Churn Prediction in Telecom Sector

Agenda

1) Project Synopsis
2) Regression Modelling Approach
3) Data Exploration and Preparation
4) Decision Tree Snapshot
5) Top Five Driving Factors for Churn
6) Logistic Model Performance
7) Data Exploration(Variance)
8) Data Exploration(Covariance)
9) Effect of Network and Service quality on churn
10) Effect of Data Usage on churn
11) Rate Plan Migration Strategy
12) Usage Based (MOU) Strategy
13) Recommendations

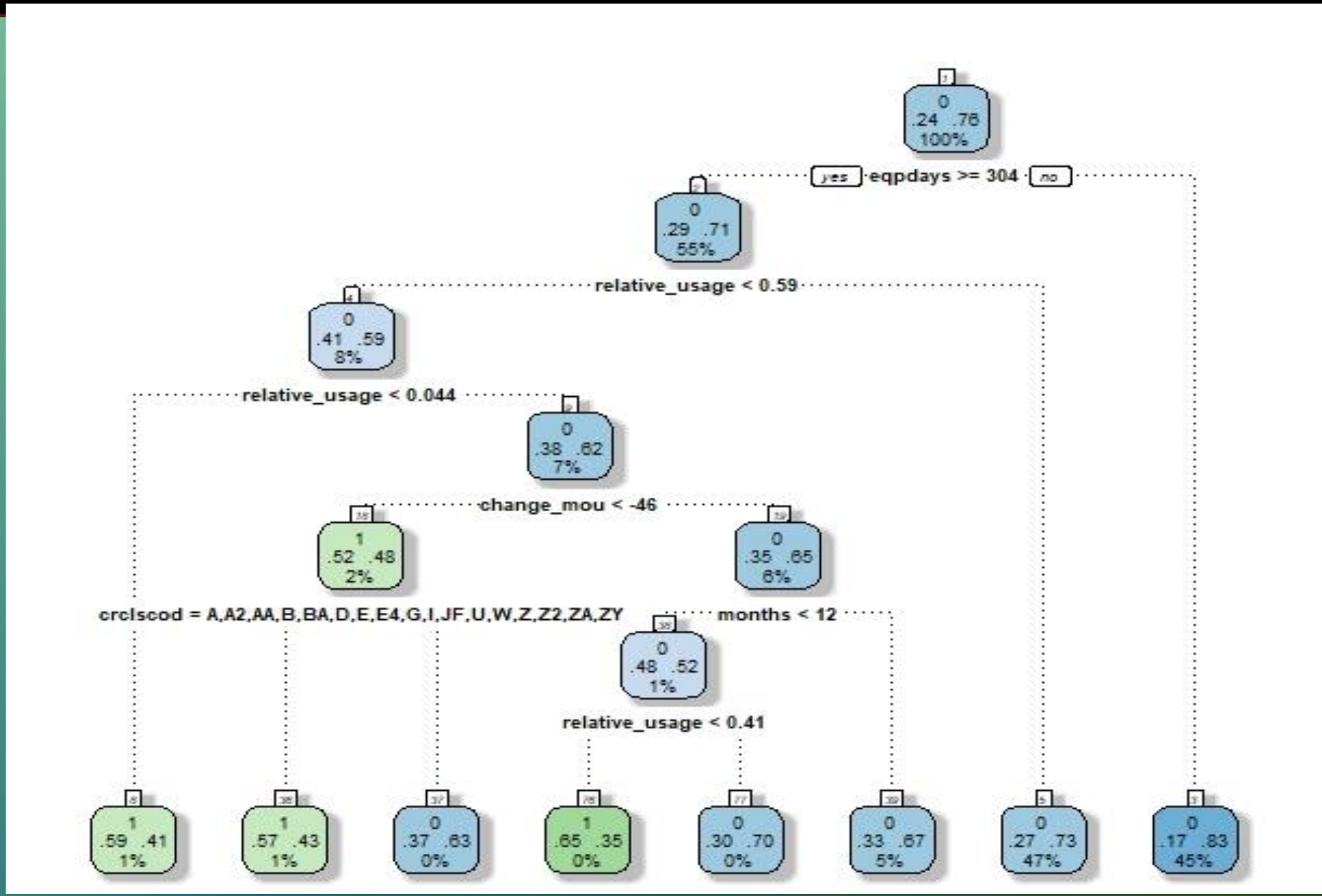
Project Synopsis

- Project Background: In telecommunications industry, customers are able to easily switch between different service providers and it costs company 5-10 times more to acquire a new customer than to retain the existing customers. Hence controlling churn has gained a lot of importance in the industry.
- Objective: The goal is to determine the ability of different factors such as minutes of usage, revenue from the customer, duration of customer etc. to drive the behaviour of churn which can be used by the company in their proactive retention programmes.
- Granularity: This study examines customer churn at the individual account level.
- Scope: To build a tree model and a logistic regression model for predicting churn as a dependent variable with independent variables like average minutes of usage, average revenue etc. to determine which factors are significant in prediction of the

Regression Modelling Approach

- The modelling approach is a combination of the Decision Tree and the Logistic Regression approach.
- First, Decision tree model was built and the tree was allowed to grow to the size of more than 1300 terminal nodes and then using the cost complexity parameter , it was pruned to the optimum size of 8 terminal nodes.
- Another approach used was Logistic regression. 'Backward selection' was used along with 'vif' factor to choose the best model parameters.
- The focus of the Logistic as well as Tree model was on 'Type I error' (or sensitivity) as the objective of this study is to help the retention programmes of the service provider by predicting churn.

Decision Tree Snapshot



Data Exploration and Preparation

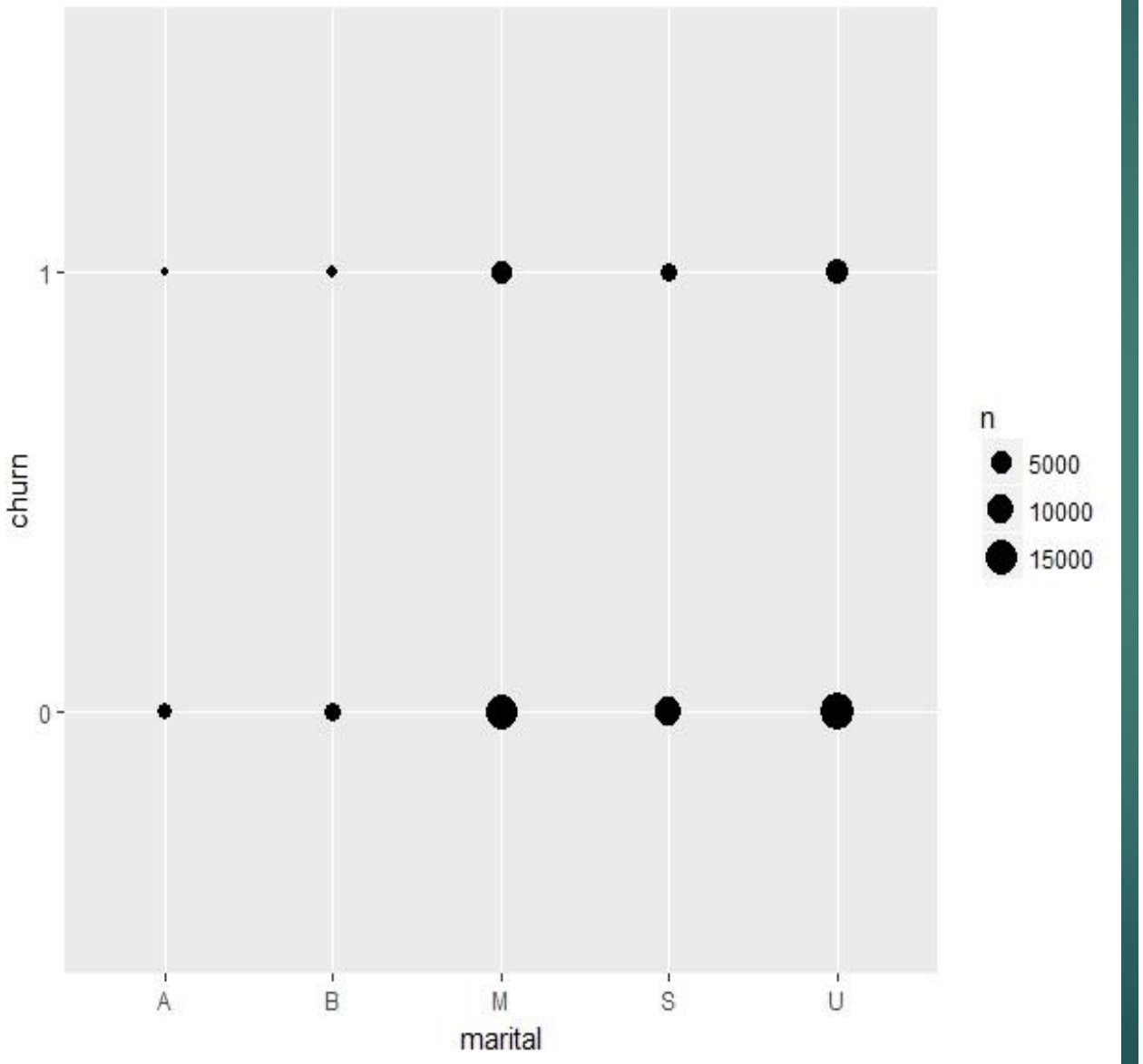
Total Number of Customers in the data	**66,297**
❑ Exclusions: Variables excluded due to missing data 1) Income 2) mailordr 3) dwlltype 4) dwllsize 5) occu1 6) mtrcycle 7) numbcars 8) retdays 9) Wrkwoman 10) solflag 11) proptype 12) mailresp 13) cartype 14) car_buy 15) csa and some more… ❑ Missing value imputation: -wrote impute_missing( ) function ❑ Outliers treatment: -wrote delete_outliers( ) function outliers=(3rd quantile+2*IQR) & (1st quantile-2*IQR)	
Derived Variables: 1) non_optimal_plan=(ovrrev_mean/avgrev) 2) Relative_usage=(mou_mean/avgmou) 3) Recency=(avg3mou/avg6mou) 4) Billing_rev_accuracy=(adjrev/totrev) 5) Dummy variables for: Ethnic, Area, Asl_flag, marital, crclscod, refurb etc.	
Total number of customers after data exploration and preparation	**65,882**

