

Project 2.1

Project - Churn Prediction



Project 2.1
Project - Churn Prediction



Project 2.1
Project - Churn Prediction



ABSTRACT

Customer value analysis is critical for a good marketing and a customer relationship management strategy. An important component of this strategy is the customer retention rate. Customer retention rate has a strong impact on the customer lifetime value, and understanding the true value of a possible customer churn will help the company in its customer relationship management. Conventional statistical methods are very successful in predicting a customer churn. The goal of this study is to apply logistic regression techniques to predict a customer churn and analyze the churning and no-churning customers by using data provided for the project

Introduction

The subject of customer retention, loyalty, and churn is receiving attention in many industries. This is important in the customer lifetime value context. A company will have a sense of how much is really being lost because of the customer churn and the scale of the efforts that would be appropriate for retention campaign. The mass marketing approach cannot succeed in the diversity of consumer business today. Customer value analysis along with customer churn predictions will help marketing programs target more specific groups of customers. Personal retail banking sector is characterized by customers who stays with a company very long time. Customers usually give their financial business to one company and they won't switch the provider of their financial help very often. In the company's perspective this produces a stable environment for the customer relationship management. Although the continuous relationships with the customers the potential loss of revenue because of customer churn in this case can be huge.

This paper will present a customer churn analysis for the data provided in the project 2. The goal of this paper is twofold. First the churning customers are analyzed in R – Logistic Regression model after applying appropriate preprocessing techniques . The second stage we need to connect with Tableau and R for visualization, probability and prediction in Tableau using the trained and testing set.

Logistic regression Binomial (binary) logistic regression is a form of regression which is used in a situation when dependent is not a continuous variable but a state which may or may not happen, or a category in a specific classification . Logistic regression can be used to predict a discrete outcome on the basis of continuous and/or categorical variables. Multinomial logistic regression exists to handle the case of dependents with more classes than two. In the logistic regression there can be only one dependent variable. Logistic regression applies maximum likelihood estimation after transforming the dependent into a logistic variable [8]. Unlike the normal regression model the dependent variable in logistic regression is usually dichotomous: the dependent variable can take value 1 with probability q and value 0 with probability 1-q.

Preprocessing

The Churn data provided for the project 2 is checked for the preprocessing requirements if any. Normally, the missing values requires creates many problems during analysis and this requires preprocessing and imputation etc., In this project the Amelia library is used to identify the missing values and as per the graph given below there is no missing value in the data.

```
library(Amelia)
```

```
## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2018 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##

any(is.na(Churn))

## [1] FALSE

# visualize the missing values using the missing map from the Amelia package
missmap(Churn,col=c("yellow","red"))

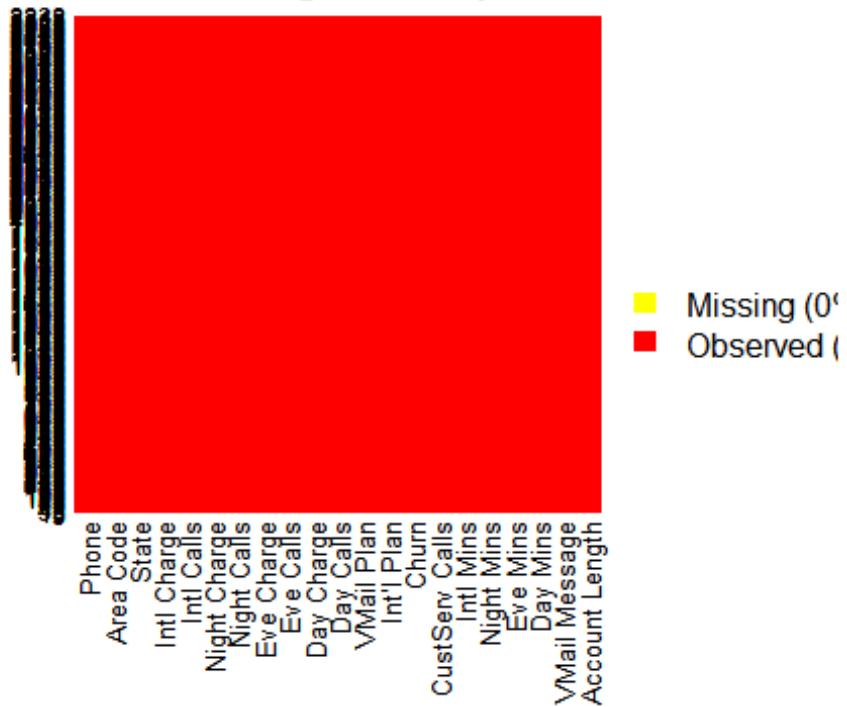
## Warning in if (class(obj) == "amelia") {: the condition has length > 1 and
## only the first element will be used

## Warning: Unknown or uninitialized column: 'arguments'.

## Warning: Unknown or uninitialized column: 'arguments'.

## Warning: Unknown or uninitialized column: 'imputations'.
```

Missingness Map



Excluding phone and state from the data set as they are not very important

```
mydata2<-Churn[, -21]
mydata<-mydata2[, -19]
sapply(mydata, function(x) sum(is.na(x)))

## Account Length  VMail Message      Day Mins      Eve Mins      Night Mins
##          0          0          0          0          0          0
##  Intl Mins  CustServ Calls      Churn  Int'l Plan  VMail Plan
##          0          0          0          0          0          0
##  Day Calls   Day Charge    Eve Calls  Eve Charge  Night Calls
##          0          0          0          0          0          0
```

```
##   Night Charge      Intl Calls      Intl Charge      Area Code
##                 0              0              0              0

mydata <- mydata[complete.cases(mydata), ]
intrain<- createDataPartition(mydata$Churn,p=0.8,list=FALSE)
set.seed(2017)
training<- mydata[intrain,]
testing<- mydata[-intrain,]
dim(training); dim(testing)

## [1] 2667 19

## [1] 666 19

library (data.table)

library (plyr)
library (stringr)

##
## Attaching package: 'stringr'

## The following object is masked from 'package:strucchange':
## 
##     boundary

LogModel <- glm(Churn ~ .,family=binomial(link="logit"),data=training)
print(summary(LogModel))

##
## Call:
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training)
## 
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.1332  -0.5222  -0.3425  -0.1992   3.2941
## 
## Coefficients:
```

```

##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -7.420e+00  1.024e+00 -7.249  4.2e-13 ***
## `Account Length`    1.141e-03  1.533e-03  0.744  0.456828
## `VMail Message`     4.207e-02  2.027e-02  2.075  0.037983 *
## `Day Mins`          -1.341e+00  3.624e+00 -0.370  0.711274
## `Eve Mins`          -6.519e-01  1.813e+00 -0.360  0.719209
## `Night Mins`         -4.968e-01  9.739e-01 -0.510  0.610007
## `Intl Mins`          -7.267e-01  5.875e+00 -0.124  0.901560
## `CustServ Calls`    5.299e-01  4.387e-02 12.079 < 2e-16 ***
## `Int'l Plan`         2.096e+00  1.610e-01 13.022 < 2e-16 ***
## `VMail Plan`         -2.319e+00  6.533e-01 -3.550  0.000385 ***
## `Day Calls`          2.586e-03  3.042e-03  0.850  0.395283
## `Day Charge`         7.962e+00  2.132e+01  0.373  0.708808
## `Eve Calls`          8.044e-04  3.055e-03  0.263  0.792319
## `Eve Charge`         7.744e+00  2.133e+01  0.363  0.716583
## `Night Calls`        -2.353e-03  3.160e-03 -0.745  0.456435
## `Night Charge`       1.109e+01  2.164e+01  0.513  0.608260
## `Intl Calls`         -1.044e-01  2.812e-02 -3.712  0.000206 ***
## `Intl Charge`         2.960e+00  2.176e+01  0.136  0.891786
## `Area Code`          -6.509e-05  1.462e-03 -0.045  0.964500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2230.1 on 2666 degrees of freedom
## Residual deviance: 1755.5 on 2648 degrees of freedom
## AIC: 1793.5
##
## Number of Fisher Scoring iterations: 6

anova(LogModel, test="Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit

```

```

##  

## Response: Churn  

##  

## Terms added sequentially (first to last)  

##  

##  

## Df Deviance Resid. Df Resid. Dev Pr(>Chi)  

## NULL 2666 2230.1  

## `Account Length` 1 1.550 2665 2228.6 0.2132053  

## `VMail Message` 1 26.557 2664 2202.0 2.559e-07 ***  

## `Day Mins` 1 92.031 2663 2110.0 < 2.2e-16 ***  

## `Eve Mins` 1 17.926 2662 2092.1 2.297e-05 ***  

## `Night Mins` 1 1.924 2661 2090.1 0.1654485  

## `Intl Mins` 1 9.890 2660 2080.2 0.0016614 **  

## `CustServ Calls` 1 130.903 2659 1949.3 < 2.2e-16 ***  

## `Int'l Plan` 1 163.266 2658 1786.1 < 2.2e-16 ***  

## `VMail Plan` 1 14.094 2657 1772.0 0.0001739 ***  

## `Day Calls` 1 0.719 2656 1771.3 0.3966047  

## `Day Charge` 1 0.051 2655 1771.2 0.8214162  

## `Eve Calls` 1 0.017 2654 1771.2 0.8948703  

## `Eve Charge` 1 0.132 2653 1771.1 0.7158769  

## `Night Calls` 1 0.536 2652 1770.5 0.4641722  

## `Night Charge` 1 0.190 2651 1770.3 0.6629360  

## `Intl Calls` 1 14.804 2650 1755.5 0.0001193 ***  

## `Intl Charge` 1 0.019 2649 1755.5 0.8911511  

## `Area Code` 1 0.002 2648 1755.5 0.9644909  

## ---  

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

testing$Churn <- as.character(testing$Churn)
testing$Churn[testing$Churn=="No"] <- "0"
testing$Churn[testing$Churn=="Yes"] <- "1"
fitted.results <- predict(LogModel,newdata=testing,type='response')
fitted.results

sapply(mydata, sd)

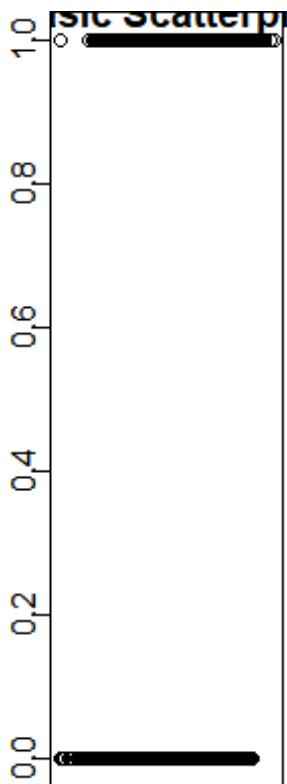
```

```

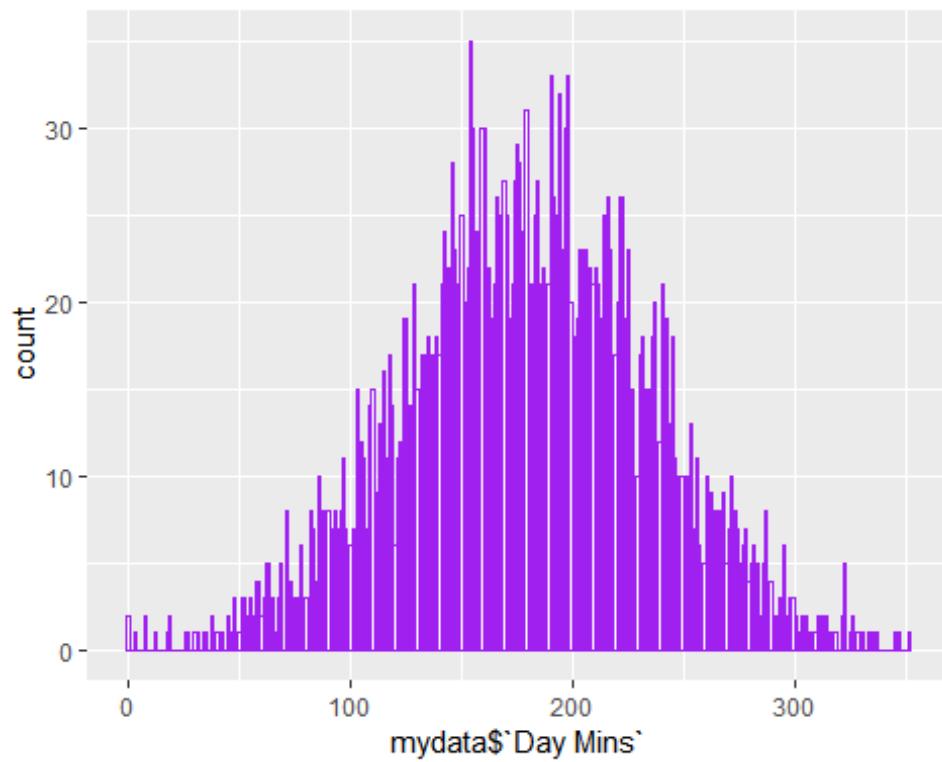
## Account Length VMail Message      Day Mins      Eve Mins      Night Mins
##   39.8221059    13.6883654    54.4673892    50.7138444    50.5738470
##   Intl Mins CustServ Calls      Churn       Int'l Plan     VMail Plan
##   2.7918395     1.3154910     0.3520674     0.2958791     0.4473979
##   Day Calls     Day Charge      Eve Calls     Eve Charge     Night Calls
##   20.0690842     9.2594346    19.9226253    4.3106676    19.5686093
##   Night Charge   Intl Calls     Intl Charge   Area Code
##   2.2758728     2.4612143     0.7537726    42.3712905

plot.new()
plot(mydata$Churn ~mydata$`Day Mins`)
title('Basic Scatterplot')

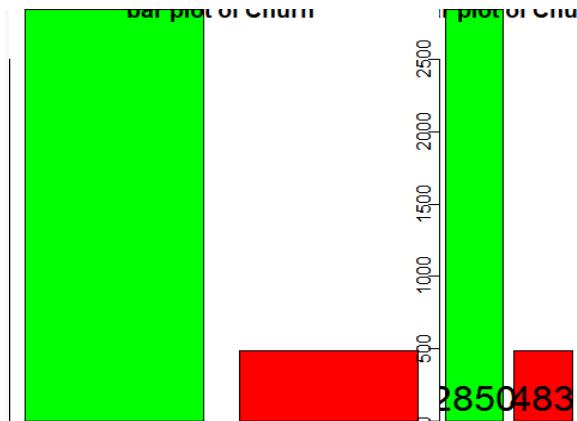
```



```
ggplot(mydata, aes(x=mydata$`Day Mins`)) + geom_histogram(binwidth = 1, fill = "white", color = "purple")
```



```
#Randomly split data into train and test set  
#80% will be assigned to train set, 20% will be assigned to tst set  
barplot(table(mydata$Churn), col= c("green", "red"), main='bar plot of Churn')  
text(barplot(table(mydata$Churn), col =c('green' , 'red'), main='bar plot of Churn'), 0,table(mydata$Churn)  
, cex =2 , pos =3)
```



```
#proportion
round(prop.table(table(mydata$Churn))*100,digits = 2)

##
##      0      1 (As per this churn is around 15%)
## 85.51 14.49

mydata_train<-mydata_n[1:2666,]
mydata_test<-mydata_n[2667:3333,]
mydata_train_labels<-mydata_n[1:2666,7]
mydata_test_labels<-mydata_n[2667:3333,7]

sapply(mydata_n, sd)

## Account.Length  VMail.Message      Day.Mins      Eve.Mins      Night.Mins
## 0.1645542     0.2683993     0.1552662     0.1394387     0.1360243
##  Intl.Mins CustServ.Calls      Churn      Int.l.Plan      VMail.Plan
## 0.1395920     0.1461657     0.3520674     0.2958791     0.4473979
##  Day.Calls    Day.Charge      Eve.Calls      Eve.Charge      Night.Calls
## 0.1216308     0.1552554     0.1171919     0.1394587     0.1378071
##  Night.Charge   Intl.Calls      Intl.Charge
## 0.1360354     0.1230607     0.1395875
```

```

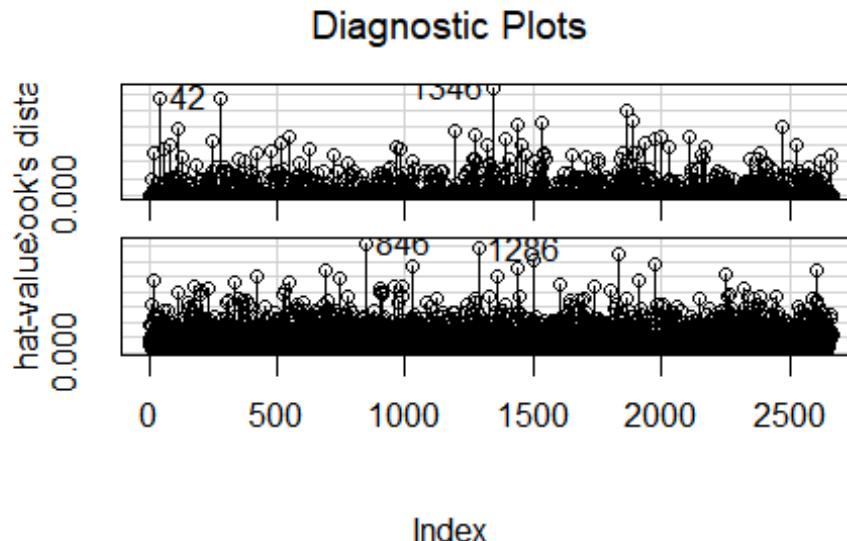
#Forward elimination
#Lower AIC indicates a better model
forward <- step(glm(Churn ~ 1, data = mydata_train), direction = 'forward', scope = ~Account.Length+VMail.Message+Day.Mins + Eve.Mins +
  Night.Mins + Intl.Mins + CustServ.Calls + Int.l.Plan + VMail.Plan +
  Day.Calls + Day.Charge + Eve.Calls + Eve.Charge + Night.Calls +
  Night.Charge + Intl.Calls + Intl.Charge)

logit<- glm(Churn ~Account.Length+Day.Mins+ Day.Charge +CustServ.Calls+VMail.Plan +Eve.Mins+ Eve.Charge+VMail.Message+Day.Calls +Eve.Calls+ Intl.Mins + Night.Calls+Intl.Calls, data = mydata_train, family = "binomial")
summary(logit)

##
## Call:
## glm(formula = Churn ~ Account.Length + Day.Mins + Day.Charge +
##       CustServ.Calls + VMail.Plan + Eve.Mins + Eve.Charge + VMail.Message +
##       Day.Calls + Eve.Calls + Intl.Mins + Night.Calls + Intl.Calls,
##       family = "binomial", data = mydata_train)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.7136   -0.5583   -0.3989   -0.2492    3.0223
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.4168    0.6846 -10.834 < 2e-16 ***
## Account.Length 0.1336    0.3696   0.362  0.71771
## Day.Mins     191.3463  1231.6229   0.155  0.87654
## Day.Charge   -187.0655  1231.7103  -0.152  0.87929
## CustServ.Calls  3.9702    0.3798  10.455 < 2e-16 ***
## VMail.Plan    -1.7024    0.6098  -2.792  0.00524 **
## Eve.Mins      462.5882  642.8816   0.720  0.47180
## Eve.Charge    -460.1275  642.7876  -0.716  0.47410
## VMail.Message  1.5122    0.9823   1.539  0.12371
## Day.Calls      0.3629    0.4940   0.735  0.46264

```

```
## Eve.Calls      0.4090    0.5063    0.808   0.41921
## Intl.Mins     2.0353    0.4447    4.576  4.73e-06 ***
## Night.Calls    0.1657    0.4403    0.376   0.70665
## Intl.Calls    -1.6524    0.5357   -3.085   0.00204 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2162.0 on 2665 degrees of freedom
## Residual deviance: 1867.4 on 2652 degrees of freedom
## AIC: 1895.4
##
## Number of Fisher Scoring iterations: 5
```



```
confint(logit)

## Waiting for profiling to be done...

##              2.5 %      97.5 %
## (Intercept) -8.7761881  -6.0914407
## Account.Length   -0.5919165   0.8578438
## Day.Mins     -2224.7782913 2605.8924991
## Day.Charge    -2601.7748293 2229.2388877
## CustServ.Calls 3.2299992   4.7198281
## VMail.Plan     -2.9307458   -0.5373346
## Eve.Mins      -796.3103582 1725.2392867
## Eve.Charge    -1722.5889512  798.5919581
```

