      0.000124        4.869     < 2e-16 ***
     drop_vce_Mean     0.230013      0.040588        2.650     < 2e-16 ***
     adjmou            0.052867      0.058942       12.437    1.51e-12 ***
     adjrev            0.077822      0.143789        5.432    1.21e-12 ***
     asl_Y_dummy      -0.406165      0.032493       10.904     < 2e-16 ***
     area_low_dummy   -0.009089      0.903321       -4.322    1.55e-05 ***
     area_high_dummy   0.019891      0.703321        3.767    1.55e-05 ***
     refurb_R_dummy    0.124311      0.0318671       5.132     0.00305 **
     ethnic_high_dummy 0.168312      0.0455892       7.026    2.12e-12 ***
     ethnic_low_dummy -0.12671       0.0238121      -7.073    1.51e-12 ***
     retdays_callmade  0.083879      0.018729        4.478    7.52e-06 ***
```

- After checking the same model on both testing and validation sample we are sure all coefficients and sign are proper.

- Creating prediction table

```
#creating prediction table
pred<-predict(mod,type="response",newdata=Tel)
summary(pred)

#if probablity is greater than churnrate pred =1
table(Tel$churn)/nrow(Tel)
```



```
> table(Tel$churn)/nrow(Tel)

        0         1
).7607993 0.2392007
```

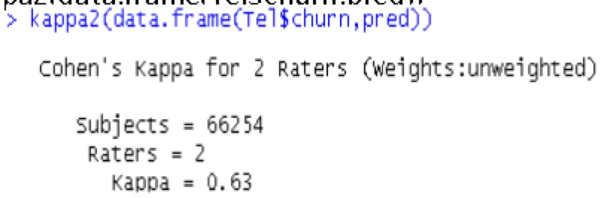
- pred<-ifelse(pred>=0.2392007,1,0)

Validation

- **Creating kappa matrix**

code

kappa2(data.frame(Tel$churn,pred))

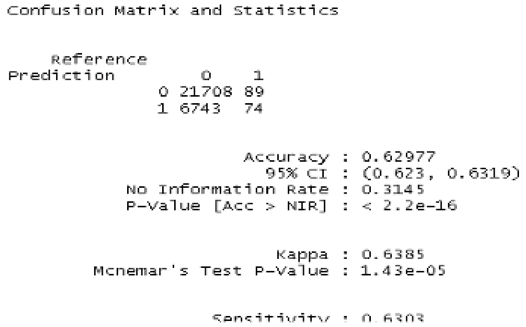


```
> kappa2(data.frame(Tel$churn,pred))

   Cohen's Kappa for 2 Raters (Weights:unweighted)

      Subjects = 66254
        Raters = 2
         Kappa = 0.63
```

as value of kappa is greater than 60 model is good

Forming confusion matrix

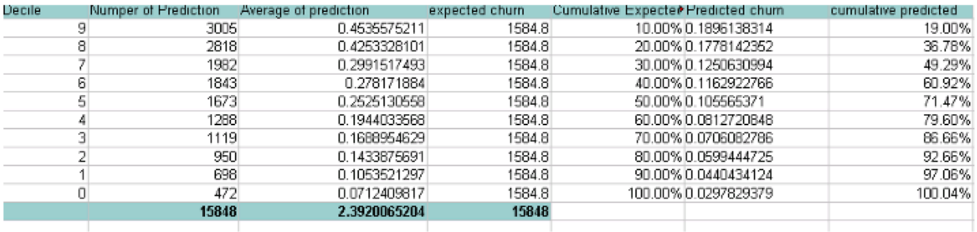
Code
confusionMatrix(pred,Tel$churn,positive ="1")



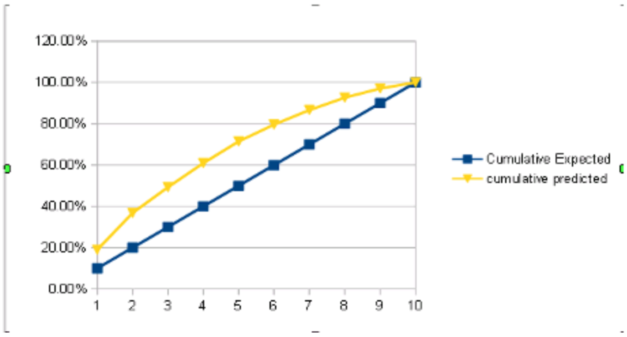
```
Confusion Matrix and Statistics


    Reference
Prediction         0    1
             0 21708 89
             1 6743   74