Top Five Driving Factors for Churn

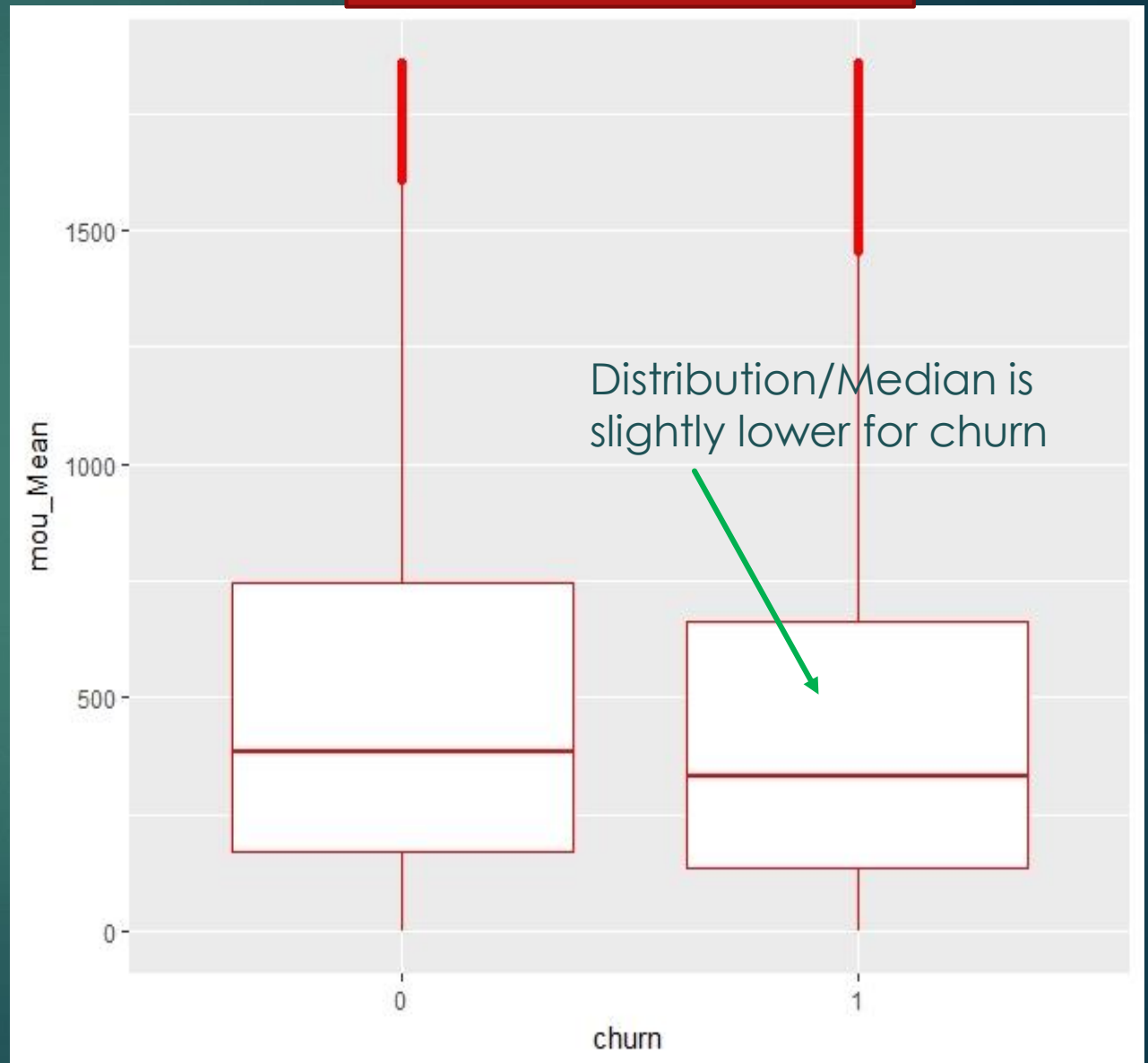
Method for ranking predictors: Reduction of Variance (Gini)

Data Exploration(Variance)