Confidence interval using Log-Likelihood

```

## VMail.Message    -0.4038228   3.4527140
## Day.Calls       -0.6043867   1.3330979
## Eve.Calls        -0.5823325   1.4036271
## Intl.Mins        1.1687246   2.9129356
## Night.Calls     -0.6973052   1.0293210
## Intl.Calls      -2.7201822   -0.6195282

exp(logit$coefficients)

##   (Intercept) Account.Length      Day.Mins      Day.Charge CustServ.Calls
## 6.010934e-04  1.142977e+00  1.260826e+83  5.734114e-82  5.299764e+01
##   VMail.Plan      Eve.Mins      Eve.Charge  VMail.Message      Day.Calls
## 1.822392e-01  7.933806e+200  1.476321e-200  4.536688e+00  1.437439e+00
##   Eve.Calls      Intl.Mins      Night.Calls      Intl.Calls
## 1.505333e+00  7.654355e+00  1.180211e+00  1.915949e-01

# Logistic regression model:
fit <- glm(Churn~.,data =mydata_train ,family = binomial(link='logit'))
summary(fit)

##
## Call:
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = mydata_train)
##
## Deviance Residuals:
##      Min        1Q        Median        3Q        Max 
## -1.9680  -0.5111  -0.3376  -0.1979   3.1864 
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -8.42021   0.76774 -10.967 < 2e-16 ***
## Account.Length 0.01841   0.38853   0.047  0.962208    
## VMail.Message  1.53996   1.01473   1.518  0.129114    
## Day.Mins      213.42834 1293.28315   0.165  0.868922    
## Eve.Mins       592.02181  674.48943   0.878  0.380088    
## Night.Mins     -110.40724  368.95516  -0.299  0.764755    
## Intl.Mins      -287.85749  120.73173  -2.384  0.017113 *  

```

```

## CustServ.Calls    4.51834   0.40343  11.200 < 2e-16 ***
## Int.l.Plan       2.06924   0.16257  12.729 < 2e-16 ***
## VMail.Plan      -1.87056   0.63215  -2.959  0.003086 **
## Day.Calls        0.44168   0.51623   0.856  0.392222
## Day.Charge     -209.05873 1293.37170  -0.162  0.871590
## Eve.Calls        0.43463   0.53648   0.810  0.417852
## Eve.Charge     -589.41466  674.38935  -0.874  0.382120
## Night.Calls     0.01869   0.46116   0.041  0.967668
## Night.Charge    111.62067  368.91878   0.303  0.762224
## Intl.Calls      -1.92925   0.56673  -3.404  0.000664 ***
## Intl.Charge     289.76049  120.71956   2.400  0.016383 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2162.0 on 2665 degrees of freedom
## Residual deviance: 1702.5 on 2648 degrees of freedom
## AIC: 1738.5
##
## Number of Fisher Scoring iterations: 6

library(MASS)
step_fit <- stepAIC(fit,method='backward')

summary(step_fit)

##
## Call:
## glm(formula = Churn ~ VMail.Message + Day.Mins + Eve.Mins + Intl.Mins +
##       CustServ.Calls + Int.l.Plan + VMail.Plan + Night.Charge +
##       Intl.Calls + Intl.Charge, family = binomial(link = "logit"),
##       data = mydata_train)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.9778  -0.5124  -0.3392  -0.2020   3.1476

```

```

## 
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -7.8769    0.5686 -13.853 < 2e-16 ***
## VMail.Message         1.5052    1.0120   1.487  0.136929
## Day.Mins              4.4084    0.4297  10.259 < 2e-16 ***
## Eve.Mins              2.5099    0.4622   5.430 5.64e-08 ***
## Intl.Mins             -291.5210 120.5655  -2.418 0.015608 *
## CustServ.Calls        4.5206    0.4022  11.240 < 2e-16 ***
## Int.l.Plan            2.0630    0.1622  12.721 < 2e-16 ***
## VMail.Plan             -1.8555   0.6300  -2.945 0.003230 **
## Night.Charge           1.2494    0.4652   2.686 0.007235 **
## Intl.Calls             -1.9404   0.5652  -3.433 0.000596 ***
## Intl.Charge            293.4337 120.5539   2.434 0.014931 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 2162.0 on 2665 degrees of freedom
## Residual deviance: 1704.9 on 2655 degrees of freedom
## AIC: 1726.9
## 
## Number of Fisher Scoring iterations: 6

confint(step_fit)

## Waiting for profiling to be done...

##                               2.5 %      97.5 %
## (Intercept)      -9.0107622 -6.7808248
## VMail.Message    -0.4650469  3.5078770
## Day.Mins          3.5752571  5.2606938
## Eve.Mins          1.6088317  3.4218238
## Intl.Mins         -528.7346134 -55.8150001
## CustServ.Calls   3.7372142  5.3151528
## Int.l.Plan        1.7458847  2.3822001

```

```

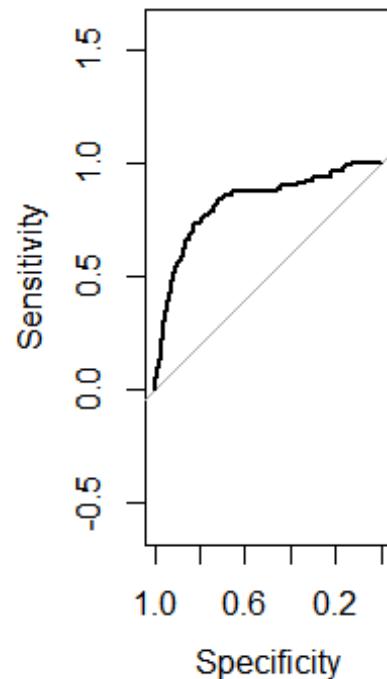
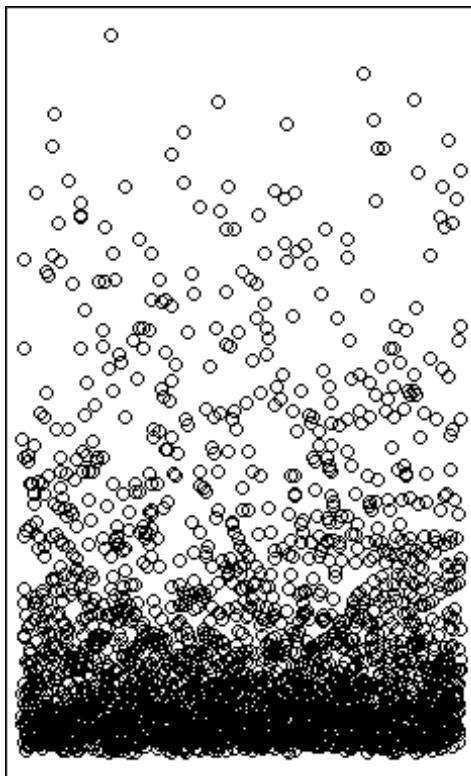
## VMail.Plan      -3.1260647  -0.6530832
## Night.Charge    0.3391697   2.1636548
## Intl.Calls     -3.0675068  -0.8511819
## Intl.Charge     57.7541777  530.6286565

#ANOVA on base model
anova(fit,test = 'Chisq')

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Churn
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL              2665      2162.0
## Account.Length  1    0.277    2664    2161.8 0.5985480
## VMail.Message    1   19.675    2663    2142.1 9.180e-06 ***
## Day.Mins         1  103.623    2662    2038.5 < 2.2e-16 ***
## Eve.Mins         1   23.581    2661    2014.9 1.197e-06 ***
## Night.Mins       1    3.199    2660    2011.7 0.0736818 .
## Intl.Mins        1   17.379    2659    1994.3 3.062e-05 ***
## CustServ.Calls   1  111.293    2658    1883.0 < 2.2e-16 ***
## Int.l.Plan        1  152.056    2657    1730.9 < 2.2e-16 ***
## VMail.Plan        1    8.313    2656    1722.6 0.0039361 **
## Day.Calls         1    1.004    2655    1721.6 0.3164303
## Day.Charge        1    0.101    2654    1721.5 0.7509001
## Eve.Calls         1    0.705    2653    1720.8 0.4010942
## Eve.Charge        1    0.752    2652    1720.1 0.3857120
## Night.Calls       1    0.000    2651    1720.1 0.9948372
## Night.Charge      1    0.088    2650    1720.0 0.7668220
## Intl.Calls        1   11.638    2649    1708.3 0.0006464 ***
## Intl.Charge        1    5.795    2648    1702.5 0.0160706 *

```

```
## ---  
pred <- predict(fit,newdata = mydata_test,type ='response')  
#check the AUC curve  
library(pROC)  
  
g <- roc( Churn~ pred, data = mydata_test)  
g  
  
##  
## Call:  
## roc.formula(formula = Churn ~ pred, data = mydata_test)  
##  
## Data: pred in 558 controls (Churn 0) < 109 cases (Churn 1).  
## Area under the curve: 0.8266  
  
plot(g)
```



```
library(caret)
#with default prob cut 0.50
mydata_test$pred_Churn <- ifelse(pred<0.8, 'yes', 'no')

table(mydata_test$pred_Churn,mydata_test$Churn)

##
##          0   1
## no      1   3
## yes    557 106

#training split of churn classes
round(table(mydata_train$Churn)/nrow(mydata_train),2)*100
```

```

##  

## 0 1  

## 86 14

# test split of churn classes  

round(table(mydata_test$Churn)/nrow(mydata_test),2)*100

##  

## 0 1  

## 84 16

#predicted split of churn classes  

round(table(mydata_test$pred_Churn)/nrow(mydata_test),2)*100

##  

## no yes  

## 1 99

#create confusion matrix  

#confusionMatrix(mydata_test$Churn,mydata_test$pred_Churn)  

#how do we create a cross validation scheme  

control <- trainControl(method = 'repeatedcv',  

                        number = 10,  

                        repeats = 3)  

seed <- 7  

metric <- 'Accuracy'  

set.seed(seed)  

fit_default <- train(Churn~.,  

                      data = mydata_train,  

                      method = 'glm',  

                      metric = NaN,  

                      trControl = control)

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to
## do classification? If so, use a 2 level factor as your outcome column.

```

```
## Warning in train.default(x, y, weights = w, ...): The metric "NaN" was not
## in the result set. RMSE will be used instead.

print(fit_default)

## Generalized Linear Model
##
## 2666 samples
##    17 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 2399, 2400, 2399, 2399, 2400, 2399, ...
## Resampling results:
##
##      RMSE      Rsquared      MAE
## 0.3171573  0.1668586  0.2165651

library(caret)
varImp(step_fit)

##          Overall
## VMail.Message  1.487325
## Day.Mins      10.258636
## Eve.Mins      5.429922
## Intl.Mins     2.417948
## CustServ.Calls 11.239954
## Int.l.Plan    12.720766
## VMail.Plan    2.945000
## Night.Charge   2.685836
## Intl.Calls     3.433245
## Intl.Charge    2.434045

varImp(fit_default)

## glm variable importance
##
```

```

## Overall
## Int.l.Plan    100.0000
## CustServ.Calls 79.36943
## Intl.Calls    21.54437
## VMail.Plan    17.99874
## Intl.Charge    15.22120
## Intl.Mins     15.11822
## VMail.Message   6.87083
## Day.Calls      5.43601
## Eve.Mins       5.43158
## Eve.Charge      5.40848
## Eve.Calls       4.87659
## Night.Charge    3.45603
## Night.Mins      3.43761
## Day.Mins        1.44730
## Day.Charge       1.42506
## Account.Length  0.02108
## Night.Calls     0.00000

library(devtools)
library(woe)

library(riv)

iv_df <- iv.mult(mydata_train, y="Churn", summary=TRUE, verbose=TRUE)

iv_df

##          Variable InformationValue Bins ZeroBins Strength
## 1      Day.Charge    0.643151413   6       0 Very strong
## 2      Day.Mins     0.643151413   6       0 Very strong
## 3 CustServ.Calls   0.158681659   2       0 Average
## 4      Eve.Mins     0.149576165   5       0 Average
## 5      Eve.Charge    0.149310982   5       0 Average
## 6      Intl.Charge   0.097357797   4       0 Weak
## 7      Intl.Mins     0.097357797   4       0 Weak
## 8 VMail.Message    0.081622650   2       0 Weak

```

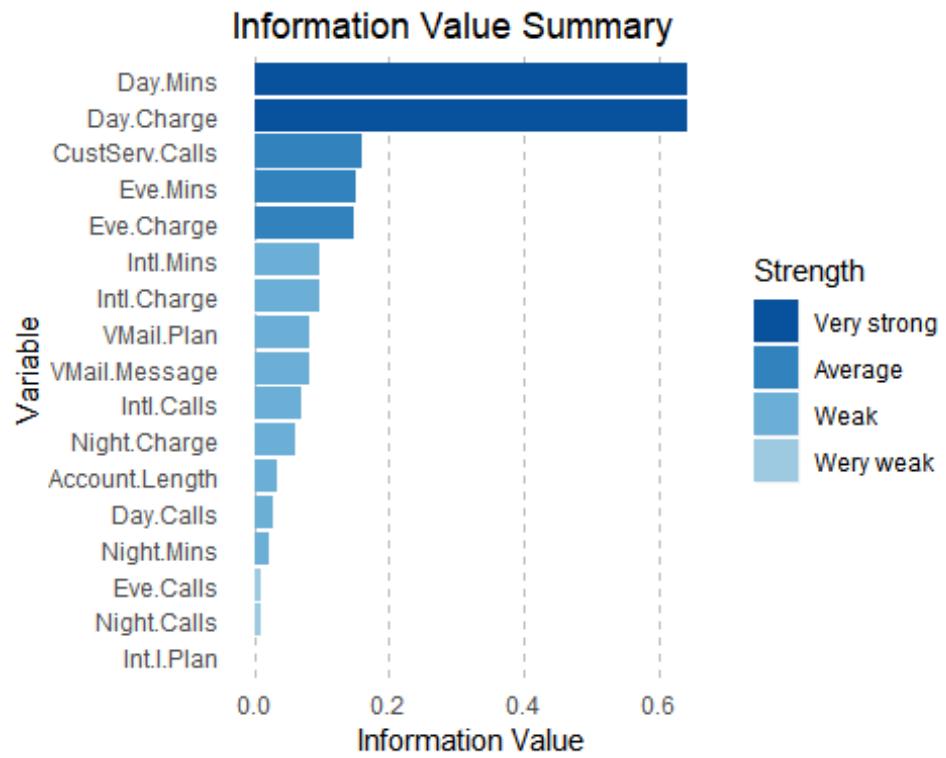
```

## 9     VMail.Plan      0.081622650 2      0      Weak
## 10    Intl.Calls     0.068633851 2      0      Weak
## 11    Night.Charge   0.060709508 6      0      Weak
## 12    Account.Length 0.033794008 4      0      Weak
## 13    Day.Calls      0.028673937 3      0      Weak
## 14    Night.Mins     0.022784889 2      0      Weak
## 15    Eve.Calls       0.010611328 2      0  Very weak
## 16    Night.Calls    0.009978104 2      0  Very weak
## 17    Int.l.Plan      0.000000000 1      0  Very weak

```

```
iv <- iv.mult(mydata_train, y="Churn", summary=FALSE, verbose=TRUE)
```

Plot information value summary



TABALEAU VISUALIZATION

Library(Rserve)

Rserve()

Tableau - churn testing

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Open Help F1
Get Support...
Check for Product Updates...
Watch Training Videos
Sample Workbooks
Sample Gallery
Choose Language
Settings and Performance
Manage Product Keys...
About Tableau

Reset Ignored Messages
Clear Saved Server Sign-ins
Enable Automatic Product Updates
Share Usage Data to Improve Tableau
Enable Autosave
Enable Accelerated Graphics
Manage External Service Connection...
Set Dashboard Web View Security
Start Performance Recording

Sheet 1 (7)

Churn

Day Mins Eve Mins

CustServ Calls Phone

prediction / Churn

Churn

0 1

Area Code

408 415 510

State	Intl Mins	Day Mins	Eve Mins	CustServ Calls	Phone
AK	4.1	2	346-4360		
	4.2	1	373-1028		
	5.3	1	404-1931		
	5.8	3	336-6533		
	6.6	0	394-4548		
	6.9	0	333-3704		
	7.2	0	382-5743		
	7.5	2	359-9454		
	7.8	1	330-5462		
	7.9	4	364-7719		
	8.1	3	365-6756		
	8.2	0	340-3409		
	8.3	3	414-7942		
	8.5	1	334-4506		
	8.6	2	411-4940		
	8.7	0	345-9153		
	8.8	1	333-5295		
	8.9	1	362-8331		
	9.0	1	394-5202		

Data Source Sheet 1 (11) Sheet 1 (13) Sheet 1 (7) Sheet 1 (8) Sheet 1 (6) Sheet 2 (2) Sheet 3 (2) Sheet 4 (2) Sheet 1 Sheet 1 (3) Sheet 1 (4) Sheet 1 (2) Sheet 2 Sheet 3 Sheet 4 Sheet 1 (12) Type here to search ENG 4:14 PM IN 15-Nov-18

Tableau - churn testing

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Show Me

External Service Connection

Select an External Service RServer

Specify a server name and a port

Server: localhost Port: 6311

Sign in with a username and password

Username:

Password:

Require SSL

Test Connection **OK** **Cancel**

Sheet 1 (7)

Churn

Day Mins Eve Mins

CustServ Calls Phone

prediction / Churn

Churn

0 1

Area Code

408 415 510

State	Intl Mins	Day Mins	Eve Mins	CustServ Calls	Phone
AK	4.1	2	346-4360		
	4.2	1	373-1028		
	5.3	1	404-1931		
	5.8	3	336-6533		
	6.6	0	394-4548		
	6.9	0	333-3704		
	7.2	0	382-5743		
	7.5	2	359-9454		
	7.8	1	330-5462		
	7.9	4	364-7719		
	8.1	3	365-6756		
	8.2	0	340-3409		
	8.3	3	414-7942		
	8.5	1	334-4506		
	8.6	2	411-4940		
	8.7	0	345-9153		
	8.8	1	333-5295		
	8.9	1	362-8331		
	9.0	1	394-5202		

Data Source Sheet 1 (11) Sheet 1 (13) Sheet 1 (7) Sheet 1 (8) Sheet 1 (6) Sheet 2 (2) Sheet 3 (2) Sheet 4 (2) Sheet 1 Sheet 1 (3) Sheet 1 (4) Sheet 1 (2) Sheet 2 Sheet 3 Sheet 4 Sheet 1 (12) Type here to search ENG 4:14 PM IN 15-Nov-18

Tableau - churn testing

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Data Analytics

Pages

Columns prediction Churn Day Mins Eve Mins

Rows State Intl Mins CustServ Calls Phone

Filters Churn

Marks All Area Color Size Label Detail Tooltip

Dimensions Area Code Day Calls Day Charge Day Mins ABC Phone State ABC Measure Names

Measures ABC Calculation2 Churn CustServ Calls Eve Calls Eve Charge Eve Mins Intl Plan Intl Calls Intl Charge Intl Mins Intl Calls Night Calls Night Charge Night Mins SUM(Eve Mi... ABC prediction

External Service Connection

Select an External Service RServe

Tableau

Successfully connected to the External Service.