                  Accuracy : 0.62977
                    95% CI : (0.623, 0.6319)
       No Information Rate : 0.3145
       P-Value [Acc > NIR] : < 2.2e-16


                     Kappa : 0.6385
    Mcnemar's Test P-Value : 1.43e-05


               Sensitivity : 0.6303
```

Model corerectly predicts churn when churn 6743 times and non event as non event 21708 times type one and type 2 error is less hence model is good

Gain chart



Decile	Numper of Prediction	Average of prediction	expected churn	Cumulative Expecte	Predicted churn	cumulative predicted
9	3005	0.4535575211	1584.8	10.00%	0.1896138314	19.00%
8	2818	0.4253328101	1584.8	20.00%	0.1778142352	36.78%
7	1982	0.2991517493	1584.8	30.00%	0.1250630994	49.29%
6	1843	0.278171884	1584.8	40.00%	0.1162922766	60.92%
5	1673	0.2525130558	1584.8	50.00%	0.105565371	71.47%
4	1288	0.1944033568	1584.8	60.00%	0.0812720848	79.60%
3	1119	0.1688954629	1584.8	70.00%	0.0706082786	86.66%
2	950	0.1433875691	1584.8	80.00%	0.0599444725	92.66%
1	698	0.1053521297	1584.8	90.00%	0.0440434124	97.06%
0	472	0.0712409817	1584.8	100.00%	0.0297829379	100.04%
	15848	**2.3920065204**	**15848**			



Q1 .What are the top five factors driving likelihood of churn at Mobicom?

- **Account Spending limit activated (asl_y_dummy):** Is observed that there are greater probability of churn if account spending limit is not activated. With account spending limit not activated observed churn rate is 25% and with activated is 17.5%

- **Average monthly minutes of use over past 6 months(avg6mou_reduced)**: It is observed that the customer having reduced average monthly mou over last 6 months have greater chances of churn.

- **Mean Number of dropped voice call (drop_vce_Mean) :** Is observerd that with increase in the mean number of dropped voice calls probablity of customer leaving is increasing

- **Number of active subscribers in household (actvsub):** Its seems like recommendation of friends and family matters churn is observed to be reduced with increase in active subscribers in home.

- **Number of days (age) of current equipments(eqpdays):**positive correlation observed between the churn and number of days. Hence older the equipment greater the probability of churn

Q2 .Validation of survey findings. a) Whether “cost and billing” and “network and service quality” are important factors influencing churn behavior b) Are data usage connectivity issues turning out to be costly? In other words, is it leading to churn?

- **Cost and Billing** :-

Billing adjusted total minutes of usage over life time of customer(adjmou) and Billing adjusted total revenue over life time of customers (adjrev) are found to be significant parameters in the model hence cost and billing is a important factor influencing churn.

- **Data usage connectivity:-**

Mean number of dropped (failed) data calls (drop_blk_Mean) as increases chance of churn also increases

Mean number of completed voice calls (comp_vce_Mean) increases chances of churn are reduced.

Mean number of roaming calls (roam_Mean) as increases chance of churn increases

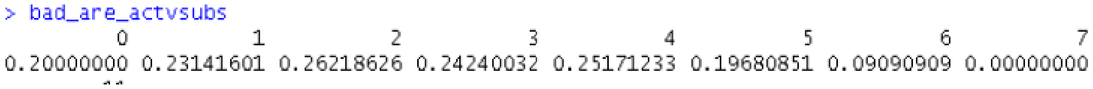
Are found to be important factors deciding churn rate. Hence connectivity issue is leading churn.

Q3. Would you recommend rate plan migration as a proactive retention strategy?

Mean overage mou (ovrmou_mean) is found to be significant as overage minute of use increases chances of churn also increases. Hence rate plan migration might be proactive retention strategy. Also some variable regarding billing and cost found to be significant hence customers can be shifted to more optimal bill plans.

Q4. What would be your recommendation on how to use this churn model for prioritization of customers for a proactive retention campaigns in the future?

- We observed that in house holds having active subscribers more than 4 have very low churn rate also observed that in houses having more uniqe subscribers but less active subscriber probability of churn is high hence we should try to implement family oriented plans to attract more subscribers from same family.