Marital Status vs Churn

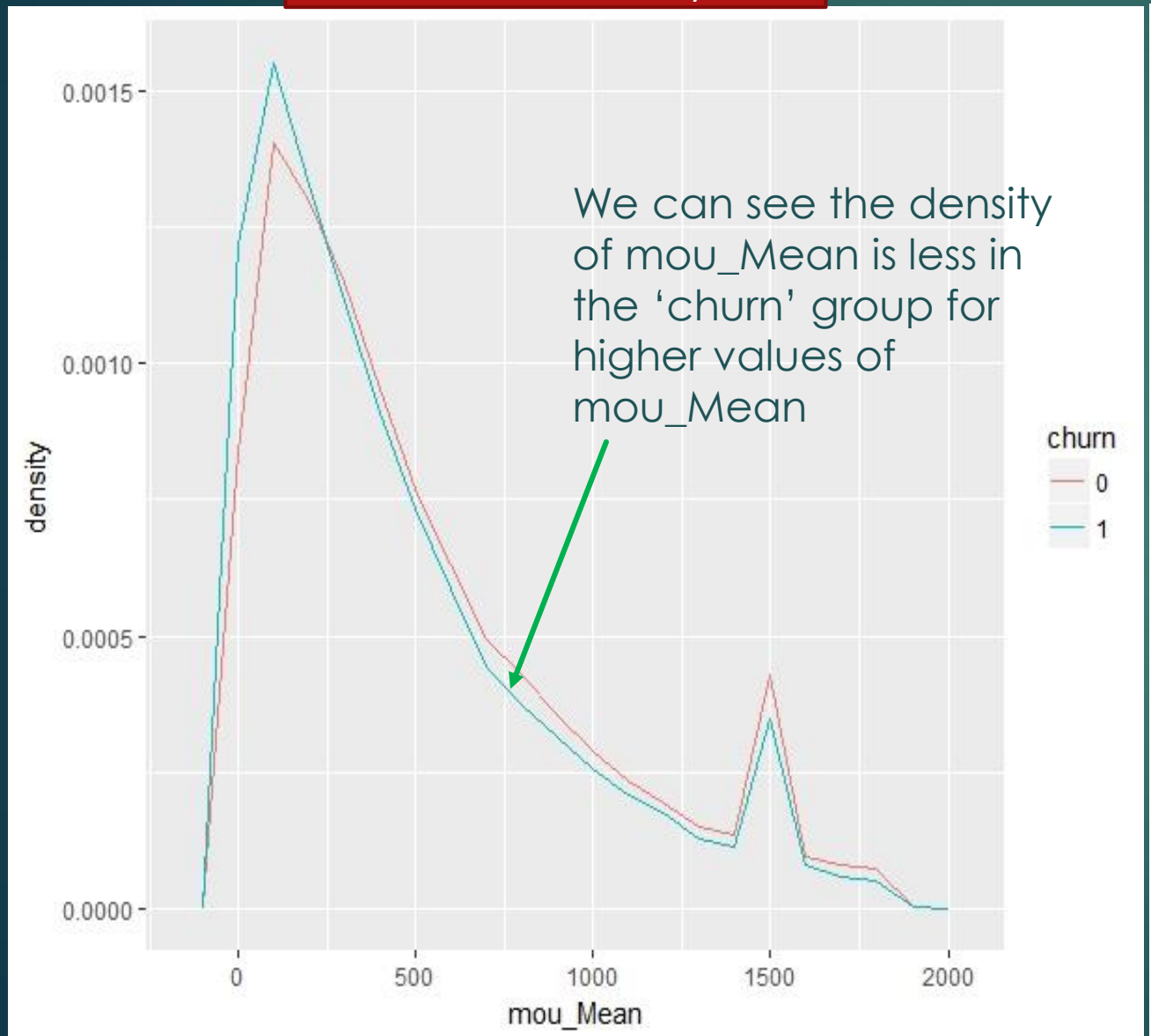


mou_Mean vs Churn

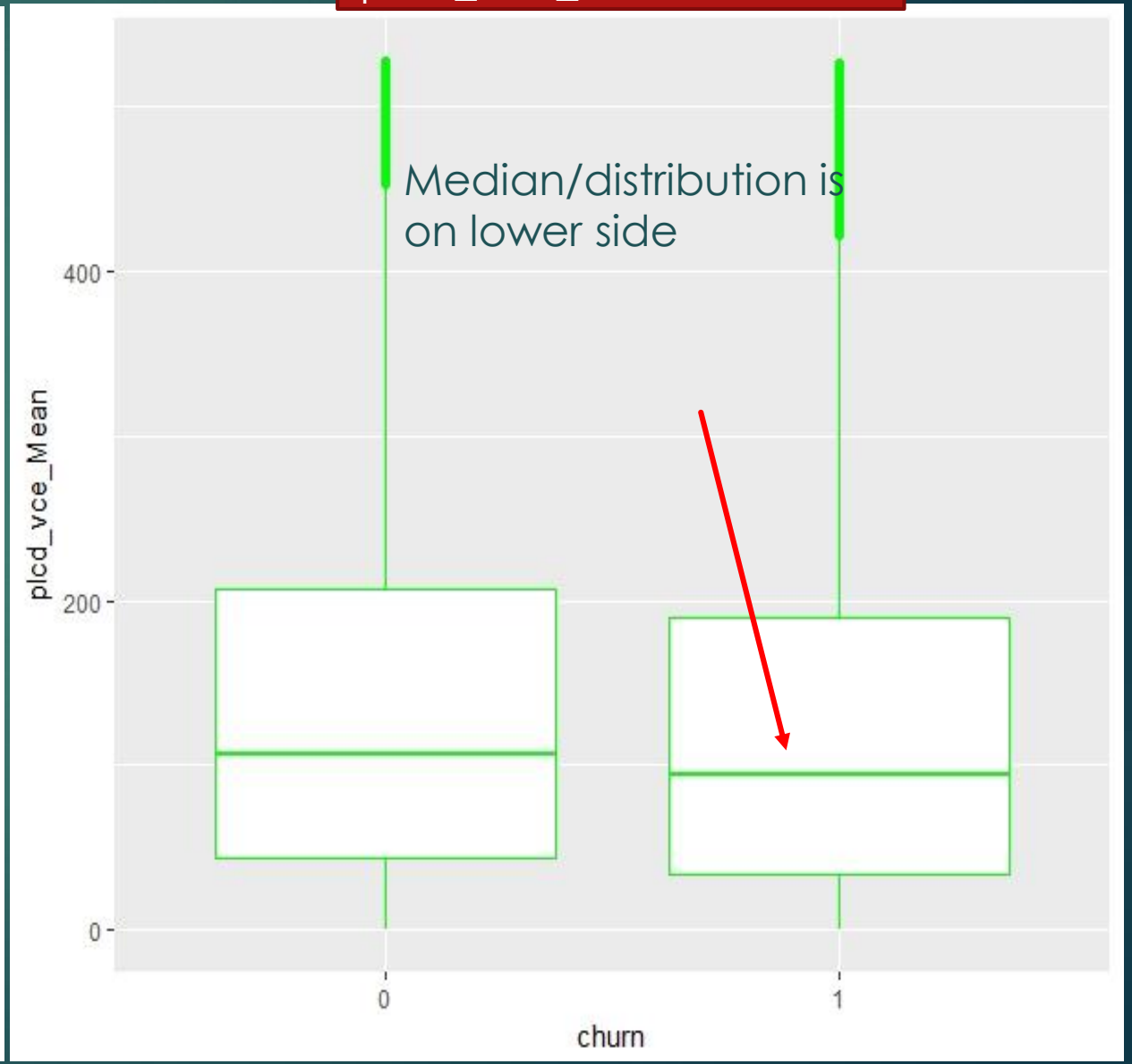


Data Exploration(Variance)