OK

Test Connection OK Cancel

Churn

0 1

Area Code

408 415 510

State

Intl Mins

CustServ Calls

Phone

Null 0

prediction / Churn

AK 41

7.8 7.9

8.1 3 414-7942

8.5 1 334-4506

411-4940

8.6 2 345-9153

8.7 0 333-5295

1 362-8331

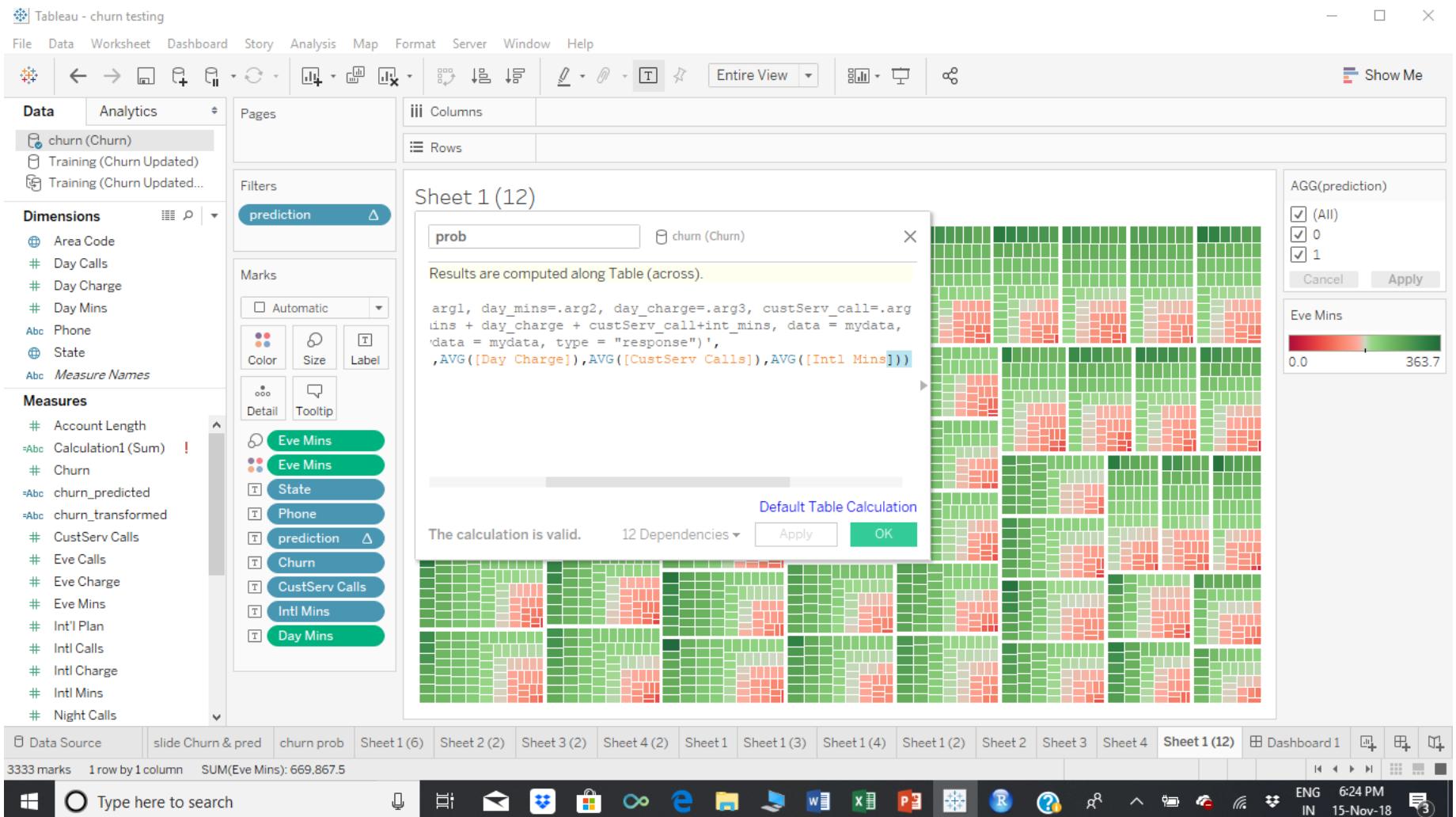
8.9 1 394-5202

Data Source Sheet1(11) Sheet1(13) Sheet1(7) Sheet1(8) Sheet1(6) Sheet2(2) Sheet3(2) Sheet4(2) Sheet1 Sheet1(3) Sheet1(4) Sheet1(2) Sheet2 Sheet3 Sheet4 Sheet1(12)

Type here to search

ENG 4:21 PM IN 15-Nov-18

The screenshot shows the Tableau desktop interface with a data source named 'churn testing'. The 'Data' tab is selected. A 'Columns' shelf contains 'prediction', 'Churn', 'Day Mins', and 'Eve Mins'. A 'Rows' shelf contains 'State', 'Intl Mins', 'CustServ Calls', and 'Phone'. A 'Filters' shelf contains 'Churn' with the value 'Churn'. The 'Marks' shelf is set to 'All' with 'Area' selected. The 'Dimensions' shelf lists 'Area Code', 'Day Calls', 'Day Charge', 'Day Mins', 'ABC Phone', 'State', and 'ABC Measure Names'. The 'Measures' shelf lists 'ABC Calculation2', 'Churn', 'CustServ Calls', 'Eve Calls', 'Eve Charge', 'Eve Mins', 'Intl Plan', 'Intl Calls', 'Intl Charge', 'Intl Mins', 'Intl Calls', 'Night Calls', 'Night Charge', 'Night Mins', 'SUM(Eve Mi...)', and 'ABC prediction'. A 'Show Me' button is visible at the top right. A modal dialog titled 'External Service Connection' is open, showing 'Select an External Service' dropdown with 'RServe' selected. Below it, another modal titled 'Tableau' shows a success message: 'Successfully connected to the External Service.' with an 'OK' button. The main workspace shows a preview of the data with columns 'State', 'Intl Mins', 'CustServ Calls', 'Phone', and 'Null' (containing '0'). A color legend for 'Area Code' shows 408 (blue), 415 (orange), and 510 (red). The preview also includes rows of data with columns '7.8', '7.9', '8.1', '8.5', '8.6', '8.7', and '8.9'. The bottom navigation bar shows tabs for 'Data Source' and various sheets. The taskbar at the bottom includes icons for File Explorer, Mail, Microsoft Edge, and Google Chrome, along with system status indicators like battery level and network signal.



Calculated field (prob)

```
SCRIPT_REAL('library(dplyr);
```

```
mydata <- data.frame(churn=.arg1, day_mins=.arg2, day_charge=.arg3, custServ_call=.arg4, int_mins=.arg5);
```

```

lrmrmodel <- glm(churn ~ day_mins + day_charge + custServ_call+int_mins, data = mydata, family = "binomial");

prob <- predict(lrmrmodel, newdata = mydata, type = "response"),

AVG([Churn]),AVG([Day Mins]),AVG([Day Charge]),AVG([CustServ Calls]),AVG([Intl Mins]))

```

Tableau - churn testing

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Data Analytics

Pages Columns Rows

Filters prediction

Dimensions

- Area Code
- Day Calls
- Day Charge
- Day Mins
- Phone
- State
- Measure Names

Measures

- Account Length
- Calculation1(Sum) !
- Churn
- churn_predicted
- churn_transformed
- CustServ Calls
- Eve Calls
- Eve Charge
- Eve Mins
- Intl Plan
- Intl Calls
- Intl Charge
- Intl Mins
- Night Calls

prediction

churn (Churn)

Results are computed along Table (across).
STR(SCRIPT_REAL('library(checkmate')

churn <- .arg1
Day_Mins <- .arg2
intl_Mins <- .arg3
set.seed(2017)
data_churn <- data.frame(churn, Day_Mins, intl_Mins)

Default Table Calculation

The calculation is valid. 16 Dependencies Apply OK

AGG(prediction)

- (All)
- 0
- 1

Eve Mins

0.0 363.0

Cancel Apply

Sheet 1 (12)

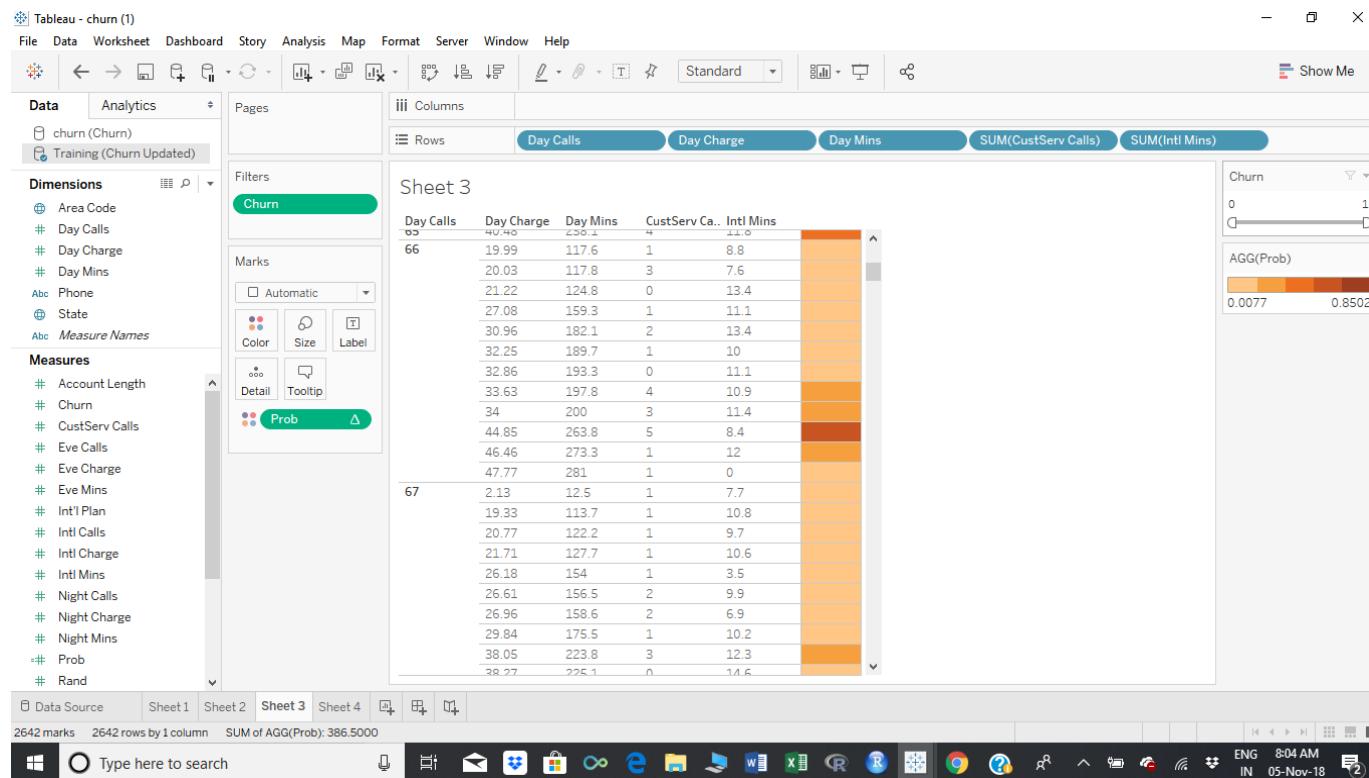
3333 marks 1 row by 1 column SUM(Eve Mins): 669,867.5

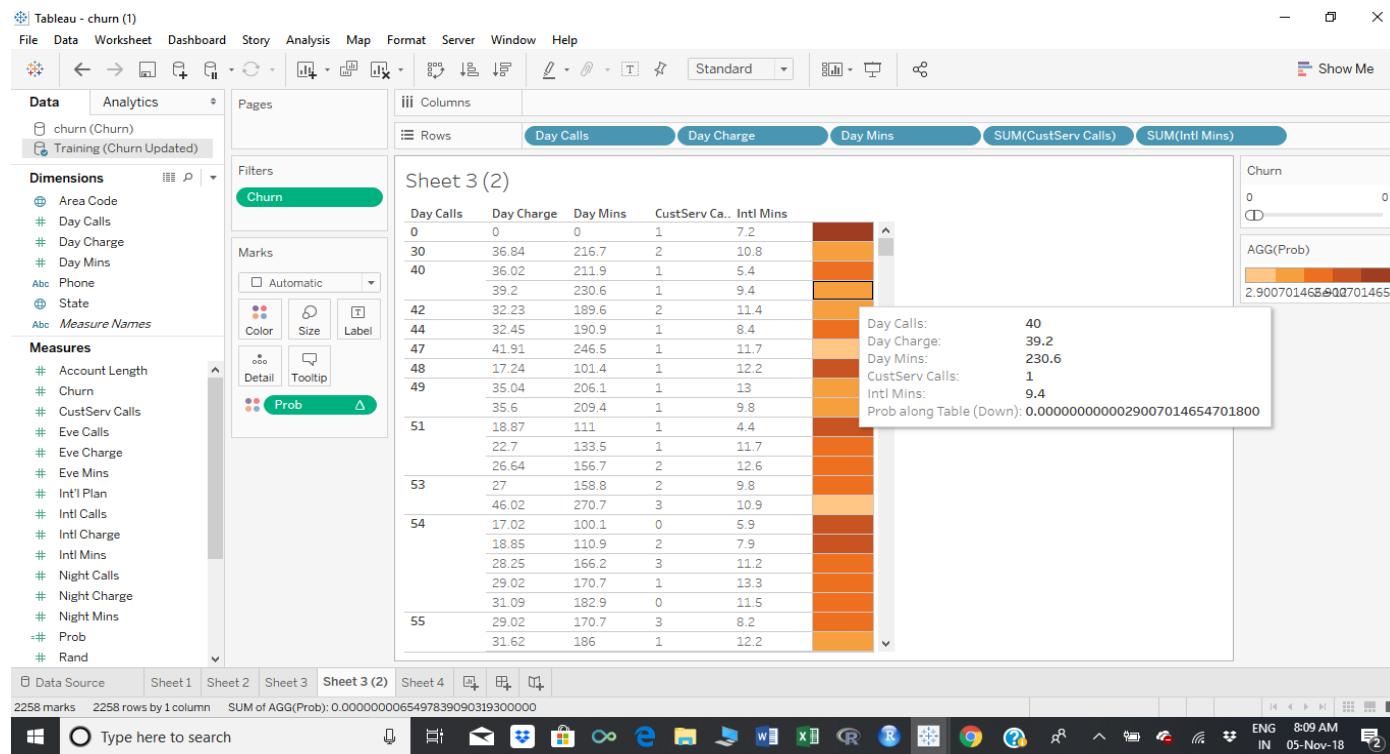
Type here to search

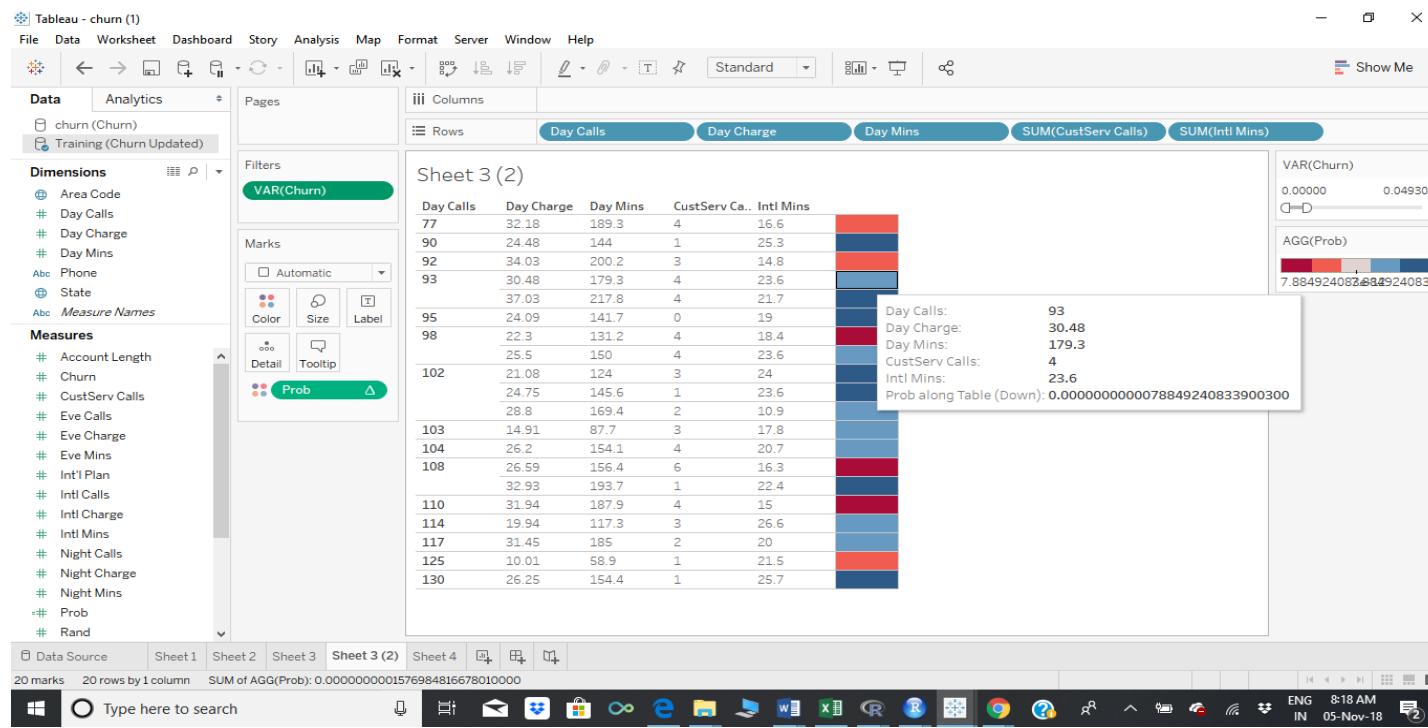
ENG 6:29 PM IN 15-Nov-18

STR(SCRIPT_REAL('library(checkmate')

```
churn <- .arg1
Day_Mins <- .arg2
intl_Mins <- .arg3
set.seed(2017)
data_churn <- data.frame(churn, Day_Mins, intl_Mins)
intrain<-sample(1:nrow(data_churn),.7*nrow(data_churn))
training<- data_churn[intrain,]
testing<- data_churn[-intrain,]
new_data<-rbind(training,testing)
LogModel <- glm(churn ~ .,family=binomial(link="logit"),data=training)
fitted.results <- predict(LogModel,newdata=new_data,type="response")
pred_val <- ifelse(fitted.results > 0.5,1,0)
pred_val
',ATTR([Churn]),SUM([Day Mins]),sum([Intl Mins])))
```







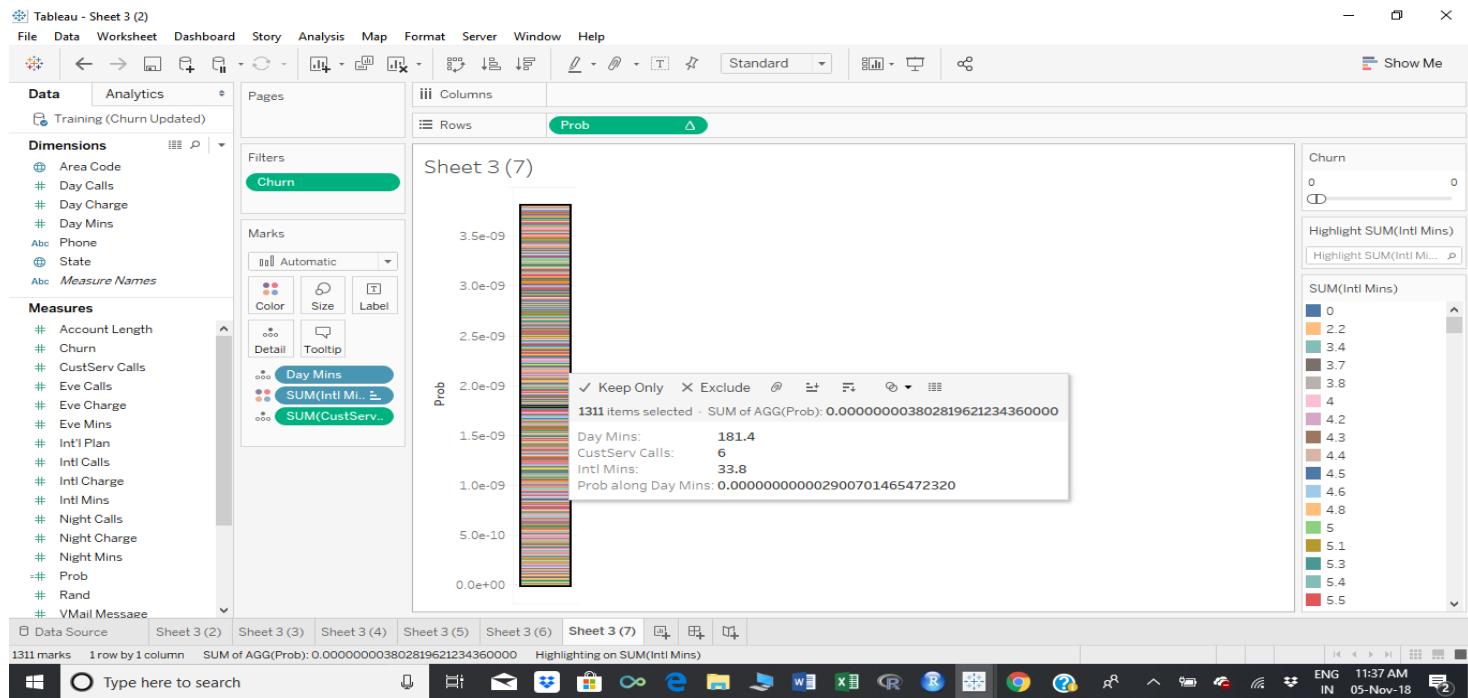
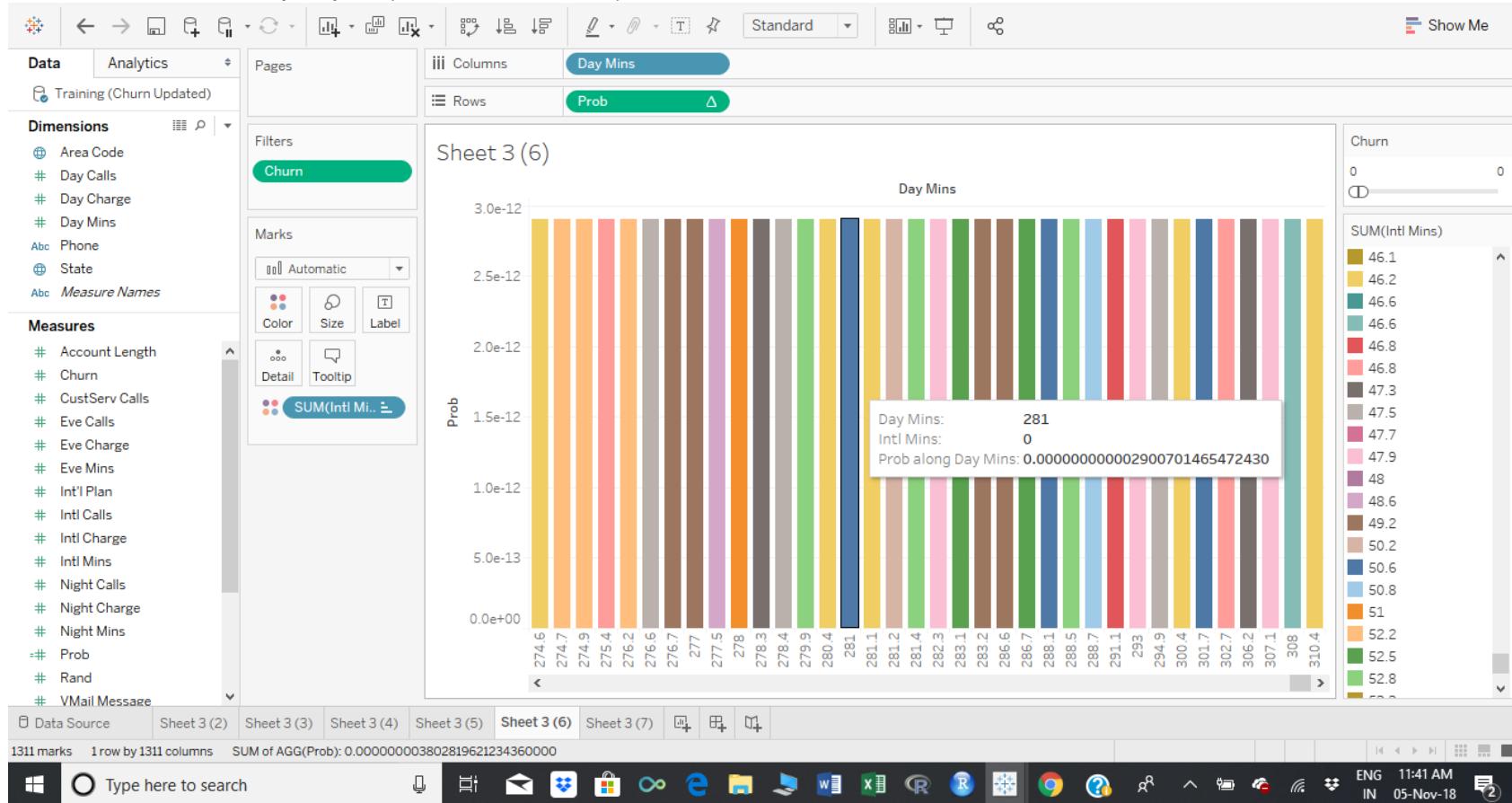