```
> bad_are_actvsubs
         0          1          2          3          4          5          6          7 
0.20000000 0.23141601 0.26218626 0.24240032 0.25171233 0.19680851 0.09090909 0.00000000 
```

- With increase in age of equipment churn rate increases. Hence mobicom can target new equipment buying customers for marketing data usage and voice call plans.

- Its observed that churn is higher in California North,New England,North Florida,Northwest/Rocky Mountain,South Florida area hence we should pay attention to this area (suggest further research to search problem area)

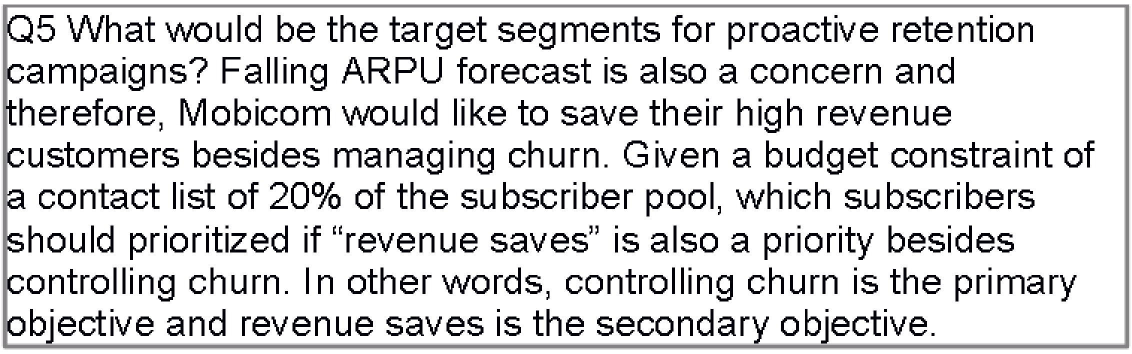
- Churn is higher with ethnicity code B D J O hence need to pay attention to this ethnic groups.

- Observe customers having significant reduce in past 3 months minutes of use as such customers tend to churn more.