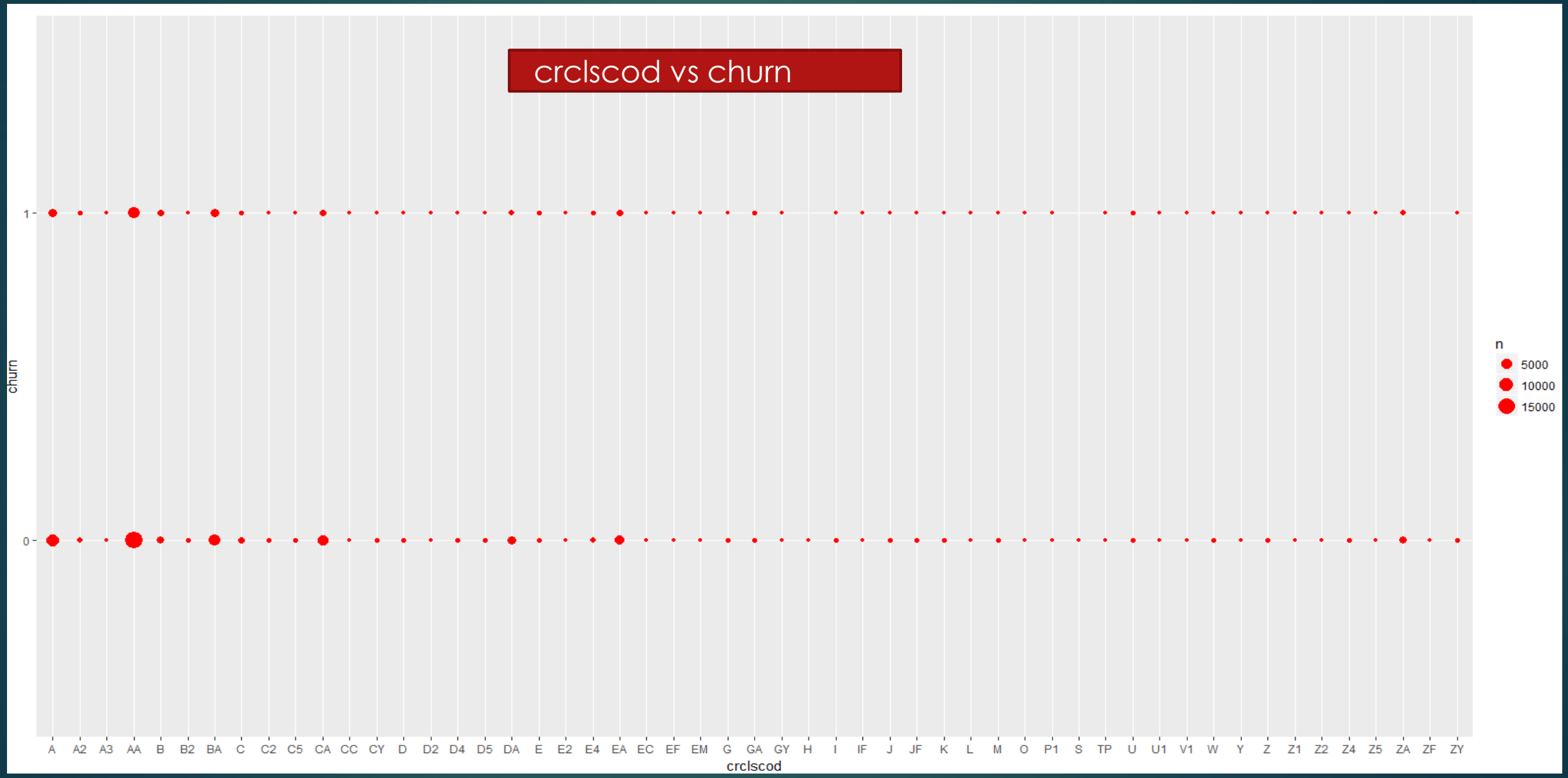
mou_Mean density



plcd_vce_Mean vs Churn

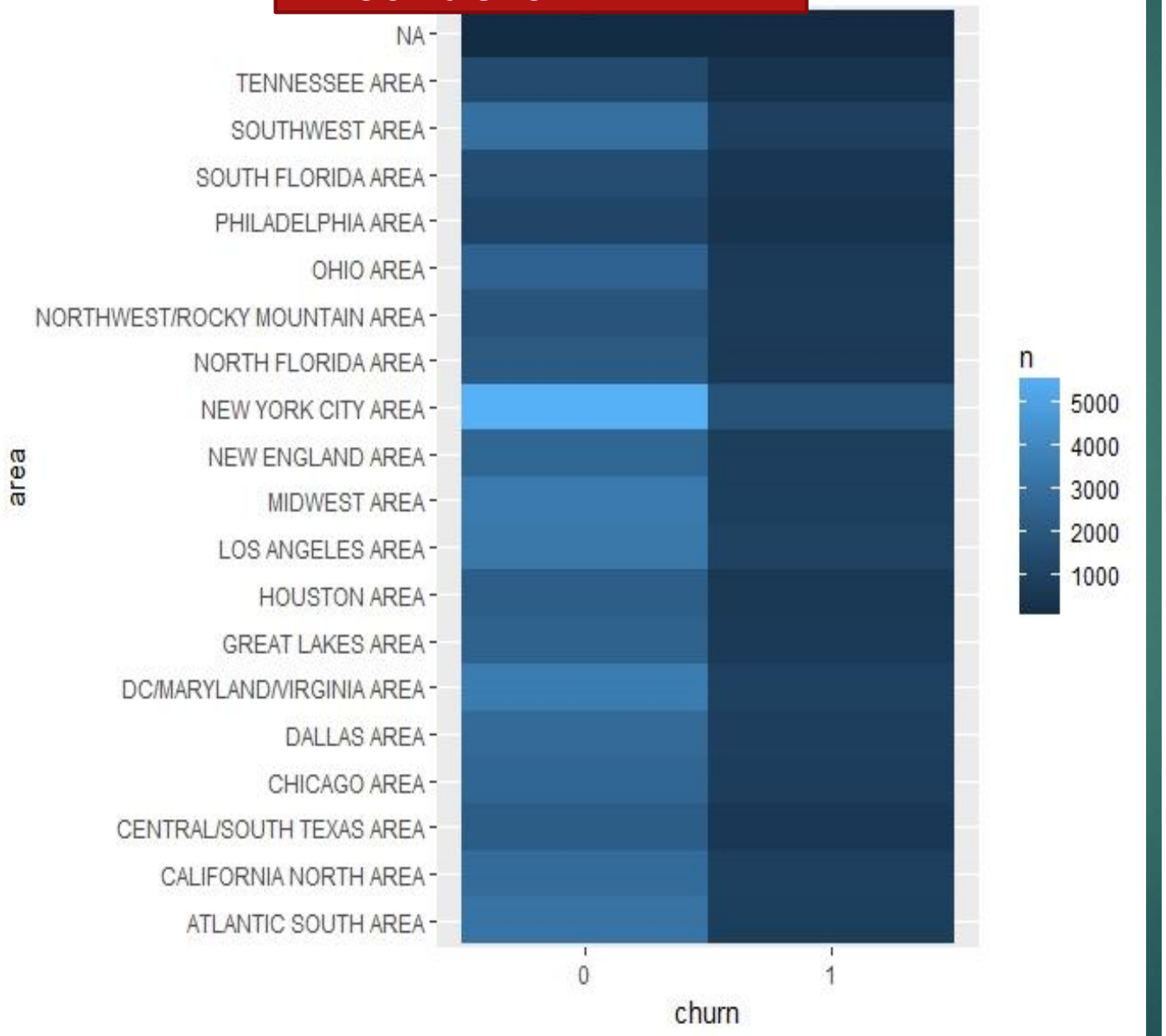


Data Exploration(Variance)

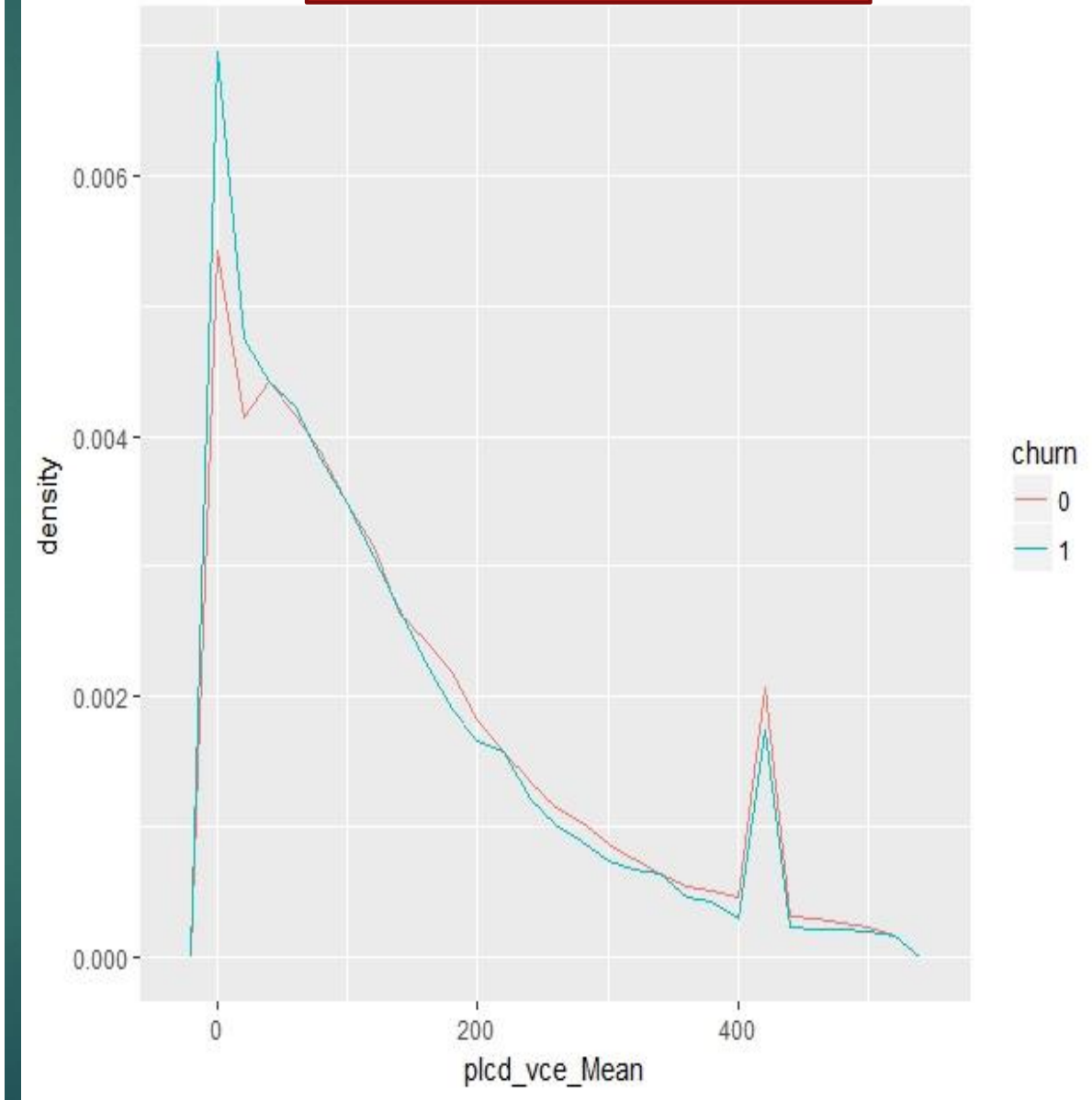


Data Exploration(Variance)