Tableau - Sheet 3 (2)

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help



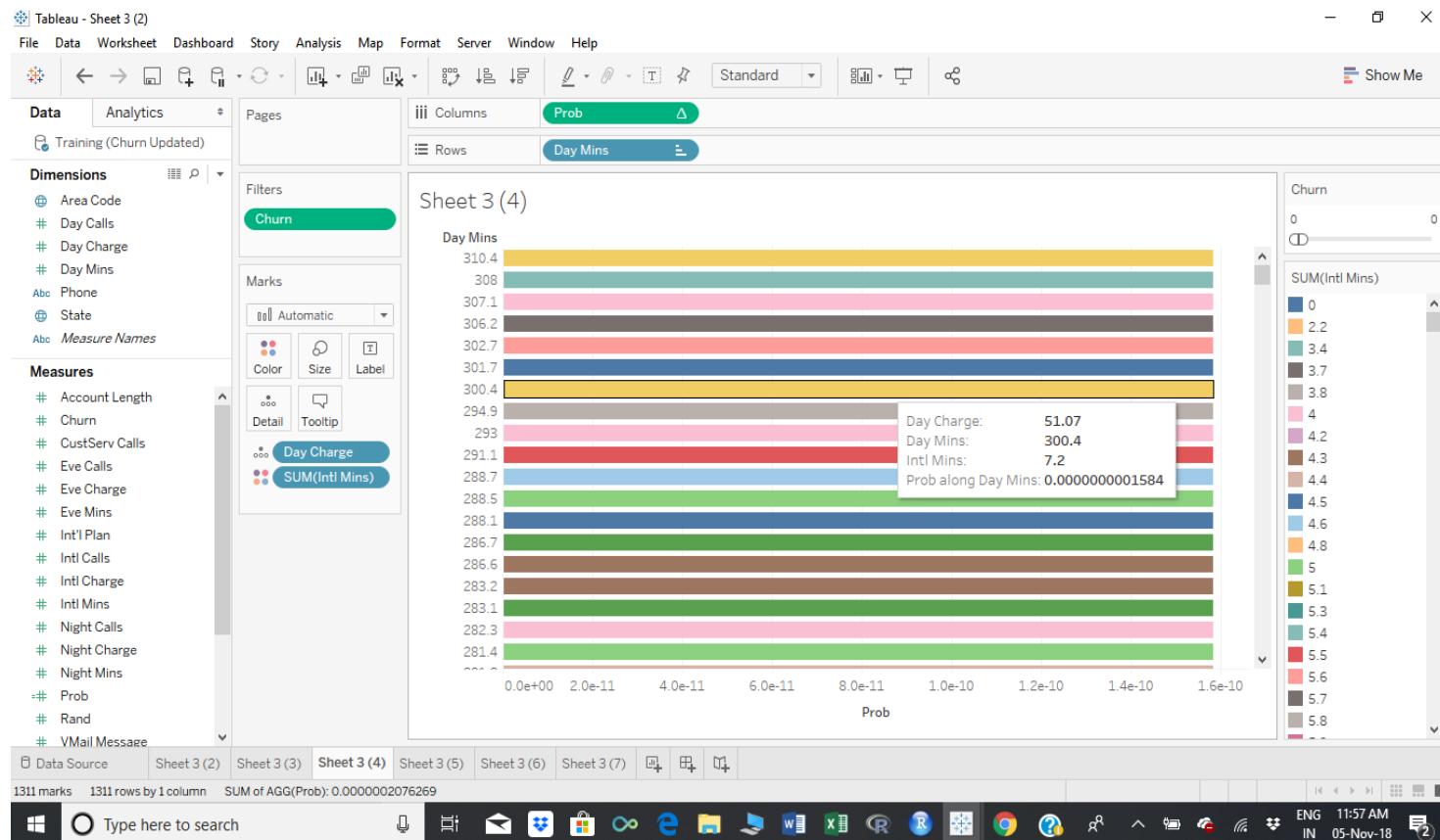


Tableau - Sheet 3 (2)

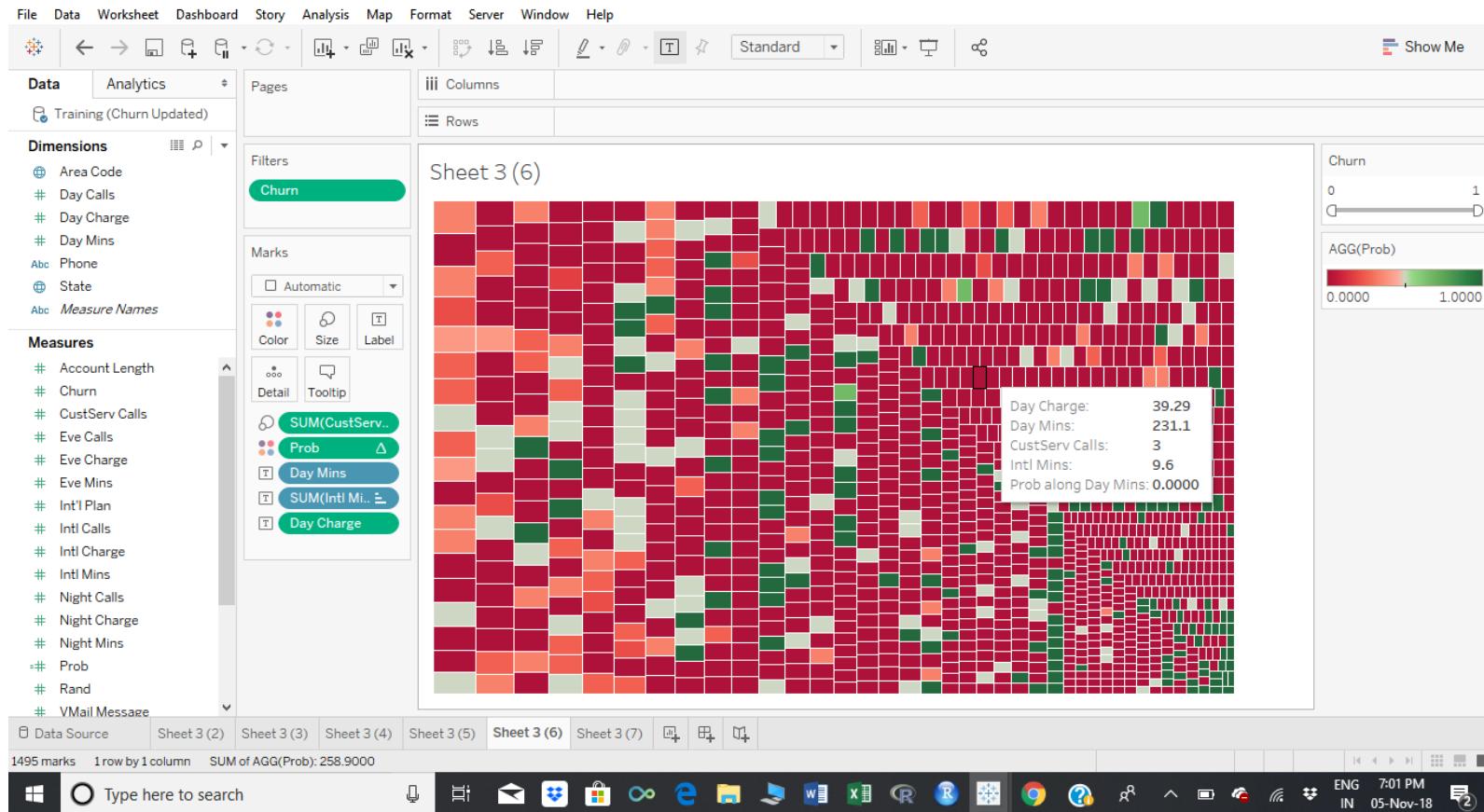
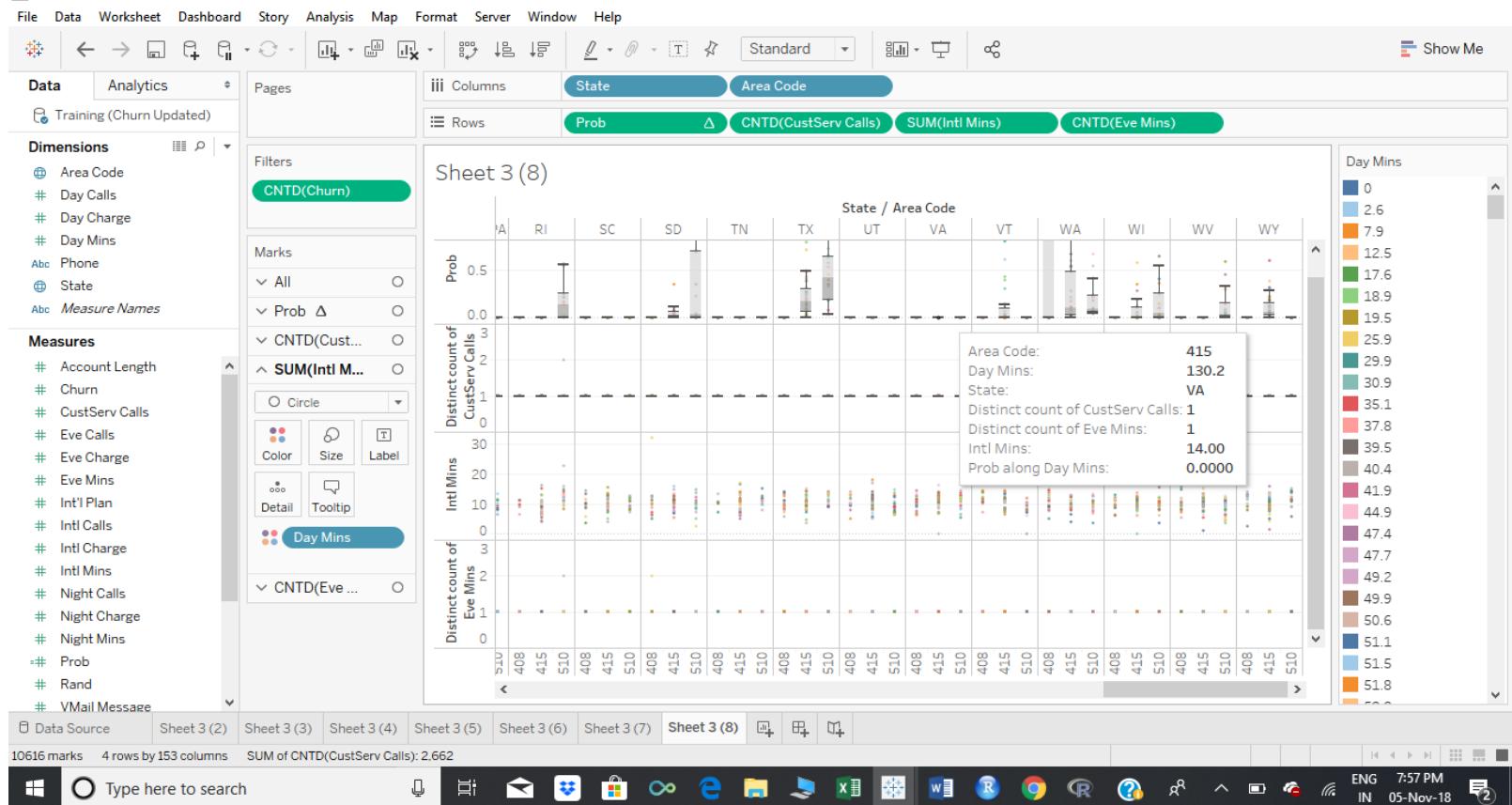
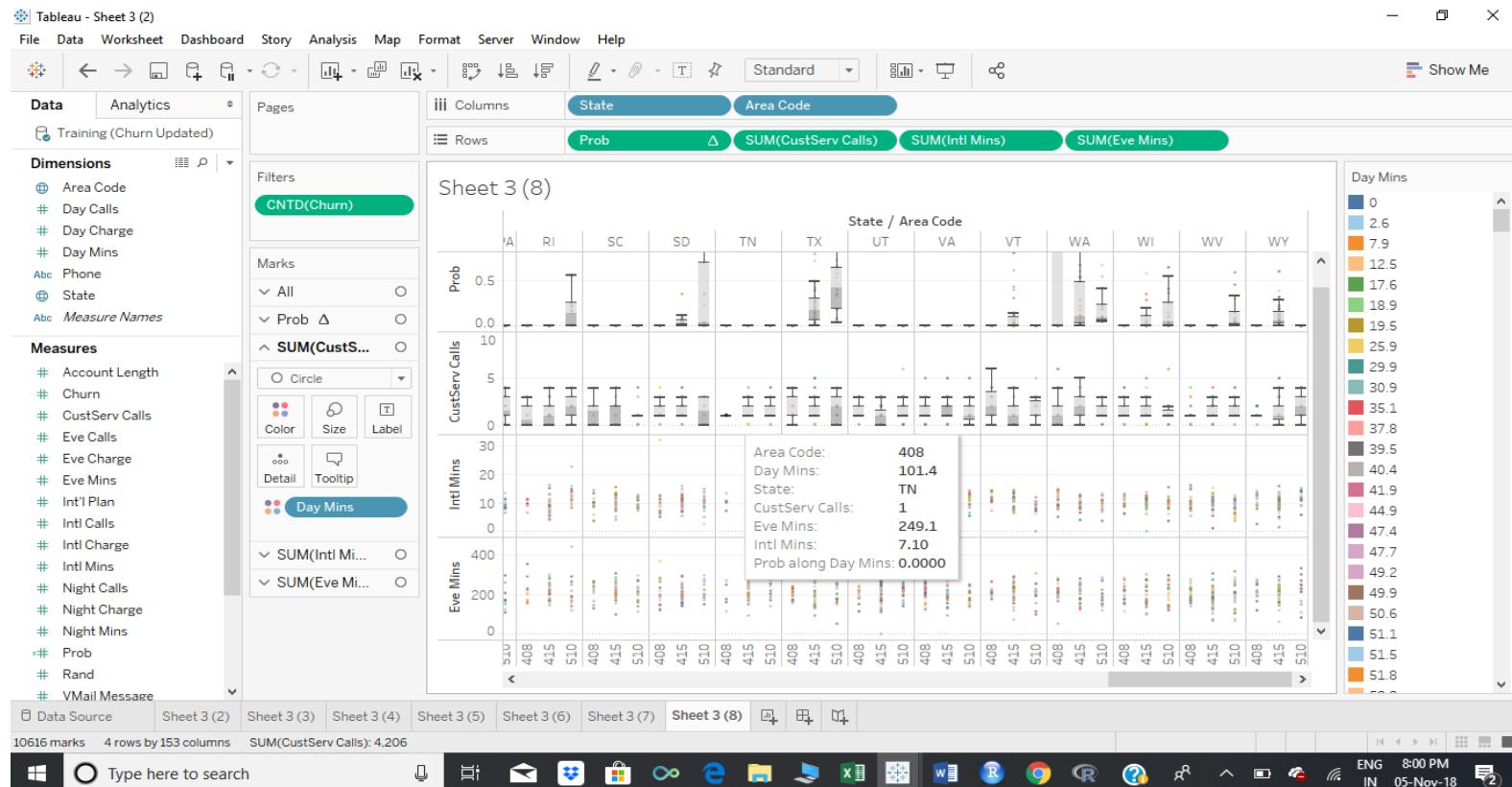
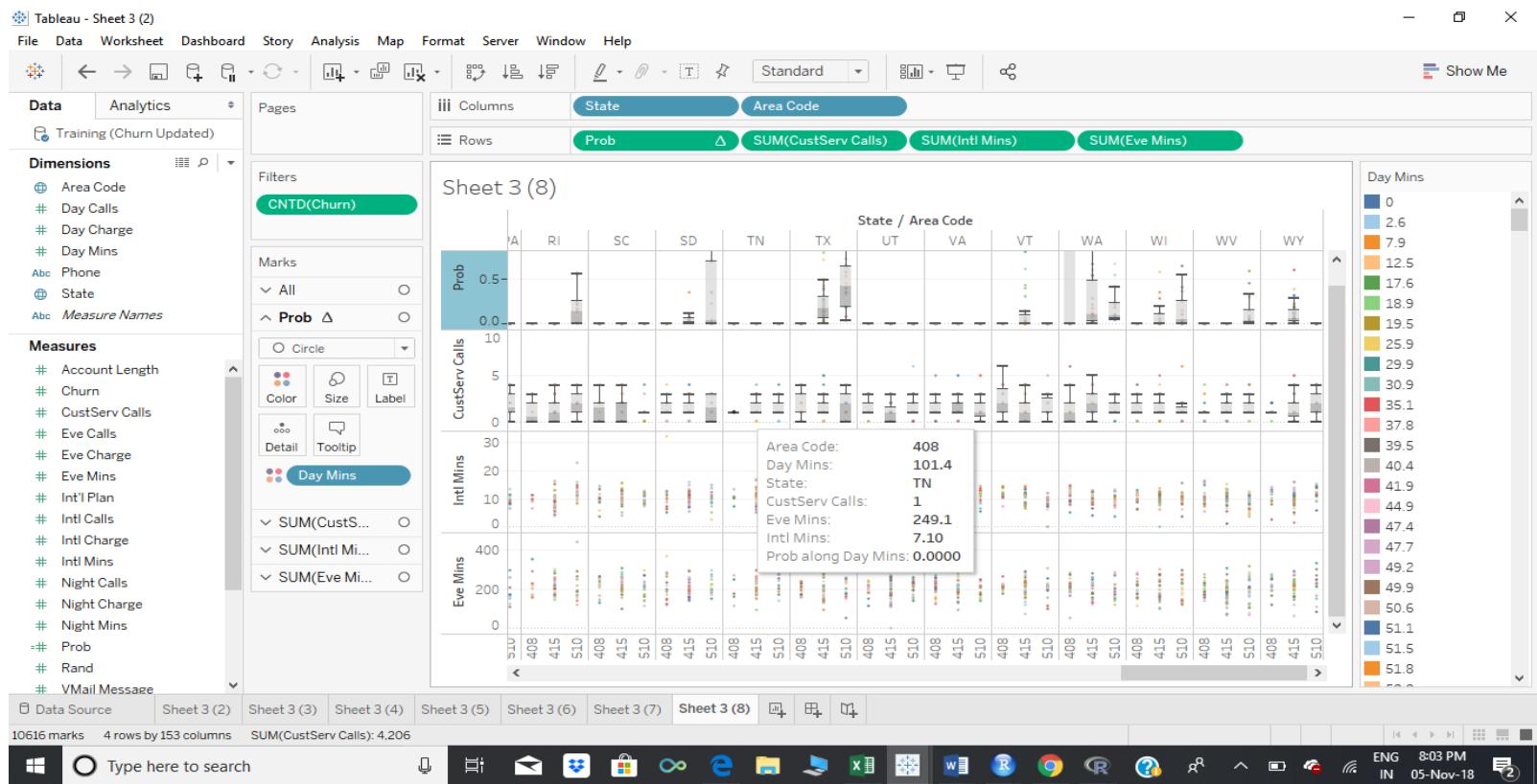
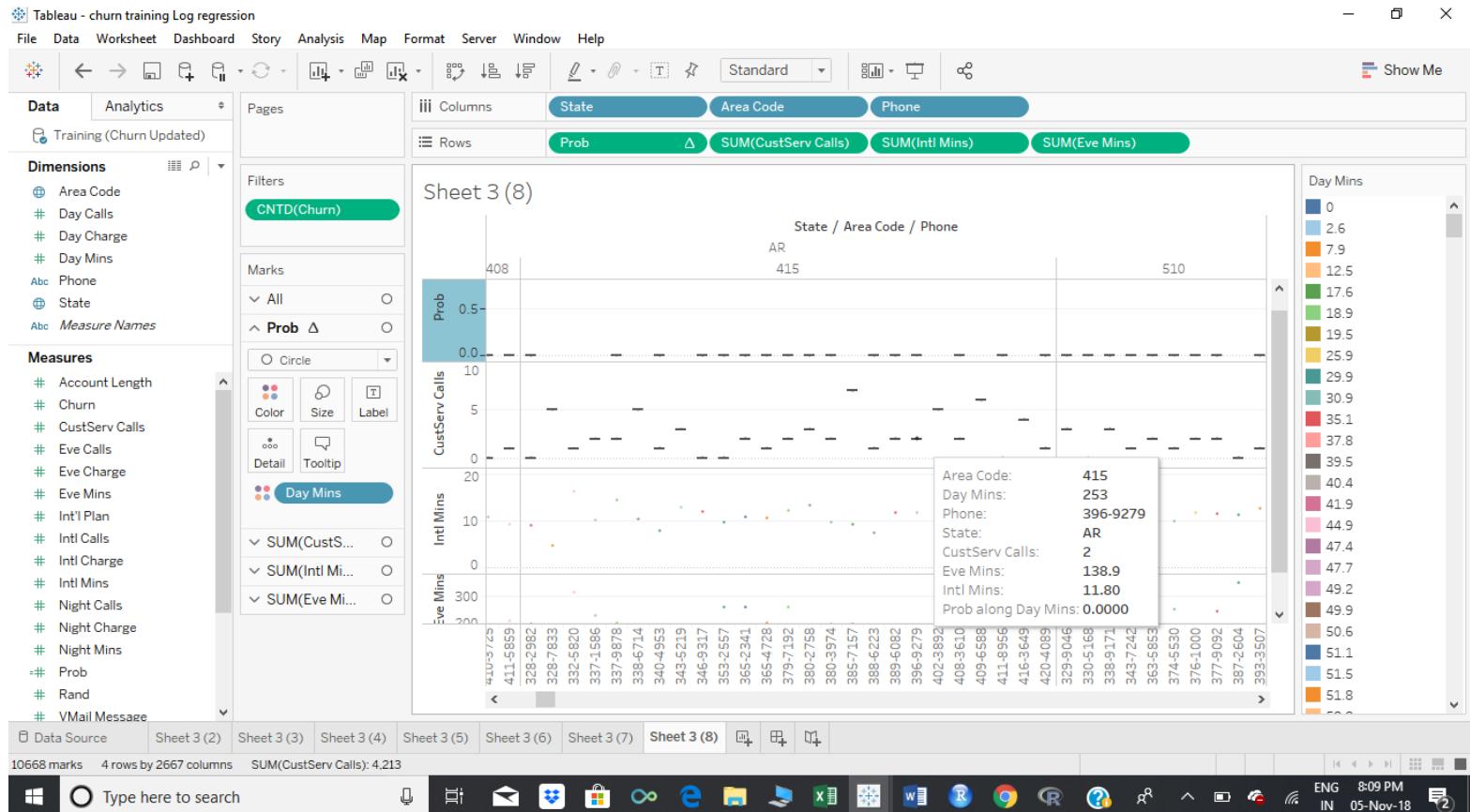