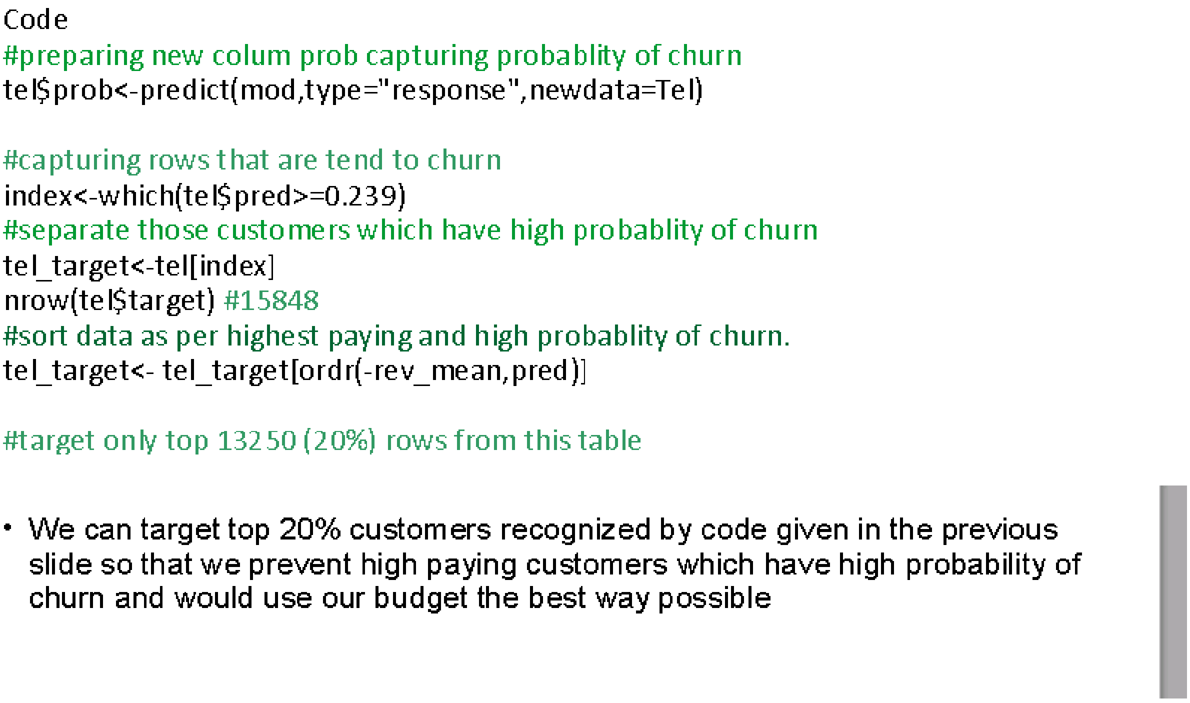
- Customers having greater adjmou can be moved to more optimal rate plans.

- Try to optimize coverage area and to reduce in roaming calls.

- Optimize network handovers to reduce drop voice calls.



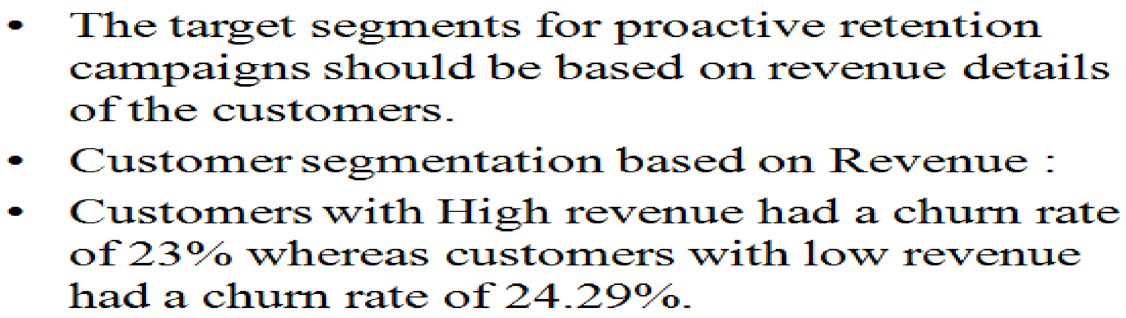
Q5 What would be the target segments for proactive retention campaigns? Falling ARPU forecast is also a concern and therefore, Mobicom would like to save their high revenue customers besides managing churn. Given a budget constraint of a contact list of 20% of the subscriber pool, which subscribers should prioritized if “revenue saves” is also a priority besides controlling churn. In other words, controlling churn is the primary objective and revenue saves is the secondary objective.



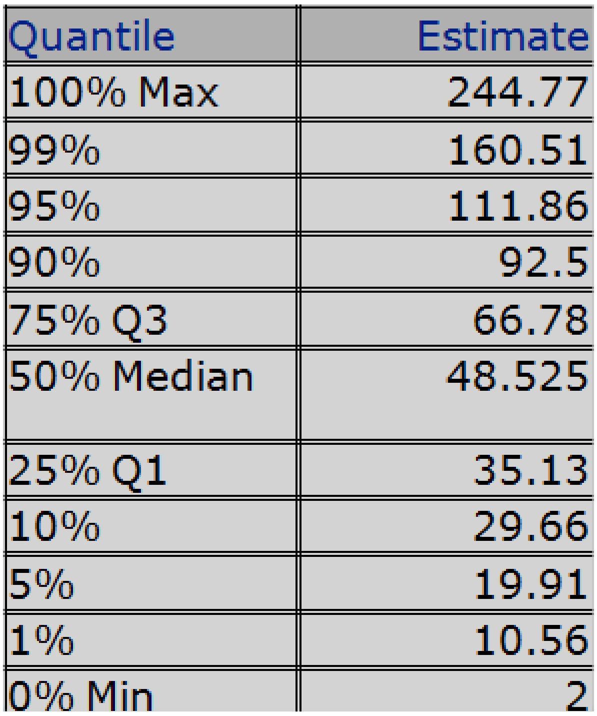
Code