Area vs churn

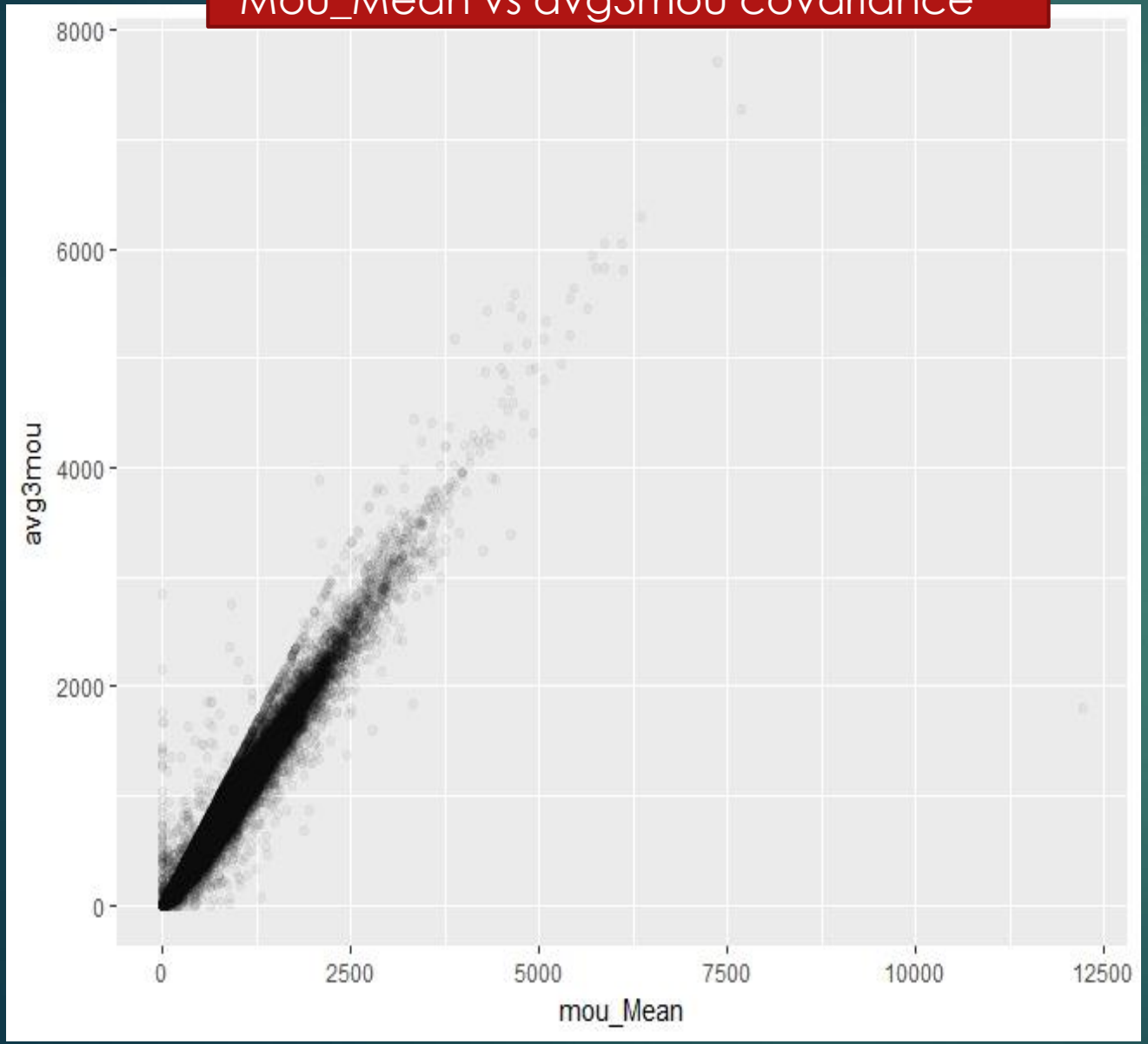


plcd_vce_Mean density

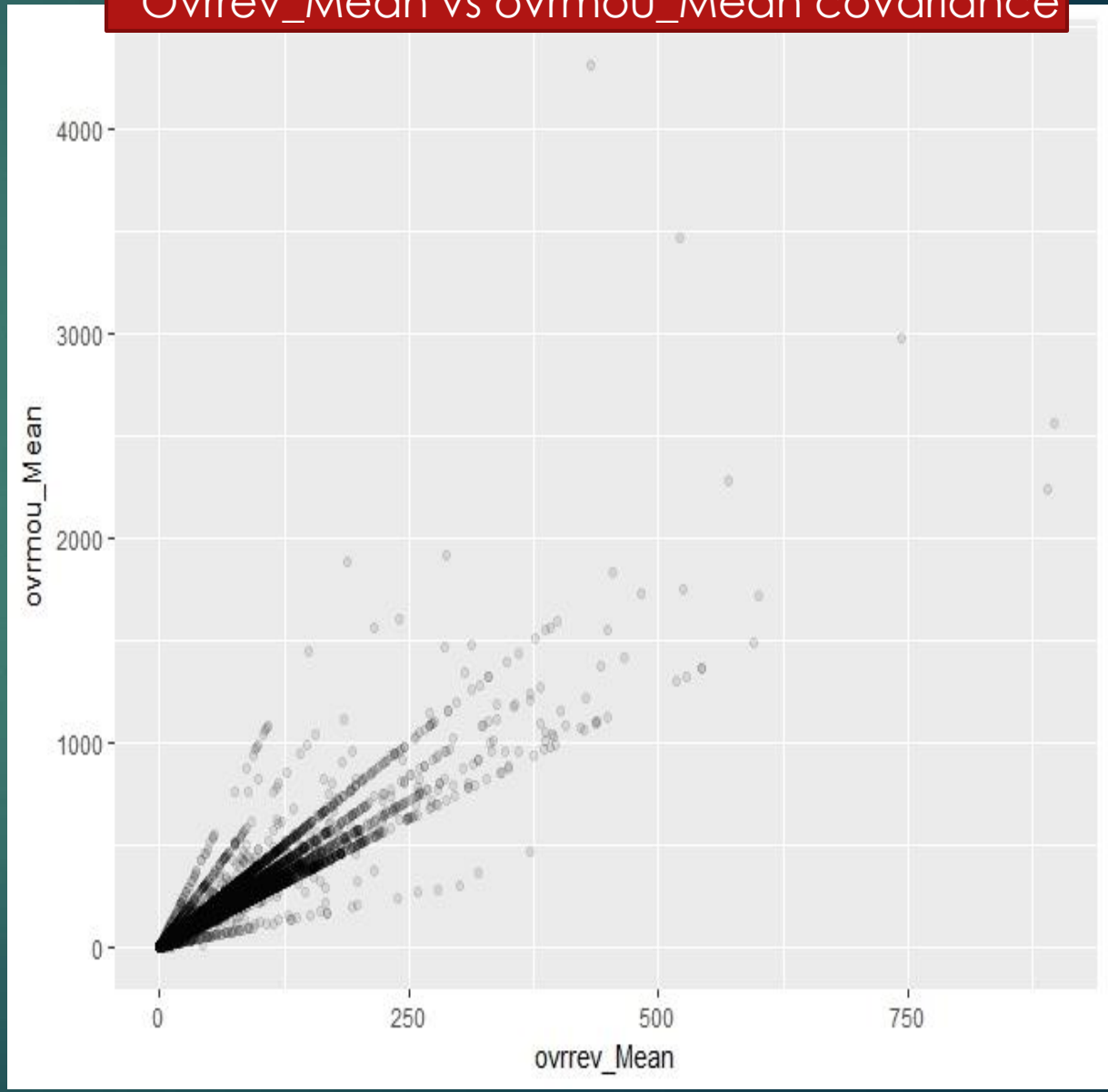


Data Exploration(Covariance)