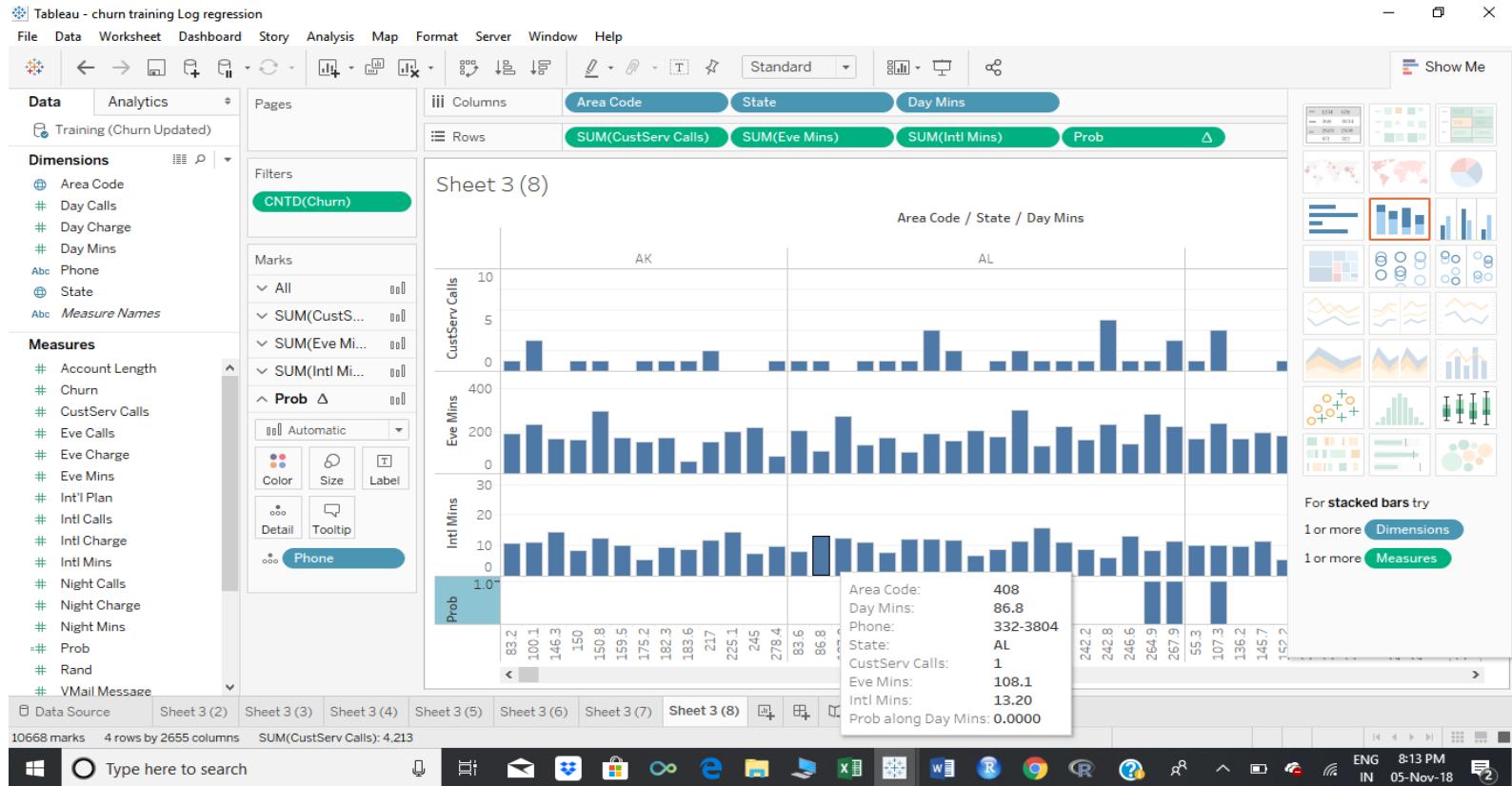
Tableau - Sheet 3 (2)