```
#preparing new colum prob capturing probablity of churn
tel$prob<-predict(mod,type="response",newdata=Tel)

#capturing rows that are tend to churn
index<-which(tel$pred>=0.239)
#separate those customers which have high probablity of churn
tel_target<-tel[index]
nrow(tel$target) #15848
#sort data as per highest paying and high probablity of churn.
tel_target<- tel_target[ordr(-rev_mean,pred)]

#target only top 13250 (20%) rows from this table
```

- We can target top 20% customers recognized by code given in the previous slide so that we prevent high paying customers which have high probability of churn and would use our budget the best way possible

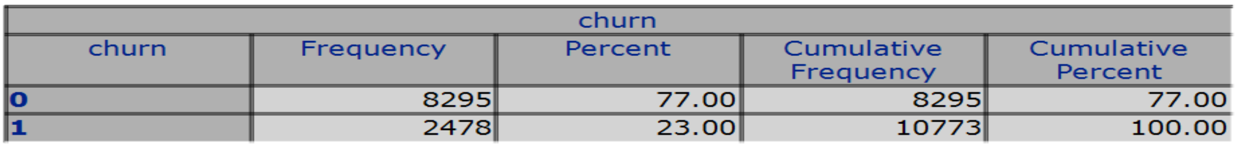


- The target segments for proactive retention campaigns should be based on revenue details of the customers.
- Customer segmentation based on Revenue :
- Customers with High revenue had a churn rate of 23% whereas customers with low revenue had a churn rate of 24.29%.



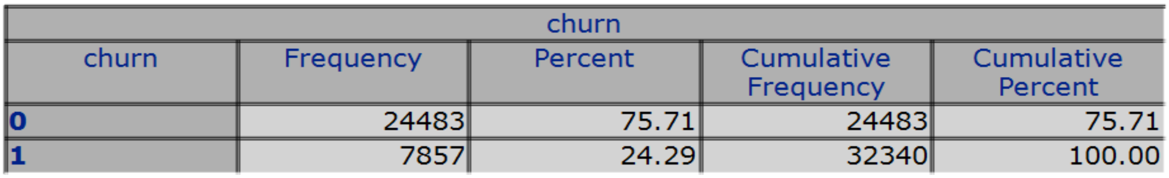
Quantile	Estimate
100% Max	244.77
99%	160.51
95%	111.86
90%	92.5
75% Q3	66.78
50% Median	48.525
25% Q1	35.13
10%	29.66
5%	19.91
1%	10.56
0% Min	2

High Revenue Customers



churn				
churn	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8295	77.00	8295	77.00
1	2478	23.00	10773	100.00

Low Revenue Customers

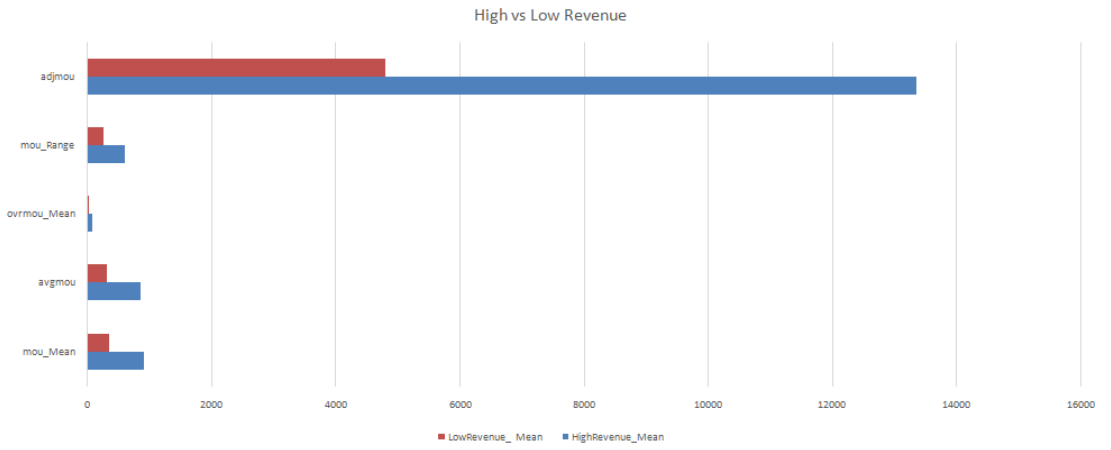


churn				
churn	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	24483	75.71	24483	75.71
1	7857	24.29	32340	100.00

Proactive Retention Strategies

- Compared low revenue customers High revenue customers will have:

1. 3 times more **Average number of Monthly minutes of usage**
2. 3 times more **Average monthly minutes of use over the life of the customer**
3. 4 times more **Mean overage minutes of use**
4. 3 times more **Billing adjusted total minutes of use over the life of the customer**
5. 2 times more **Range of number of minutes of use.**



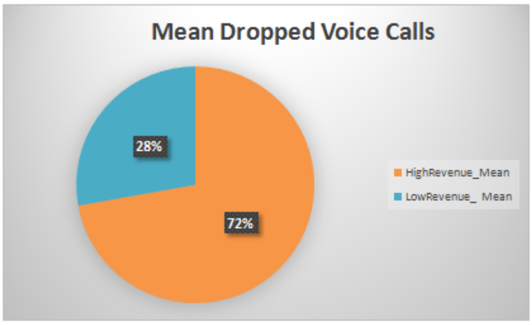
Strategy 1

- So to reduce the churn rates and increase ARPU,Retention and Marketing team should target these High Revenue customers with offers which includes :

1. Lower overage charges.
2. Quick Billing Adjustments with acknowledgement messages so that customers are aware of the proactive actions taken by Mobicom.