Mou_Mean vs avg3mou covariance

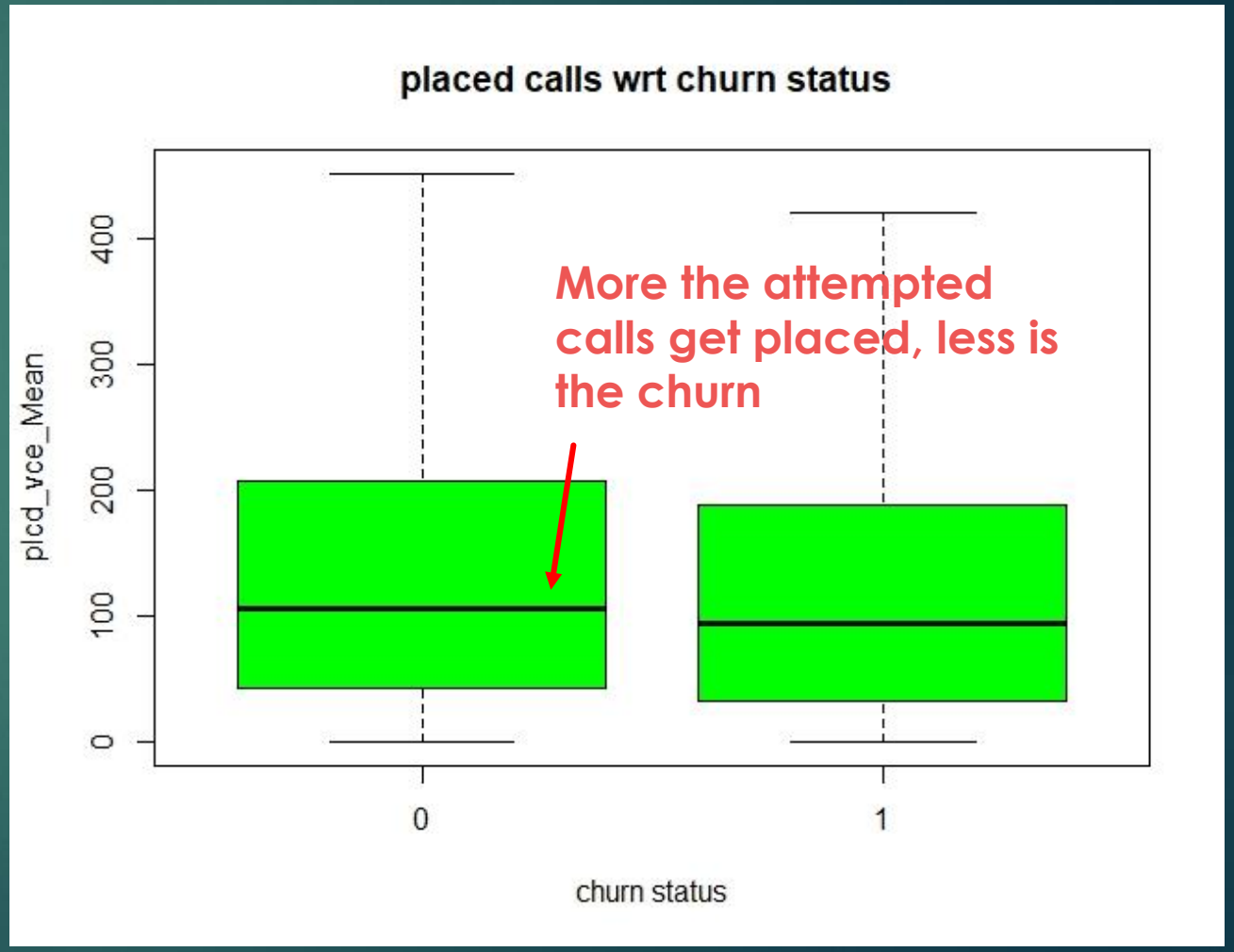


Ovrrev_Mean vs ovrmou_Mean covariance



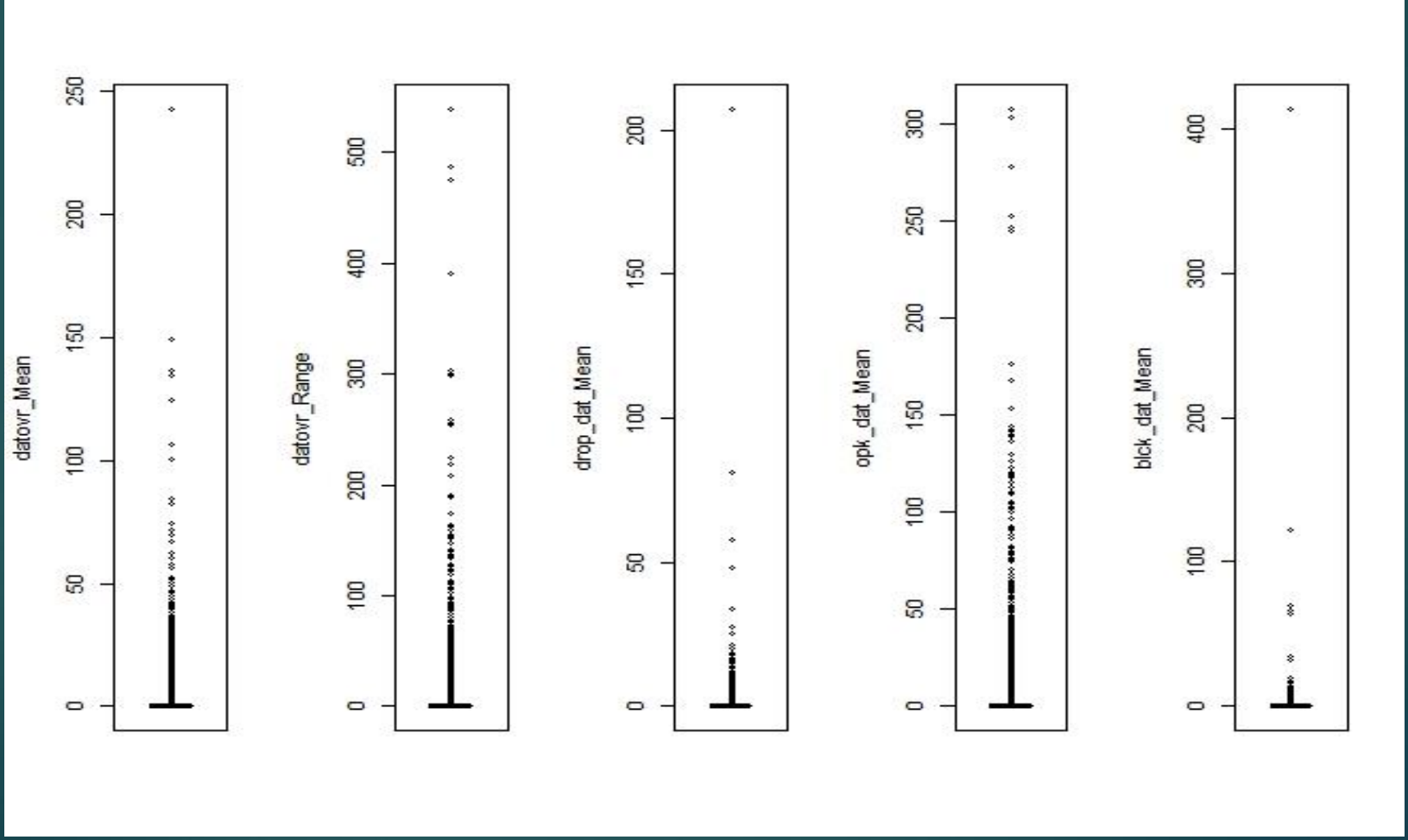
Effect of Network and Service quality on churn

- Analysis of variable 'plcd_vce_Mean':
  - If there are more dropped calls or blocked calls, it can very well be interpreted as having issues with network and should logically reflect into having more customers leaving the network provider for lack of quality service.
  - This model corroborates this fact by showing positive correlation of churn with drop_blk mean and drop_vce_range but when it comes to plcd_vce_Mean there is a negative correlation with churn. It means network and service quality have significant impact on the customer's decision to stay with the company or leave.