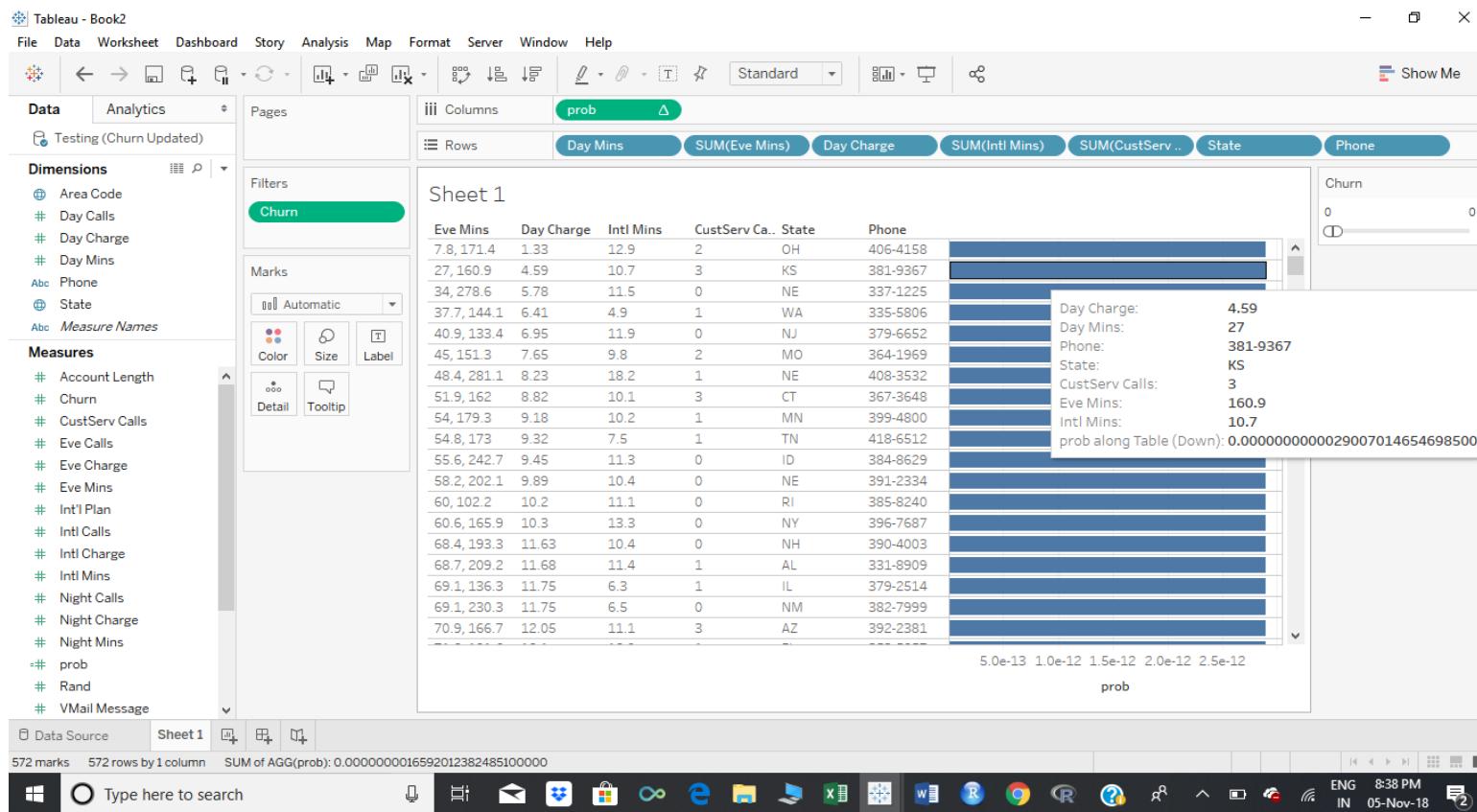


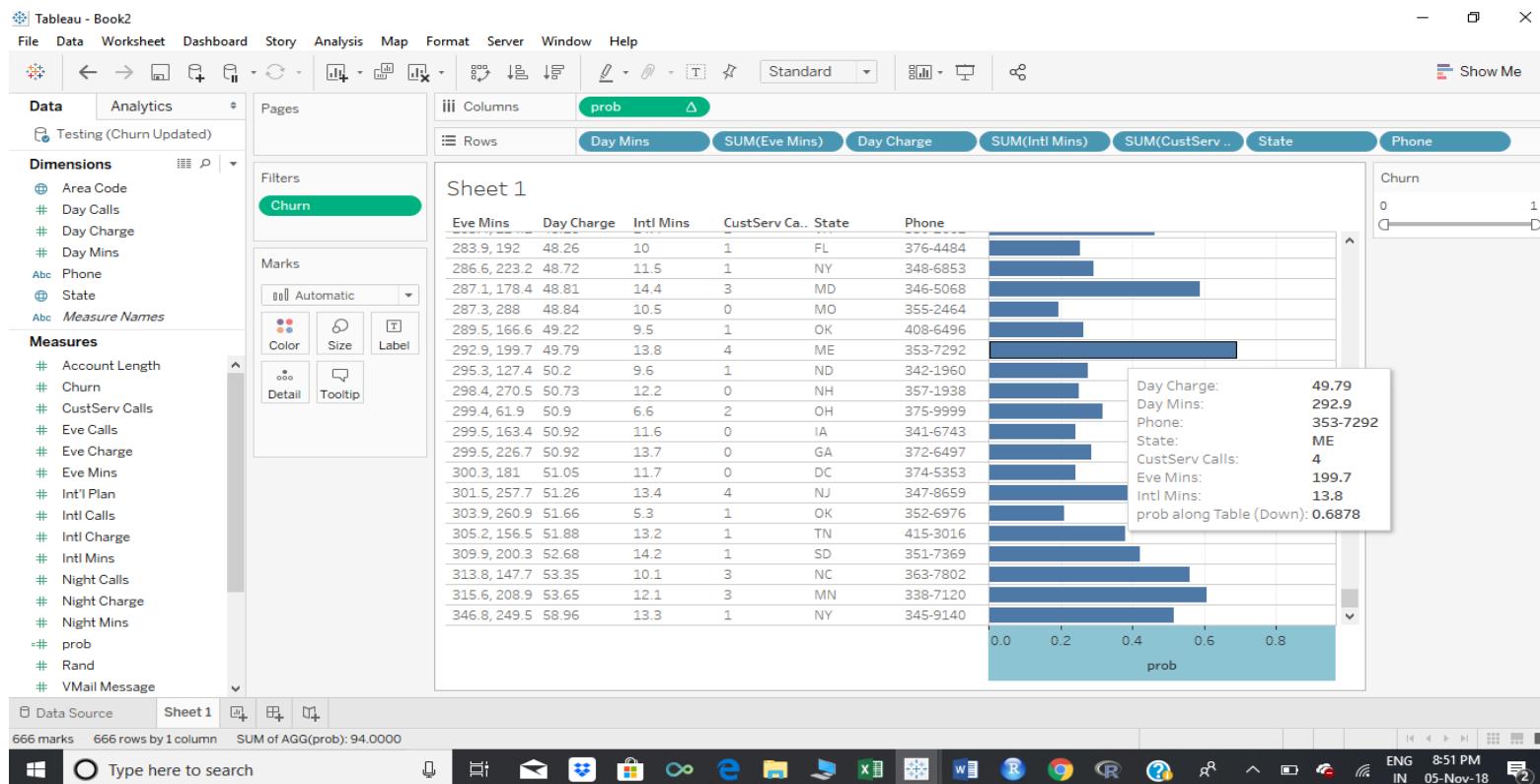


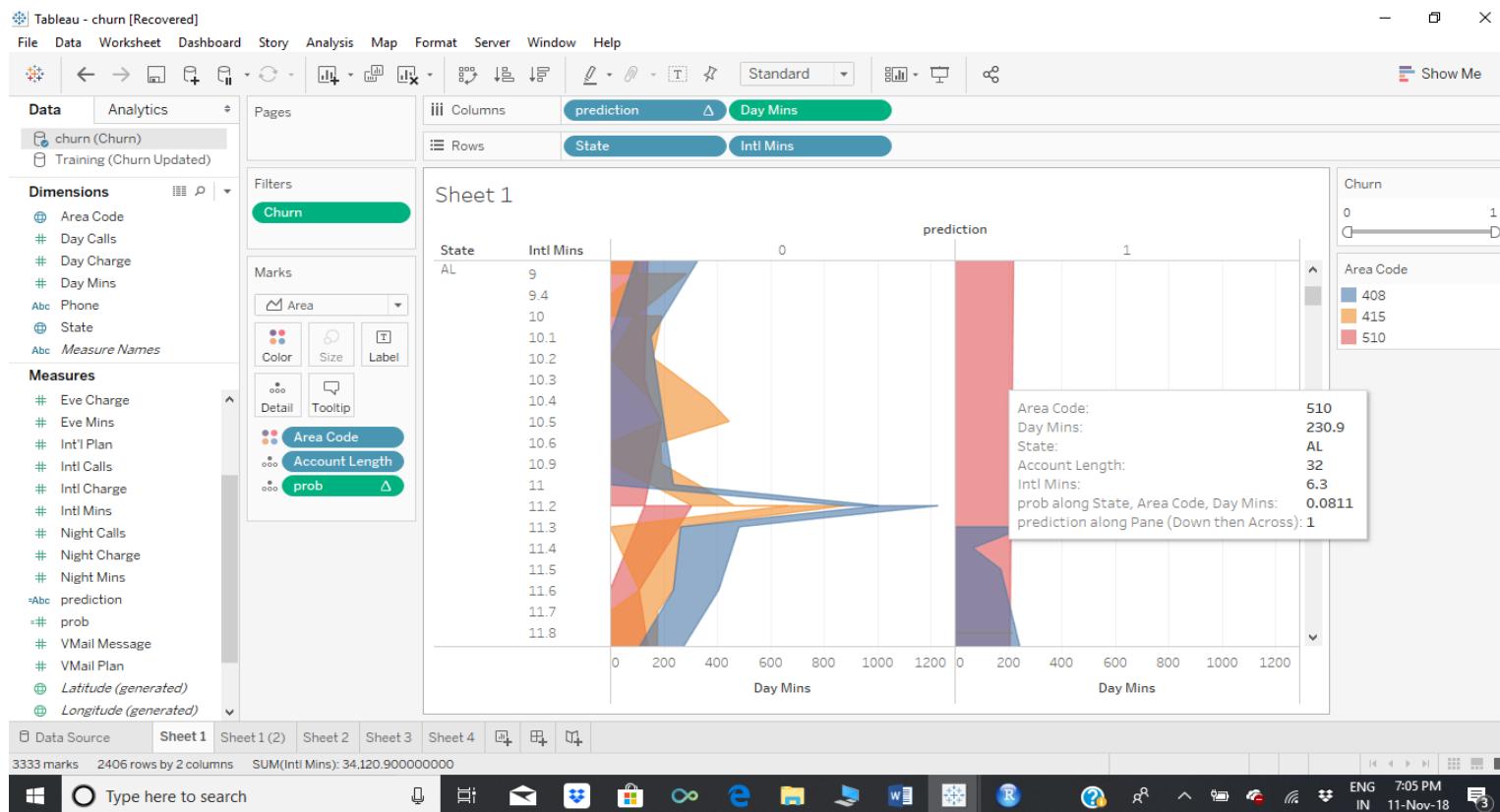


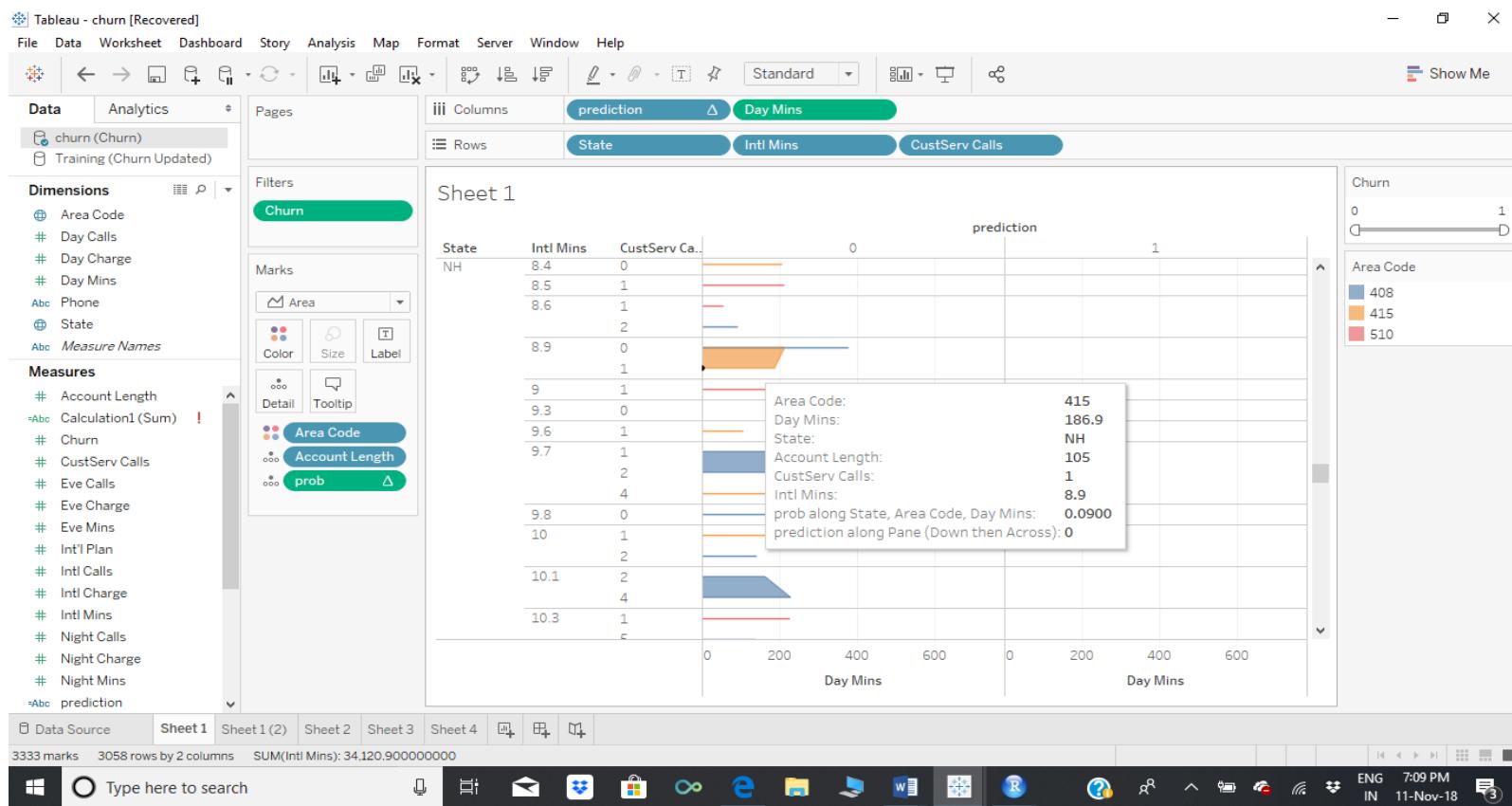


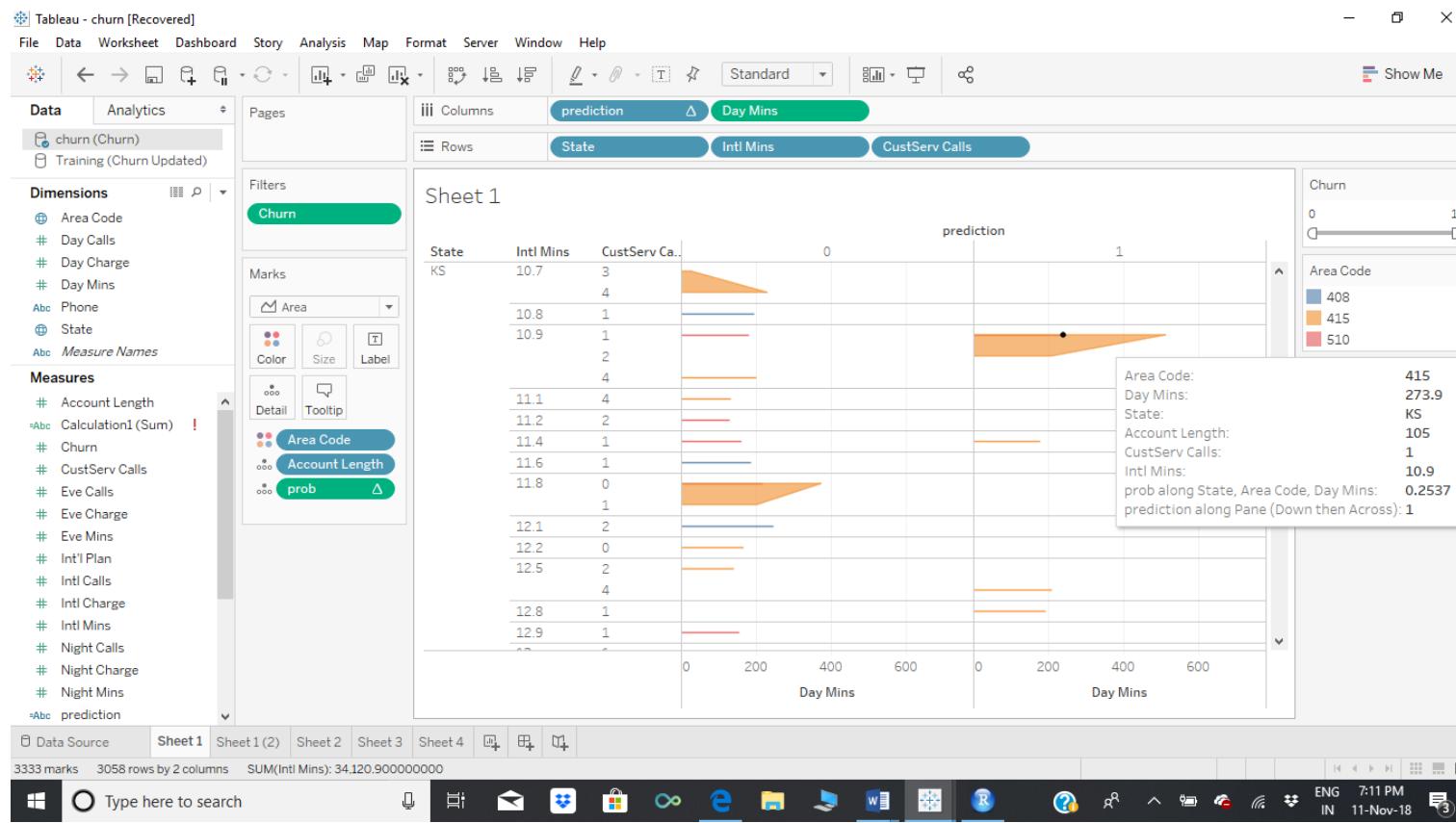


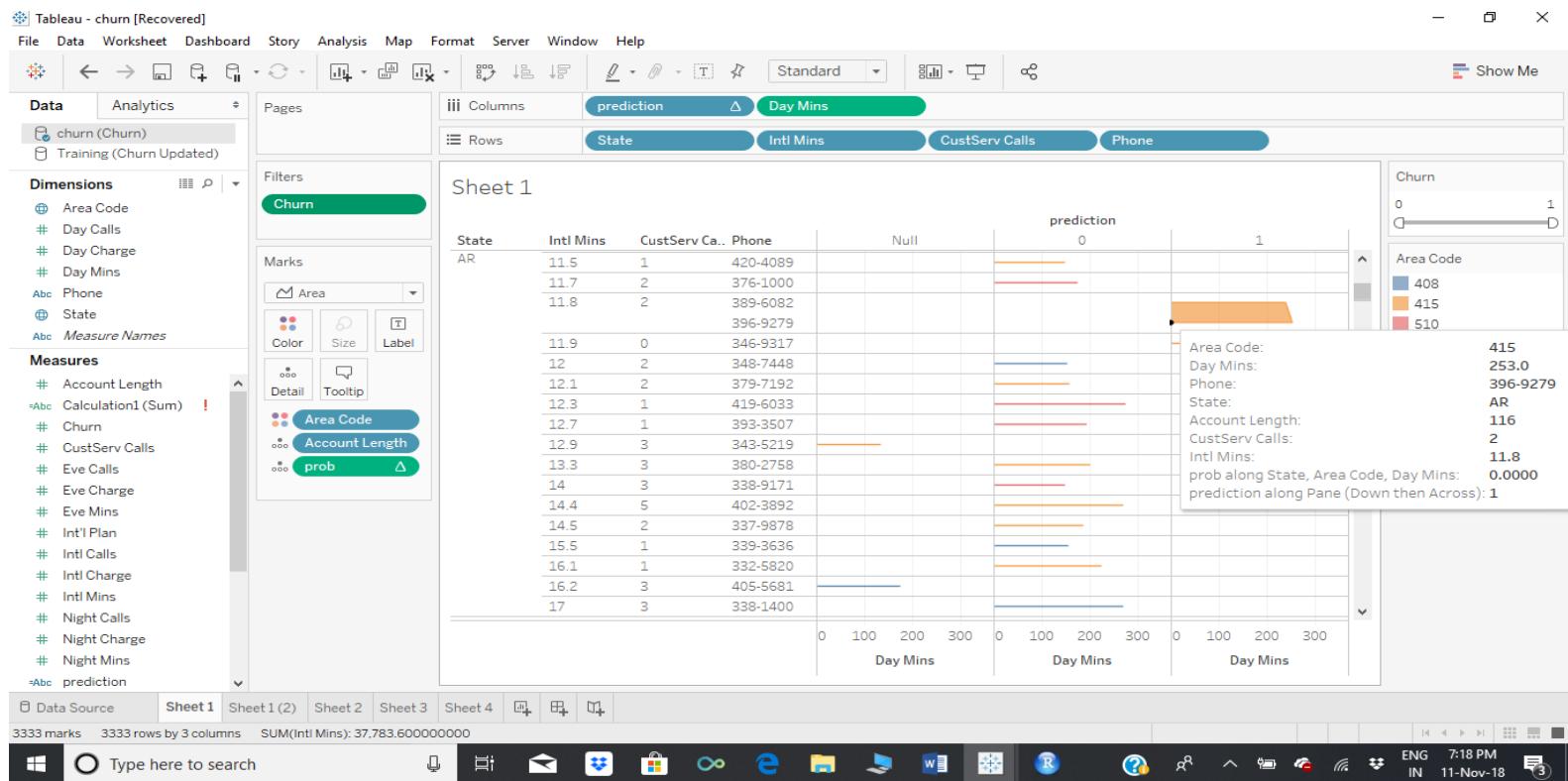


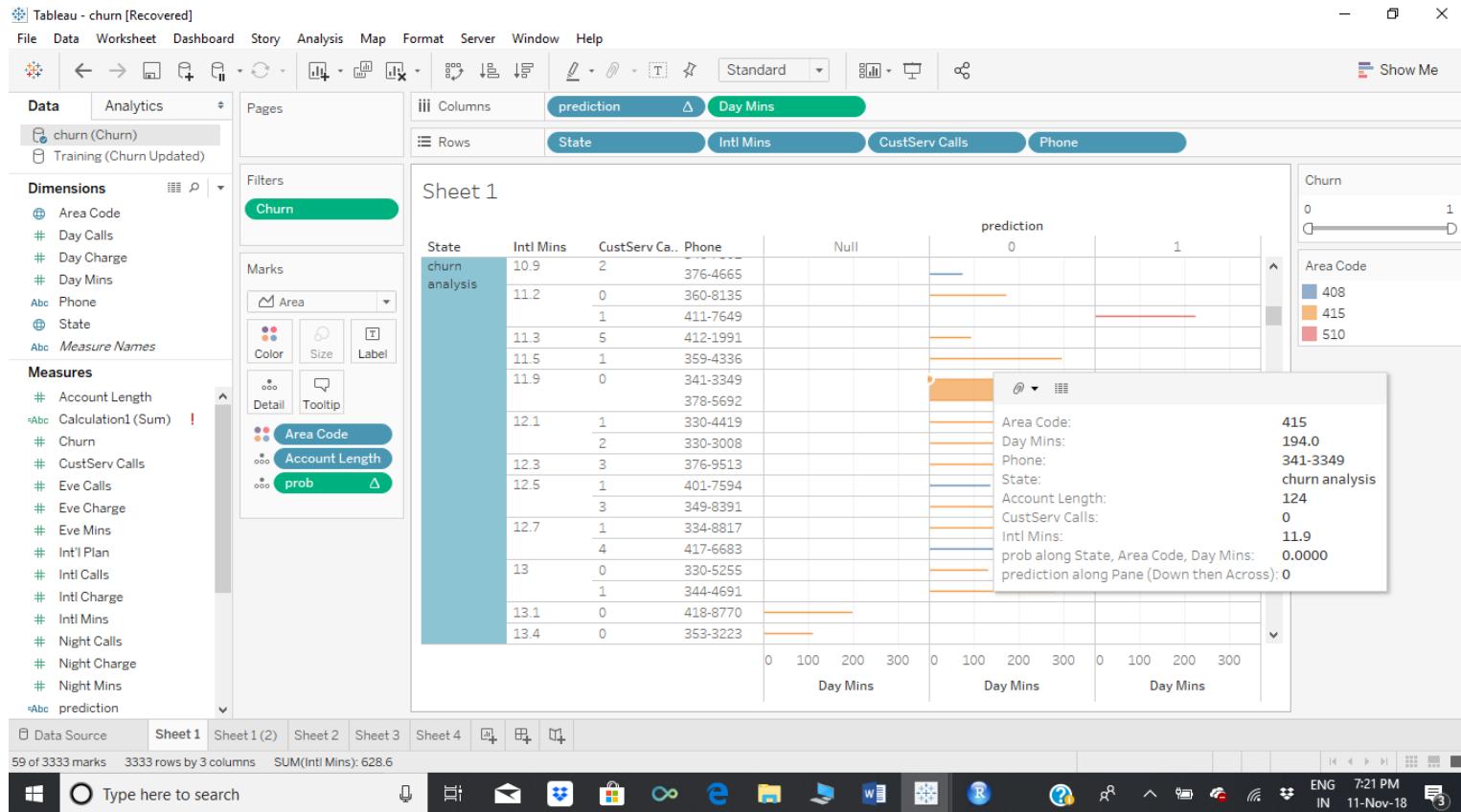












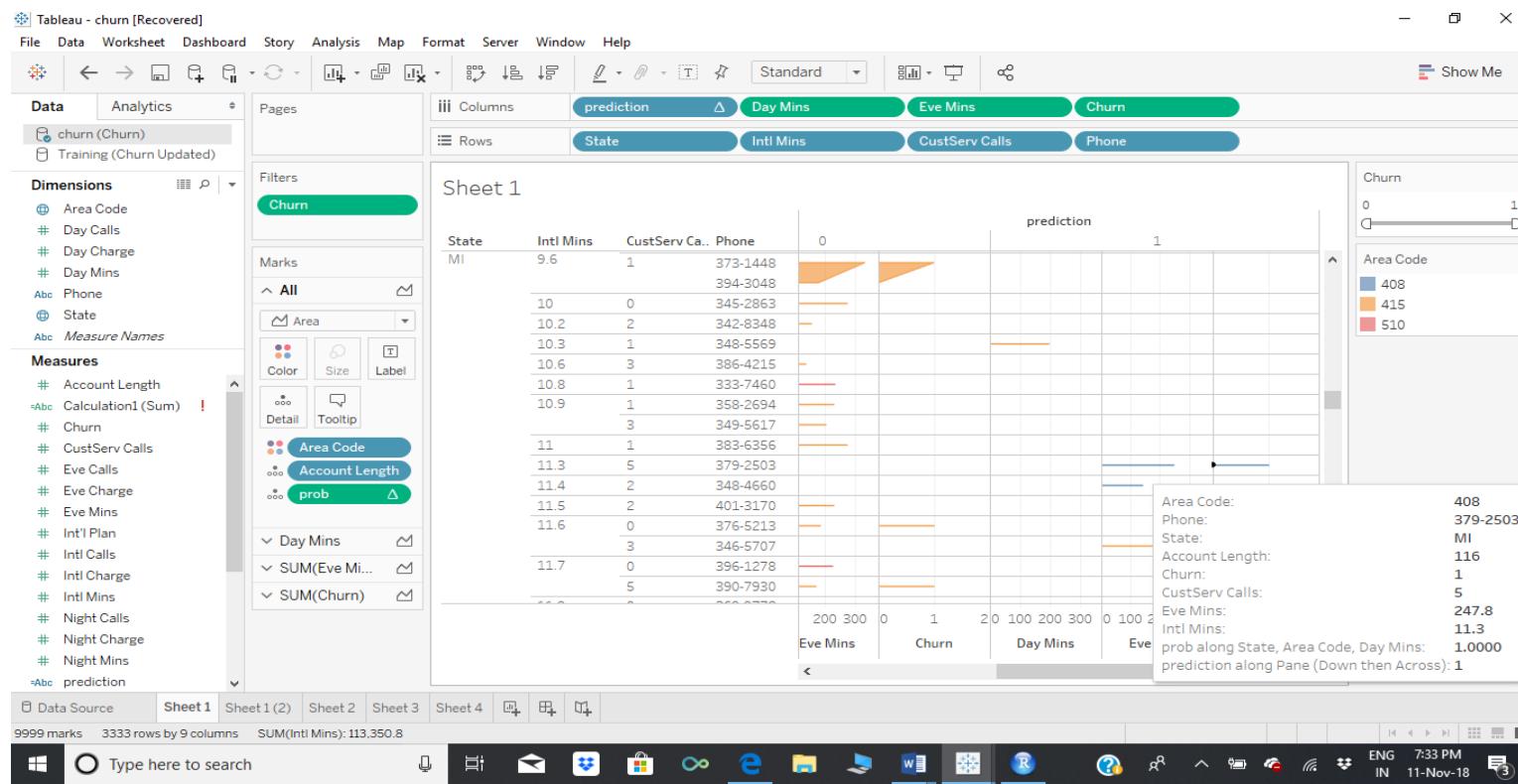


Tableau - churn [Recovered]

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Data Analytics Pages

churn (Churn) Training (Churn Updated)

Dimensions

- Area Code
- Day Calls
- Day Charge
- Day Mins
- Phone
- State
- Measure Names

Measures

- Account Length
- Calculation1 (Sum) !
- Churn
- CustServ Calls
- Eve Calls
- Eve Charge
- Eve Mins
- Int'l Plan
- Intl Calls
- Intl Charge
- Intl Mins
- Night Calls
- Night Charge
- Night Mins
- prediction

prediction chun (Churn)

Results are computed along Table (across).
 intl_Mins <- .arg3
 set.seed(2017)
 data_churn <- data.frame(churn, Day_Mins, intl_Mins)
 intrain<-sample(1:nrow(data_churn), .7*nrow(data_churn))
 training<- data_churn[intrain,]
 testing<- data_churn[-intrain,]

Sheet 1

State	Intl Mins	CustServ Ca..	Phone
MI	9.6	1	373-1448
			394-3048
	10	0	345-2863
	10.2	2	342-8348
	10.3	1	348-5569
	10.6	3	386-4215
	10.8	1	333-7460
	10.9	1	358-2694
		3	349-5617
	11	1	383-6356
	11.3	5	379-2503
	11.4	2	348-4660
	11.5	2	401-3170
	11.6	0	376-5213
		3	346-5707
	11.7	0	396-1278
		5	390-7930
			200 300 0 1 20 100 200 300 0 100 200 300 0 1
			Eve Mins Churn Day Mins Eve Mins Churn

The calculation is valid. 2 Dependencies ▾

Show Me

Churn

Area Code

408 415 510

Default Table Calculation

OK

Sheet 1 Sheet 1 (2) Sheet 2 Sheet 3 Sheet 4

9999 marks 3333 rows by 9 columns SUM(Intl Mins): 113.350.8

Type here to search ENG 7:34 PM IN 11-Nov-18

Tableau - churn [Recovered]

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Data Analytics

churn (Churn) Training (Churn Updated)

Dimensions

- Area Code
- Day Calls
- Day Charge
- Day Mins
- Phone
- State
- Measure Names

Measures

- Account Length
- Calculation1 (Sum) !
- Churn
- CustServ Calls
- Eve Calls
- Eve Charge
- Eve Mins
- Int'l Plan
- Intl Calls
- Intl Charge
- Intl Mins
- Night Calls
- Night Charge
- Night Mins
- prediction

Pages

iii Columns prediction Day M

Rows State Int'l M

Sheet 1

Results are computed along Table (across).

```
day_mins=.arg2, day_charge=.arg3, custServ_call=.arg4, int_
day_charge + custServ_call+int_mins, data = mydata, family
mydata, type = "response"',
Day Charge],AVG([CustServ Calls]),AVG([Intl Mins]))
```