3. Value added services for higher minutes of usage i.e. more than 800 or 850

Strategy 2

The average Dropped voice calls in High Revenue customers are also 3 times more than that of low revenue customers.



Strategy 2

- More Dropped voice calls was also significant predictors for the churn rate in our Model.
- Hence higher the dropped voice calls higher are the chances of churn rates.
- Moreover, for every dropped voice calls Mobicom have to reimburse its customer by billing adjustments, thereby affecting ARPU.
- Hence Mobicom's Retention and Marketing team should classify such High Revenue customers who are facing higher average dropped voice calls and take corrective actions to solve this technical issue. Also, from our Analysis, we have found that around **7%** of these customers are of Afro-American Ethinicity.

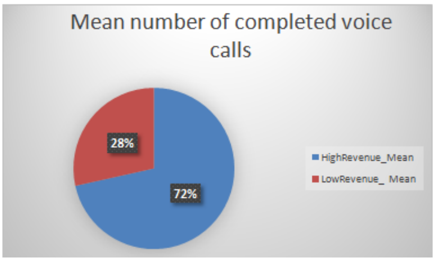
Strategy 3

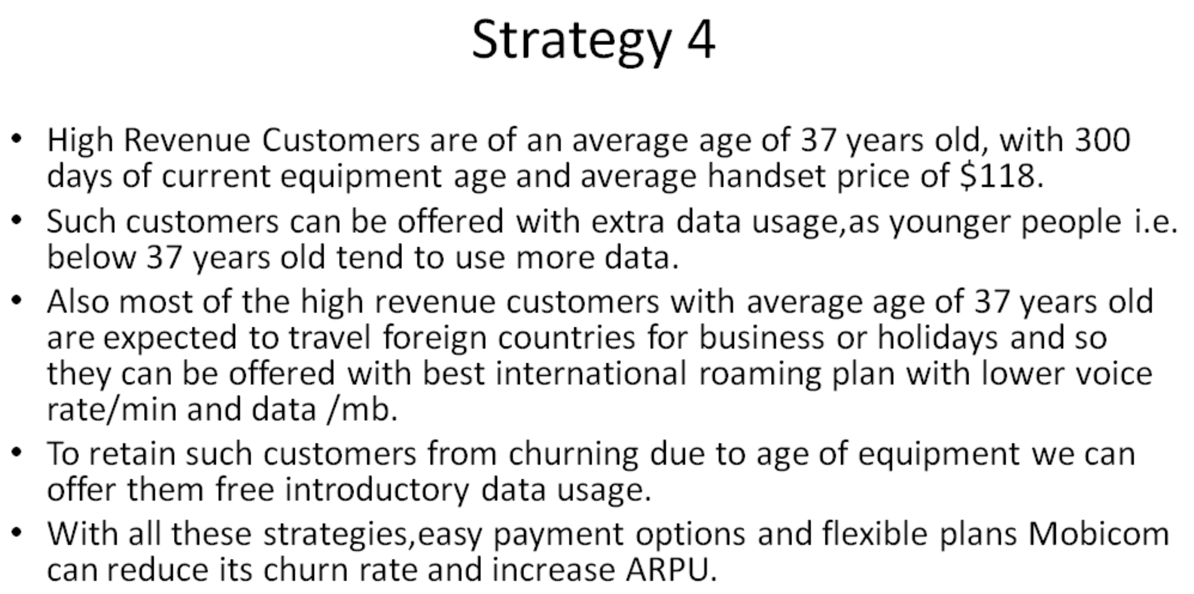
- High revenue customers has 2 times more Mean number of completed voice calls as compared to Low revenue customers.
- Hence such customers should be provided with

Offers related to price of voice with some other VAS.

- This will not only retain customers but also help in

Generating revenue





Strategy 4

- High Revenue Customers are of an average age of 37 years old, with 300 days of current equipment age and average handset price of $118.
- Such customers can be offered with extra data usage,as younger people i.e. below 37 years old tend to use more data.
- Also most of the high revenue customers with average age of 37 years old are expected to travel foreign countries for business or holidays and so they can be offered with best international roaming plan with lower voice rate/min and data /mb.
- To retain such customers from churning due to age of equipment we can offer them free introductory data usage.
- With all these strategies,easy payment options and flexible plans Mobicom can reduce its churn rate and increase ARPU.

The project was built on the jigsaw academy lMS cloud and not on personal desktop hence could not copy pest code here

However important code is provided in the document this is screenshots of actual presentation built in a cloud