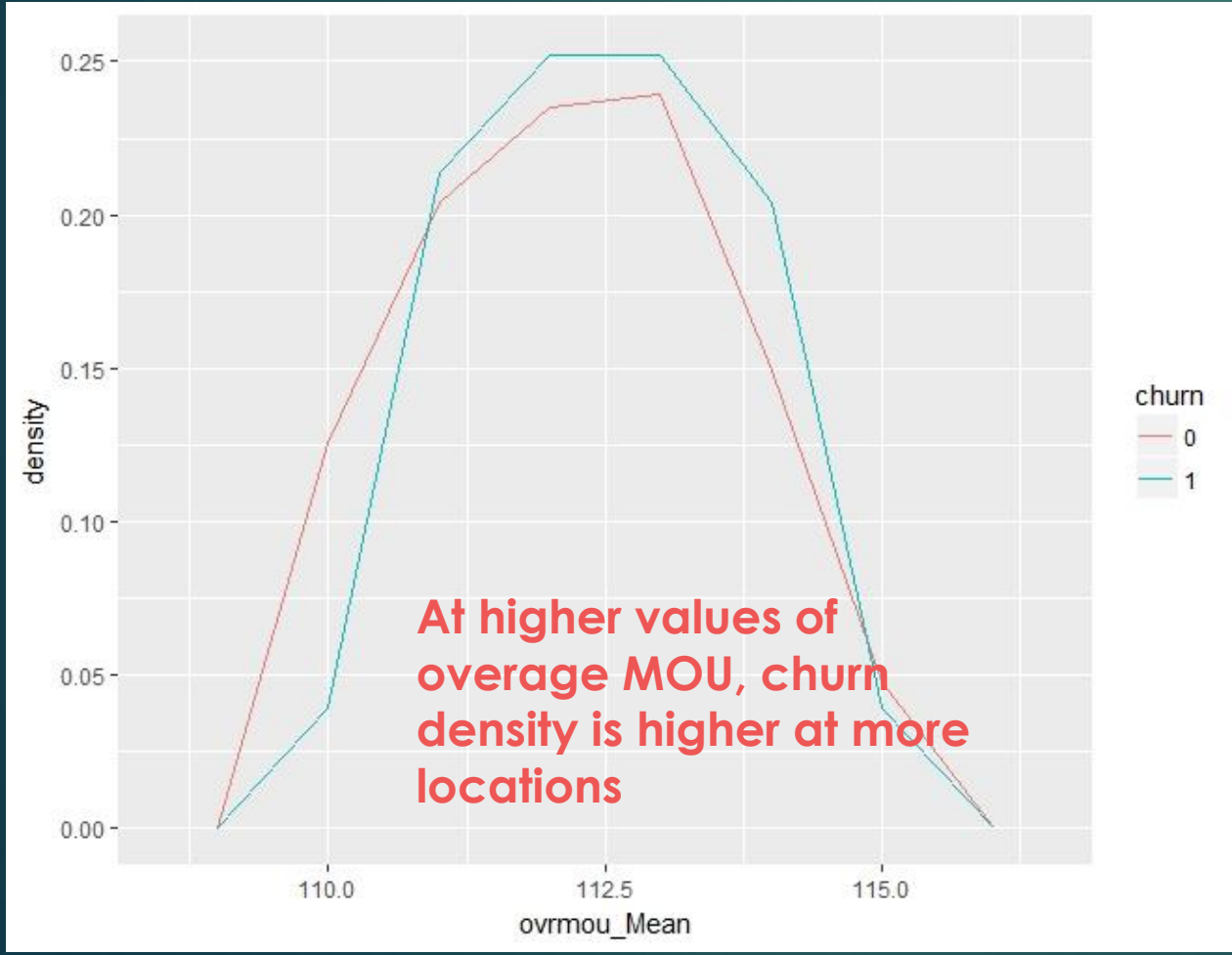
Effect of Data Usage on churn

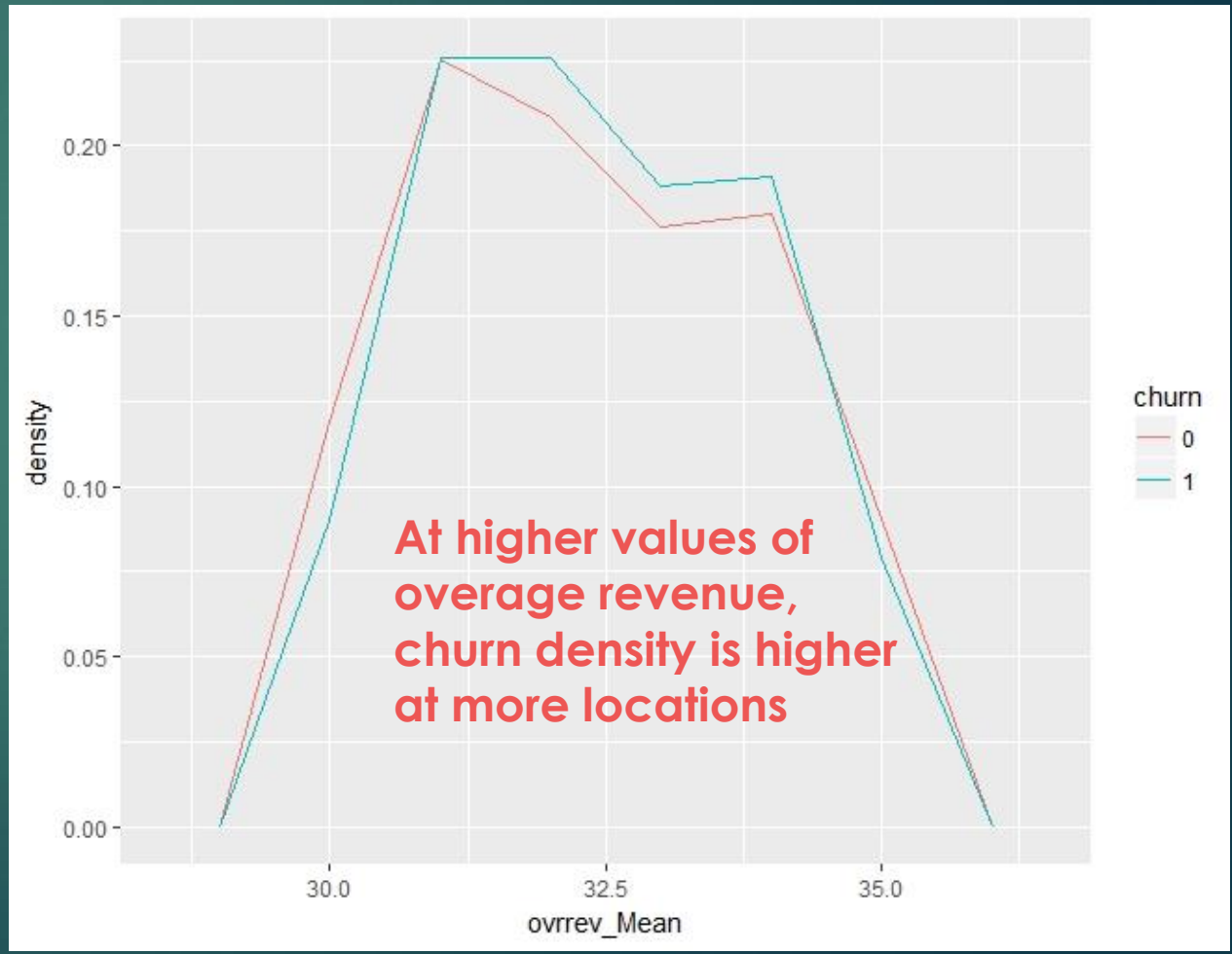
- All the data usage related variables have very skewed data (possibly incorrect), hence the sensitivity of churn to internet usage was not reflected in this model. But this may not give the right picture about the effect of internet on churn.



Rate Plan Migration Strategy

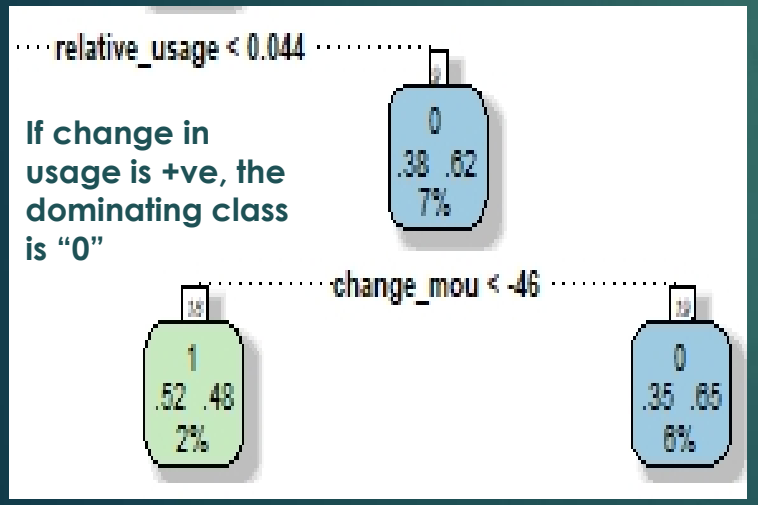
- Analysis of variables ovrrev_Mean and ovrmou_Mean on churn:
  - The higher the overage revenue and overage MOU usage, the more non-optimal is the existing plan for the customer. Therefore, rate plan migration strategy is an effective one.