The calculation is valid. 2 Dependencies ▾ Apply OK

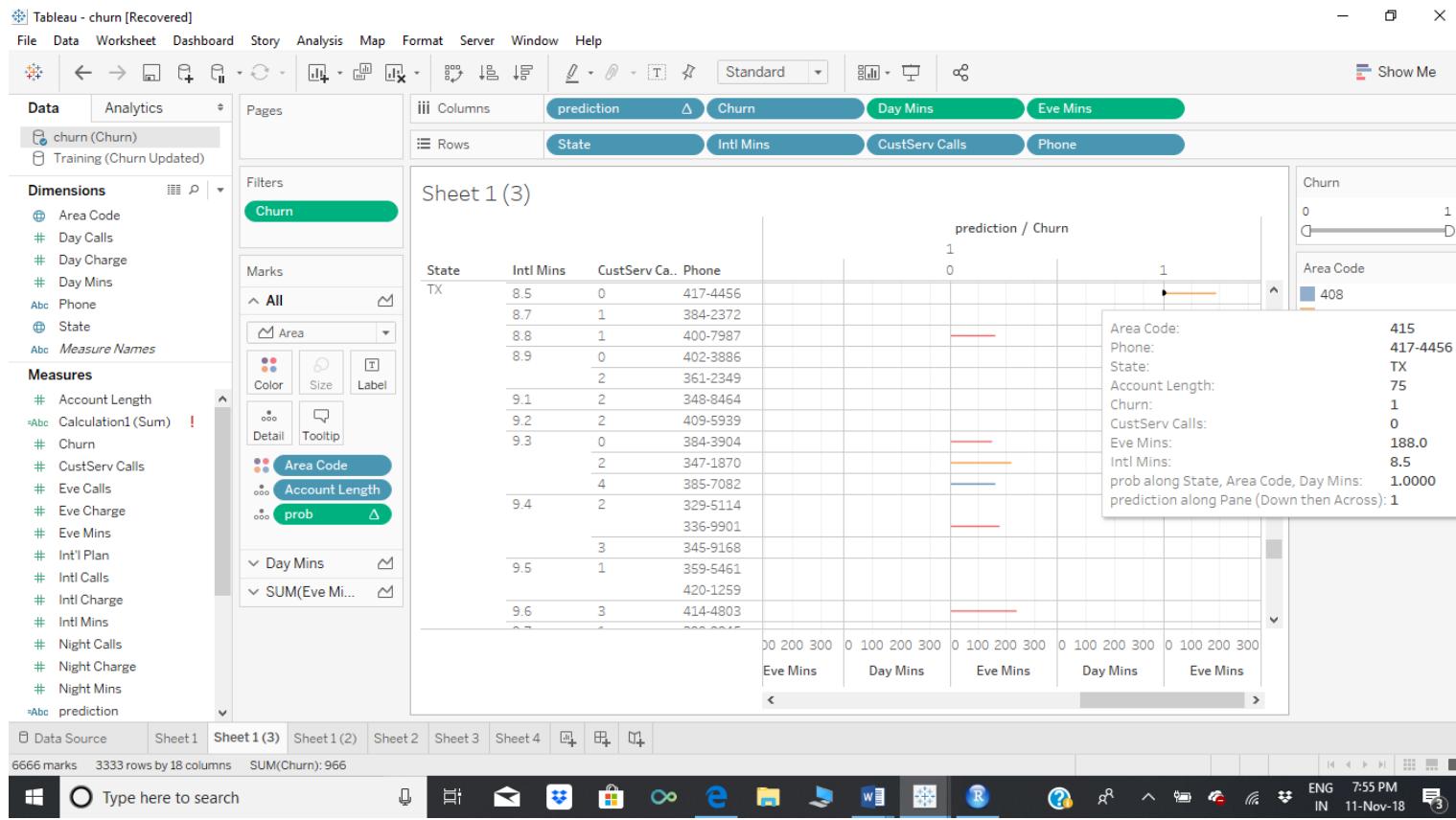
Default Table Calculation

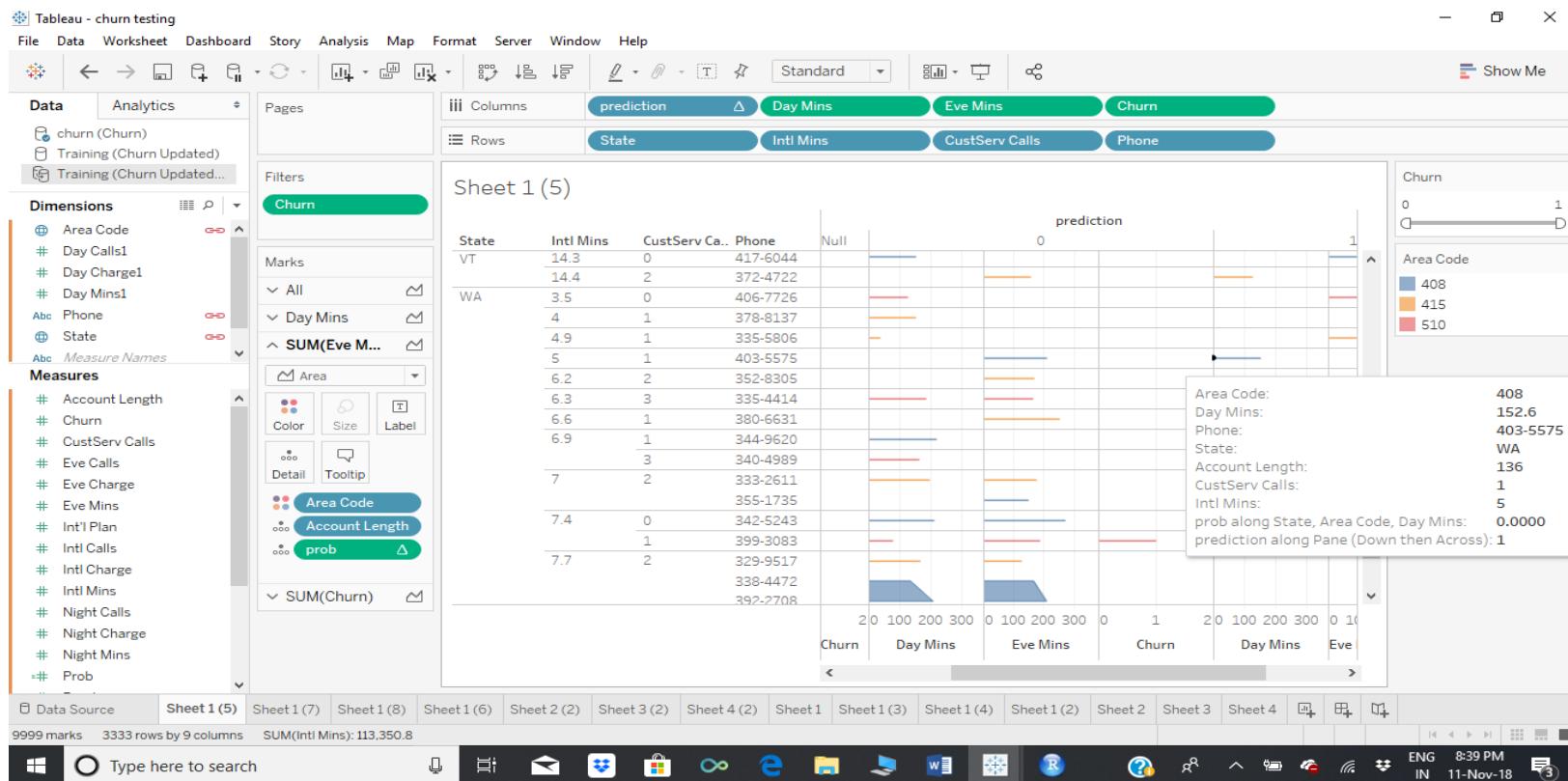
State	Intl Mins	CustServ Ca.. Phone
MI	9.6	373-1448 394-3048
	10	345-2863
	10.2	342-9348
	10.3	348-5569
	10.6	386-4215
	10.8	333-7460
	10.9	358-2694 3 349-5617
	11	383-6356
	11.3	379-2503
	11.4	348-4660
	11.5	401-3170
	11.6	376-5213 3 346-5707
	11.7	396-1278 5 390-7930

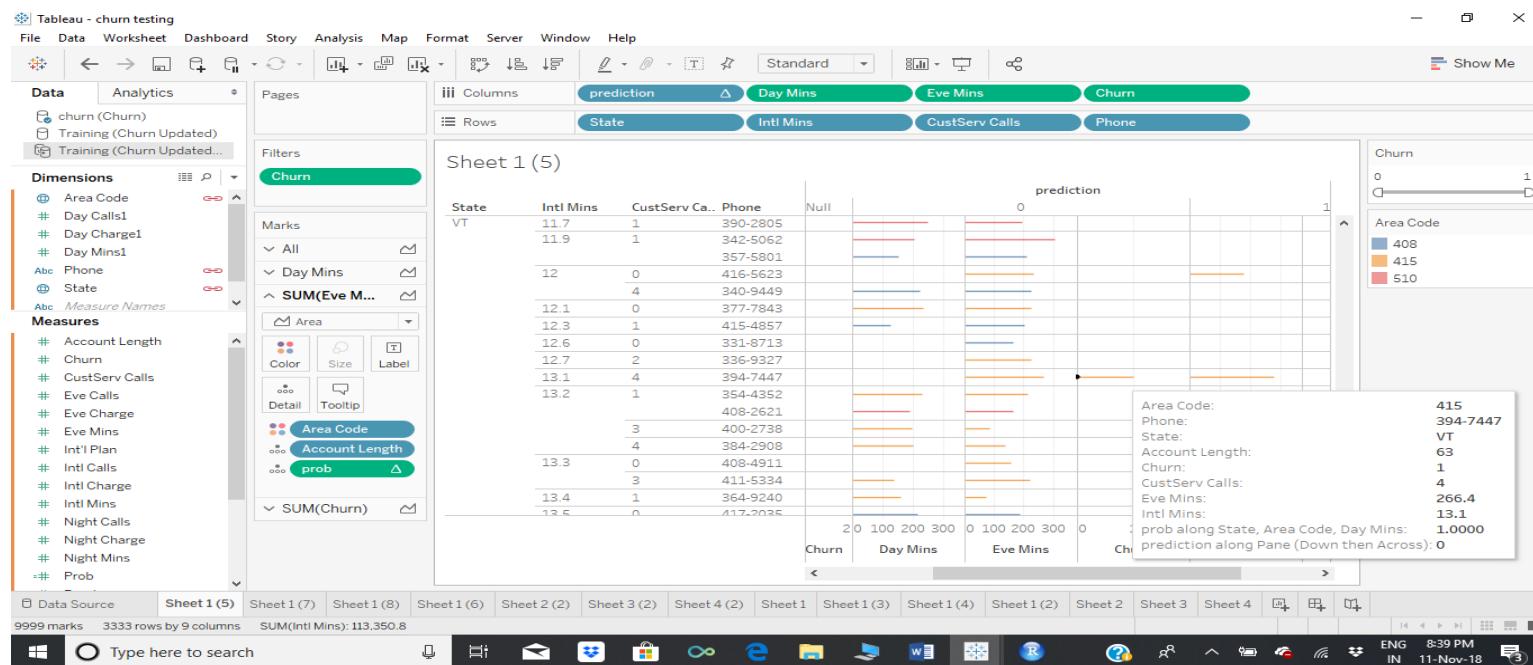
Eve Mins Churn Day Mins Eve Mins Churn

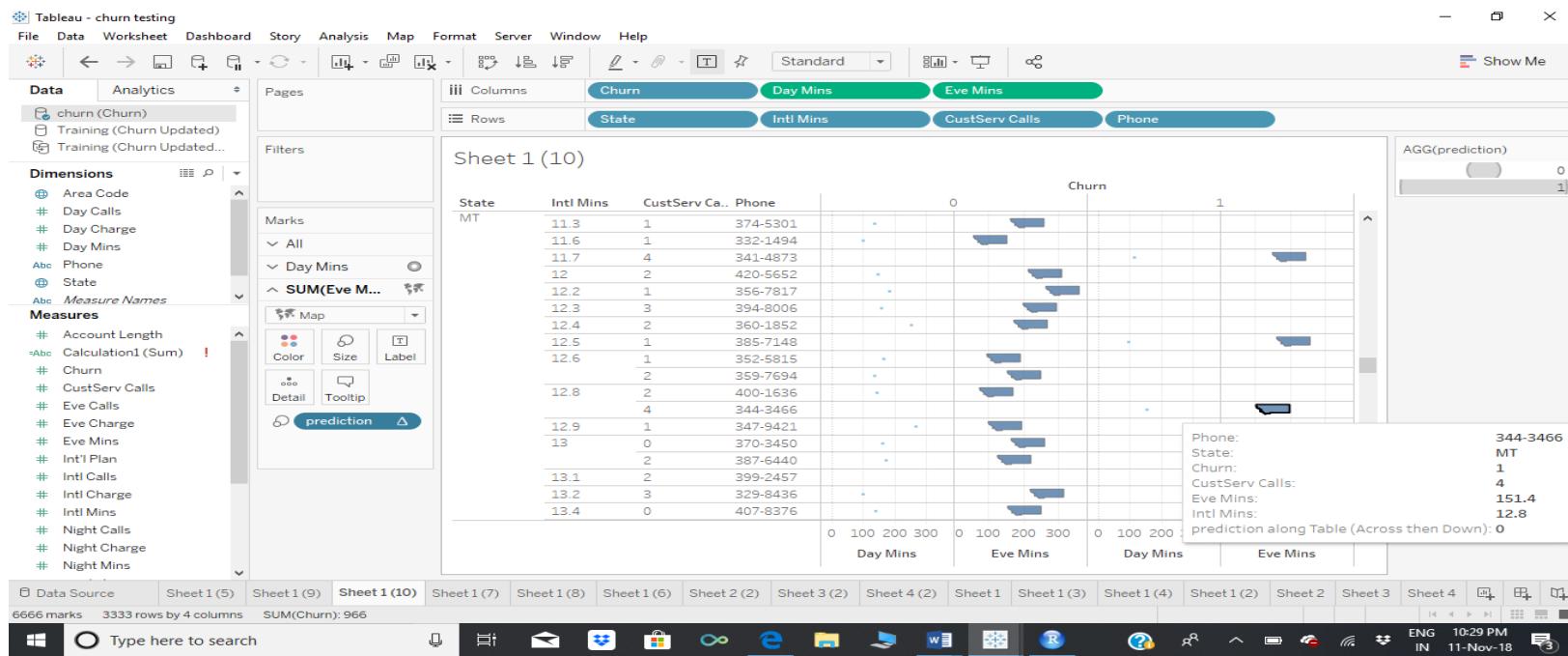
9999 marks 3333 rows by 9 columns SUM(Int'l Mins): 113,350.8

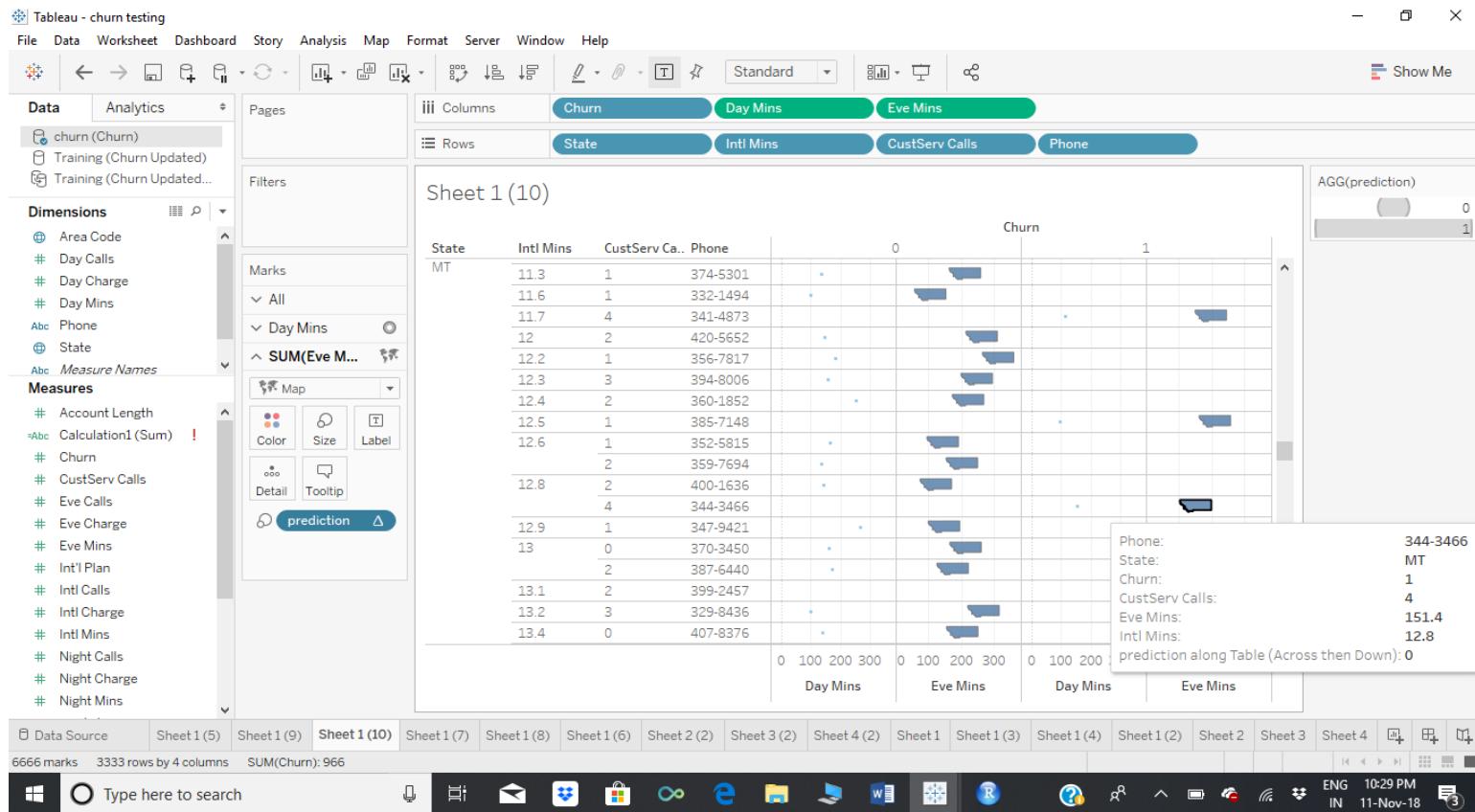
Type here to search ENG 7:35 PM IN 11-Nov-18

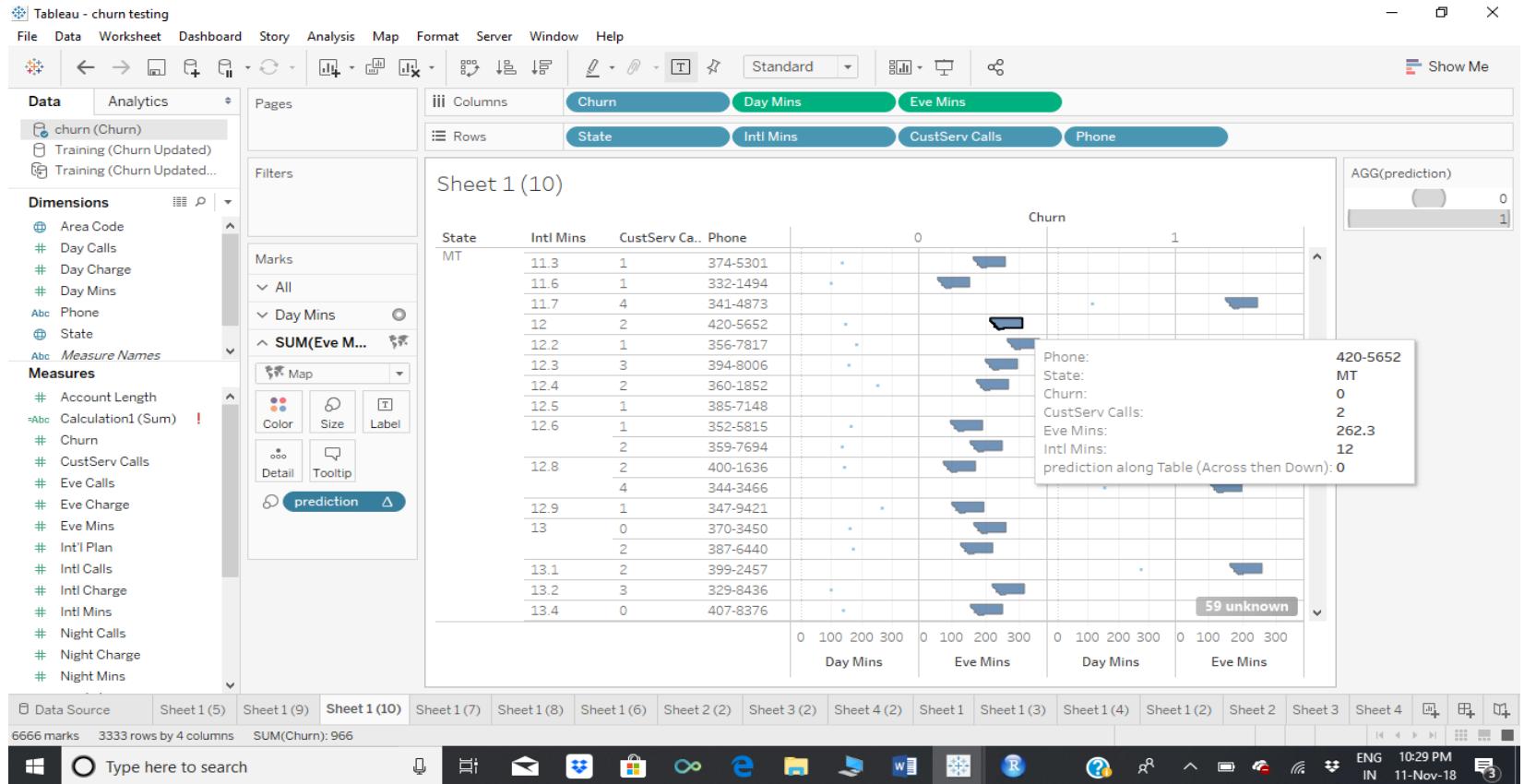


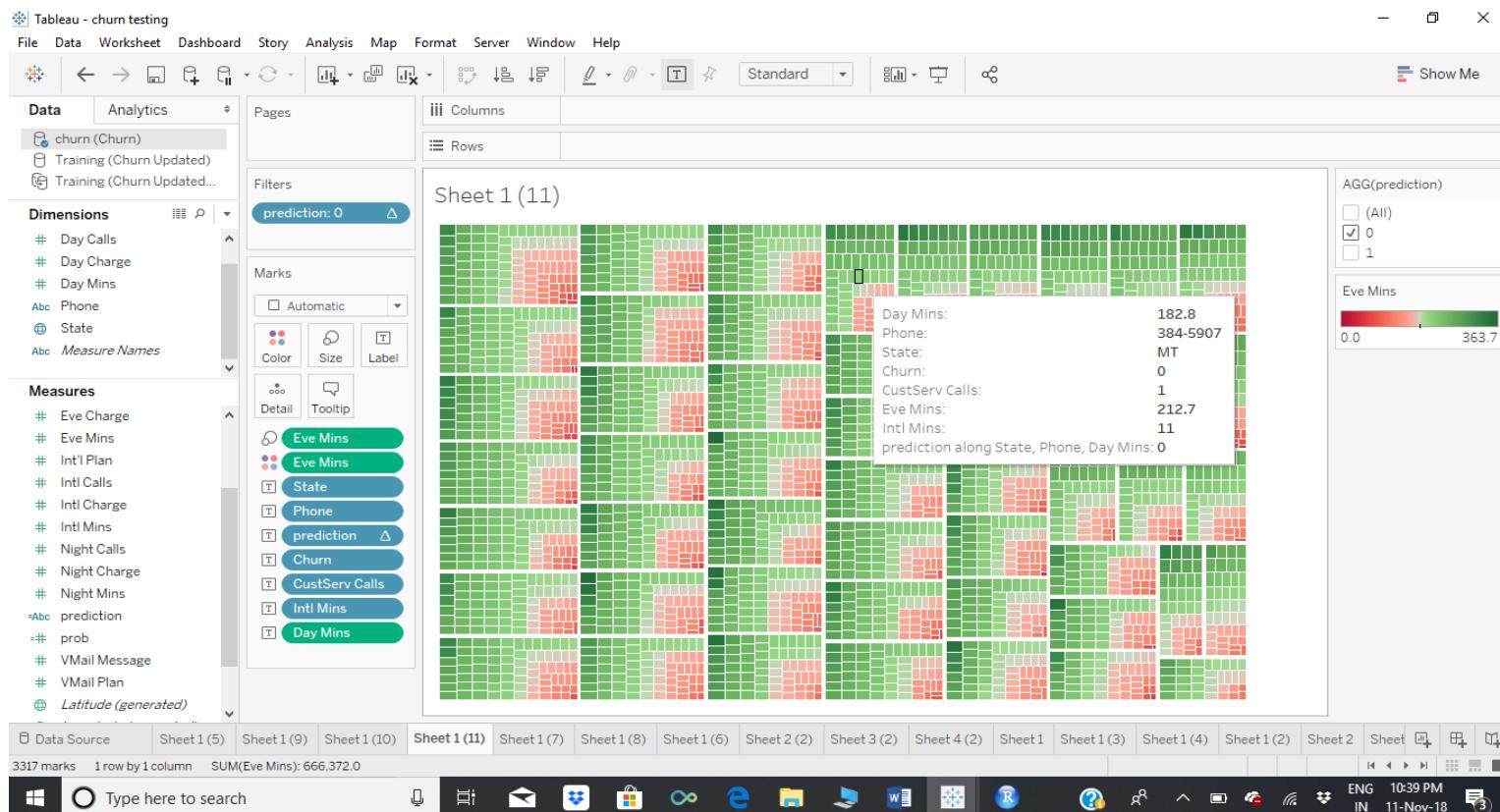


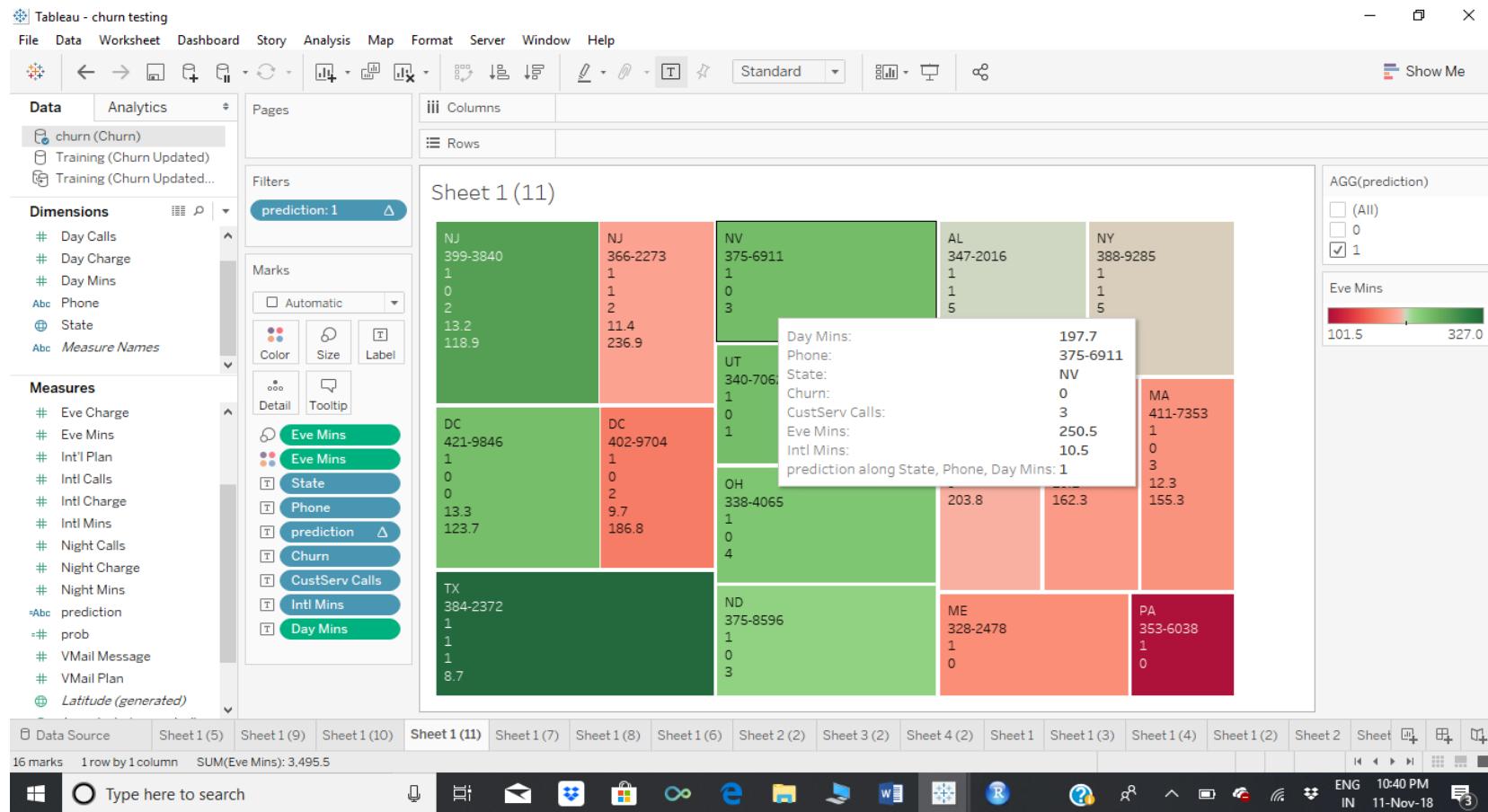






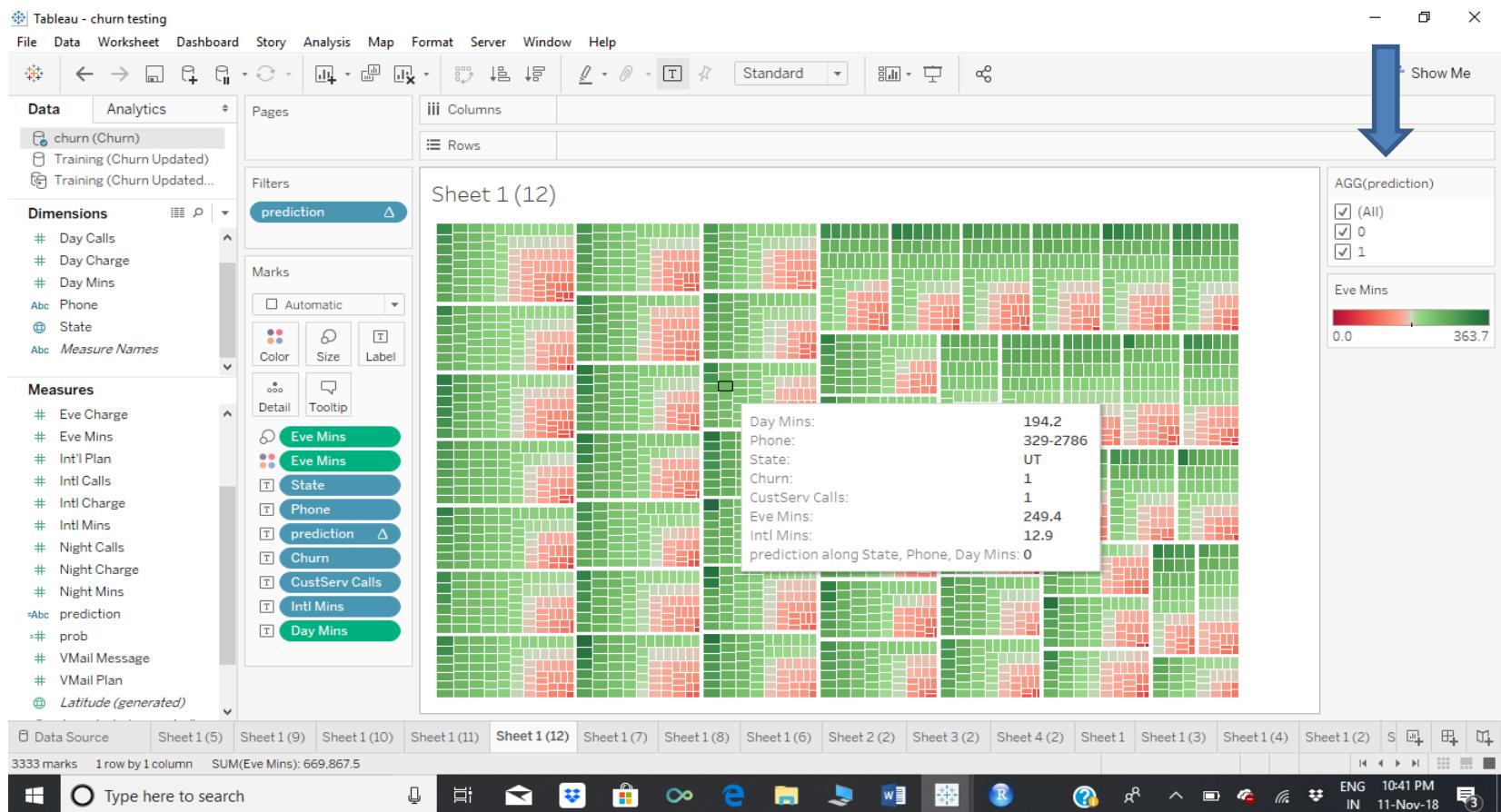






Prediction in slider

we can select 0,1 or All



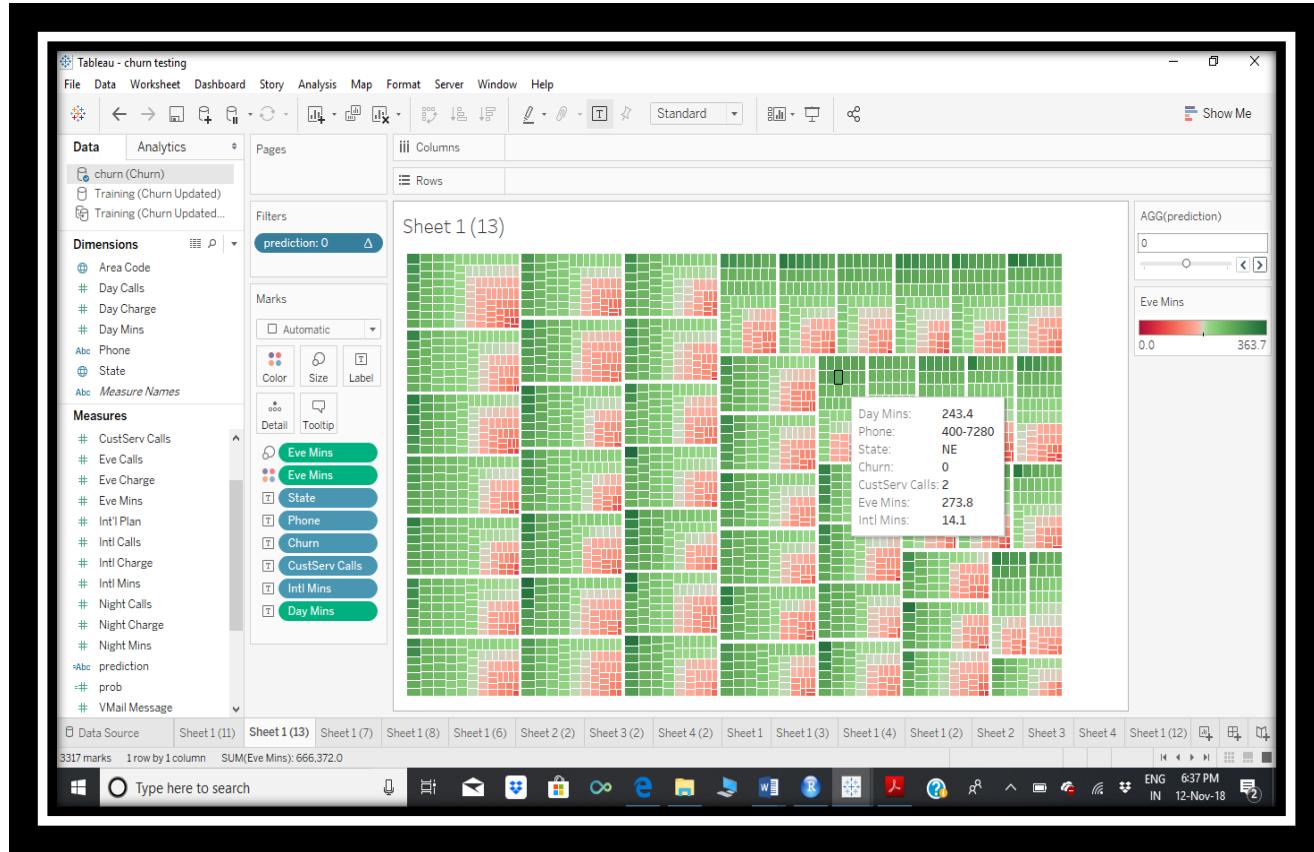


Tableau - churn testing

File Data Worksheet Dashboard Story Analysis Map Format Server Window Help

Show Me

Data Analytics

Pages

Columns

Rows Day Mins SUM(CustServ Call..) Area Code State SUM(Intl Mins) Day Mins

Filters prediction Phone

Marks Area Color Size Label Detail Tooltip

slider churn 0 & 1

Day Mins	CustServ Ca..	Area Code	State	Intl Mins	Day Mins
291.6	2	415	HI	14	100 0
291.8	0	510	MN	9.9	300 200 100 0
292.9	4	510	ME	13.8	300 200 100 0
293	2	510	IN	10	300 200 100 0
293.3	0	415	KS	14.5	300 200 100 0

AGG(prediction)

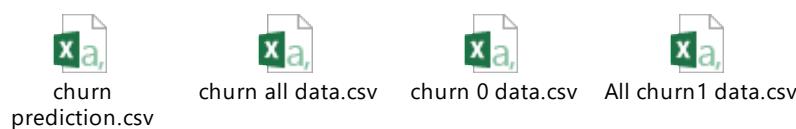
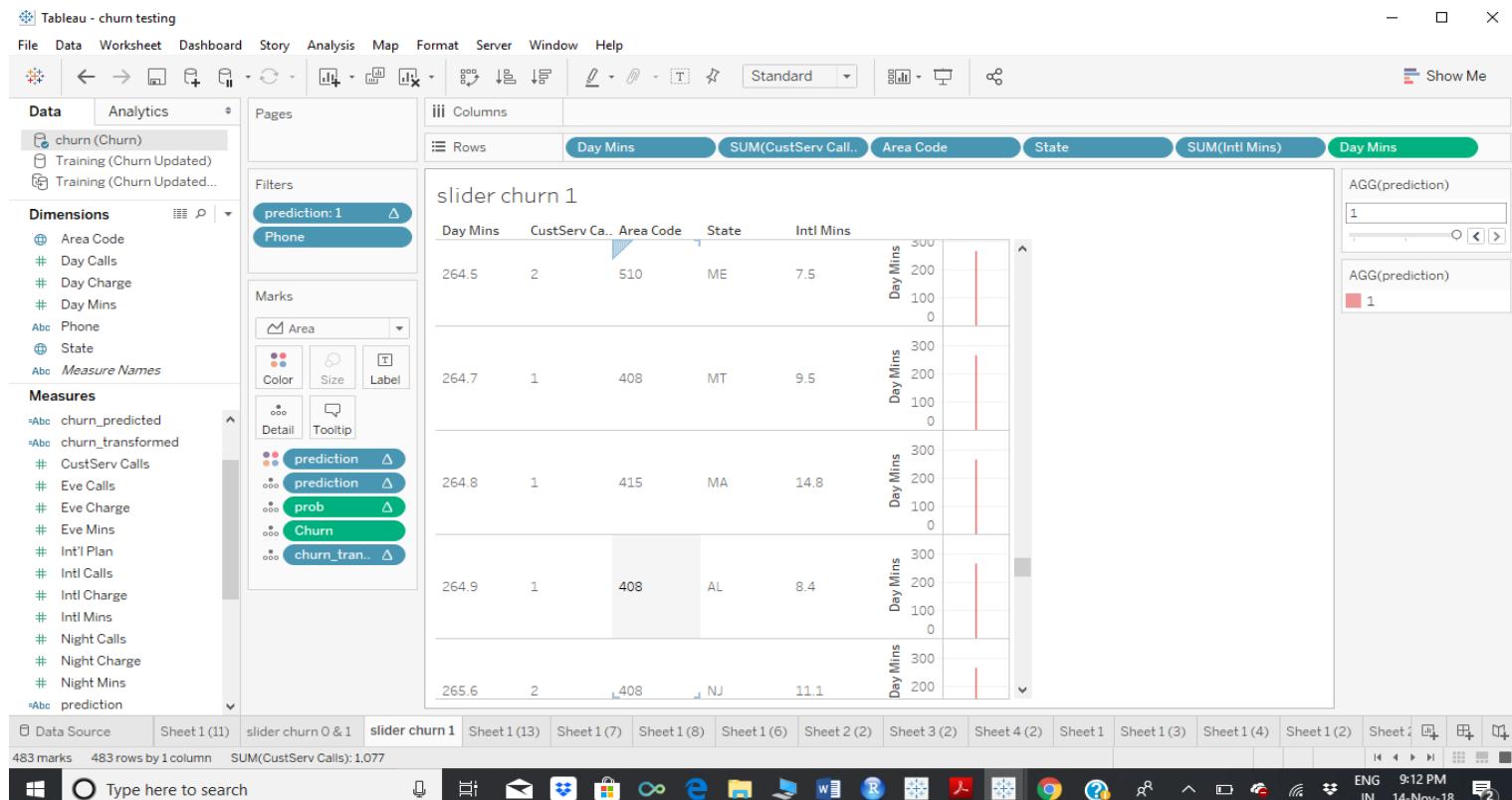
AGG(prediction)

0 1

Data Source Sheet1(11) slider churn 0 & 1 slider churn 1 Sheet1(13) Sheet1(7) Sheet1(8) Sheet1(6) Sheet2(2) Sheet3(2) Sheet4(2) Sheet1 Sheet1(3) Sheet1(4) Sheet1(2) Sheet1(1) 3315 marks 3315 rows by 1 column SUM(CustServ Calls): 5,209

Type here to search

ENG 9:11 PM IN 14-Nov-18



In this project a customer churn analysis was presented for churn data provided in the project 2. The analysis focused on churn prediction based on logistic regression. The different models predicted the actual churners relatively well.

The differences between the models input data (the significance level in case of each of the variables) indicates the dynamic nature of the churning customer profile. This makes it hard to formulate one standard model that could be used as the predictive model in the future. The findings of this study indicate that, in case of logistic regression model, the user should update the model to be able to produce predictions with high accuracy. It is interesting for a company's perspective whether the churning customers are worth retaining or not. And also in marketing perspective what can be done to retain them. For Visualization Tableau desktop is used. Calculated fields is used to calculate the probability and the fitted results prediction. Using the Library Reserve in R Logistic regression model is connected with Tableau.

Through this model we can identify easily the churn 1 or 0 and the probabilities through a slider to visualize the churn 1 , churn 0 and all. The visualization in Tableau provides separately state wise, Area code, specific telephone no and their probability to predict and visualize them nicely. The data can be imported and stored as excel data for further analysis and churn predictions. These files are exported and stored and attached in this project file.

Acknowledgement

This is a quite interesting project and I have gained a lot of knowledge about breast cancer and the identification of tumors through Machine Learning classification Model. I thank the institute Acadgild and the Mentors Dr.Vinod and Mr. Sunil who taught us the R coding and other subjects to understand the Algorithms. I thank the support coordinator Mr. Anuj for guiding me to understand the project related queries and complete the project on time. I once again

thank Acadgild for enlighten me on Machine learning through online teaching and various coding support through the support coordinators. Thank you Acadgild.