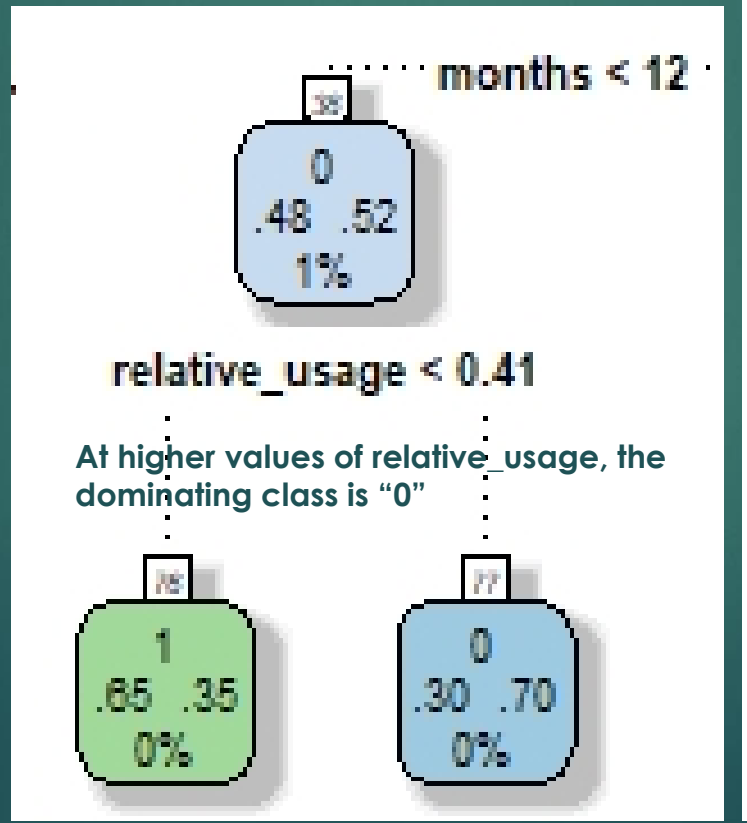


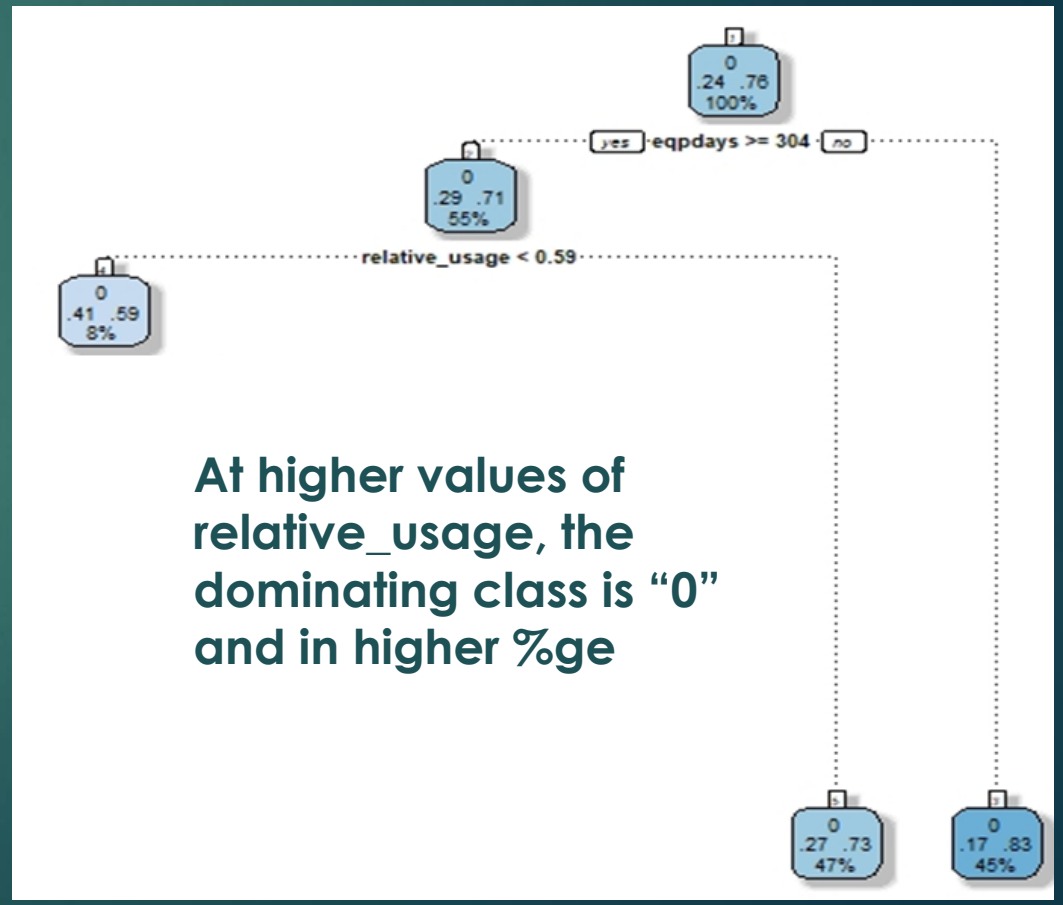


Usage Based (MOU) Strategy

- Analysis of variables change_mou and relative_usage on churn:
  - Variable relative_usage can be defined as average usage per month by the subscriber with current operator as a percentage of monthly usage over lifetime. It has negative correlation with churn. Usage Based Strategy is highly effective one as below.







Recommendations

- From the model, Usage Based Promotions (MOU) and Rate Plan Migration strategies are strongly recommended.
- For cut-off probability of 50%, the model accurately predicts customers who are going to default and not default 77% of the times. It means the accuracy rate is on the higher side. Hence, it is recommended that top 50%le of the customers who are likely to switch as predicted by this model should be set aside for any retention programme.
- Revenue Saving: As there is a constraint on the budget and only 20% of the subscribers can be contacted, we can further rank the subscribers among the above 50%le group with high revenue by sorting them by 'totrev' or 'rev_mean' values and the top 40%le subscribers among those can be selected for retention programme. This achieves both the objectives of company; 1) Controlling Churn 2) Revenue Saving

Getting Target Customers for Retention Campaigns

```
##.....................................................................................##
##Customer Targeting
##.....................................................................................##

test$probs<-probs

quantile(probs,probs=seq(0.1,1,0.1))

targeted<-test[test$probs>0.229,] ##Separate top 50%le of customers for churn

targeted_by_revenue<-arrange(targeted,-totrev) ##Arrange these 50%le customers by revenue

targeted_customers<-targeted_by_revenue[as.numeric(row.names(targeted_by_revenue))<0.40 * nrow(targeted_by_revenue),"Customer_ID"]
##40% of 50% is 20%. Hence we obtain 20% of the total customers to be targeted.
```

APPENDIX

```
delete_outliers<-function(x){
  data<-x
  for (i in 1:ncol(data)){
    if(class(data[,i])=="numeric" | class(data[,i])=="integer"){
      quantile<-quantile(data[,i],prob=c(0.25,0.75),na.rm=T)
      q_range<-2*IQR(data[,i])
      Min<-quantile[1]-q_range
      Max<-quantile[2]+q_range
      data[,i][data[,i]>Max]<-NA
      data[,i][data[,i]<Min]<-NA

    }else{
      x
    }
  }
  data
}
```

delete_outliers Function

```
delete_missing<-function(x){
  data<-x
  for (i in 1:ncol(data)){
    if(any(is.na(data[,i]))){
      data%>%filter(!is.na(data[,i]))->data
    }
  }
  data
}
```

delete_missing Function

```
impute_missing<-function(x){
  data<-x
  for (i in 1:ncol(data)){
    if(any(is.na(data[,i]))){
      if(class(data[,i])=="numeric" | class(data[,i])=="integer"){
        data%>%group_by(quantile=ntile(data[,i],10))->data2
        tbl1<-table(data2$quantile,data2$churn)/(nrow(data2)/10)
        index<-which(is.na(data2[,i]))
        tbl2<-table(data2$churn[index])/length(index)
        quant<-quantile(data[,i],p=(0:10)/10,na.rm=T)
        condition1<-tbl1[,1]>0.85*tbl2[1] & tbl1[,1]<1.15*tbl2[1]
        if(any(condition1) && sum(condition1)>1){
          diff<-abs(tbl1[condition1,1]-tbl2[1])
          condition2<-diff==min(diff)
          ind<-as.numeric(names(condition2[condition2]))
        }
        else{
          ind<-which(condition1)
        }
        condition<-length(ind)==1
        if (condition){
          data[,i][index]<-(quant[ind+1]+quant[ind])/2
        }
      } else if(class(data[,i])=="character"){
        tbl1<-table(data[,i],data$churn)
        good_rate<-tbl1[,1]/rowSums(tbl1)
        index<-which(is.na(data2[,i]))
        tbl2<-table(data2$churn[index])/length(index)

        condition1<-good_rate>0.9*tbl2[1] & good_rate<1.1*tbl2[1]
        if(any(condition1) && sum(condition1)>1){
          diff<-abs(good_rate[condition1]-tbl2[1])
          condition2<-diff==min(diff)
          ind<-names(condition2[condition2])
        }
        else{
          ind<-names(condition1)
        }
        condition<-length(ind)==1
        if (condition){
          data[,i][index]<-ind
        }
      }else{
        data
      }
    }

  }
  data
}
```

impute_missing Function