|  |  |
| --- | --- |
|  | CHURN analysis |
|  |  |
| 7/2/2017 | Predictive Modeling |
|  | **Group 2**  Shruthi  Meena  Krishna  Manish  Santhosh |

Contents

[DESCRIPTION. 2](#_Toc486762361)

[Approach 3](#_Toc486762362)

[Understanding the Data 3](#_Toc486762363)

[Exploratory Data Analysis of Raw Data: 3](#_Toc486762364)

[Check for Correlations 3](#_Toc486762365)

[Can the Linear Regression (OLS) solve the problem? 5](#_Toc486762366)

[Perform Logistic Regression 5](#_Toc486762367)

[Testing “Goodness of Fit” 6](#_Toc486762368)

[Log likelihood Ratio Test 6](#_Toc486762369)

[Testing for Pseudo R Square 6](#_Toc486762370)

[Interpreting Summary from logistic regression model 6](#_Toc486762371)

[Odds and Probability 7](#_Toc486762372)

[Results from Goodness of fit. 7](#_Toc486762373)

[Predict using the Model 7](#_Toc486762374)

[Test for Accuracy 8](#_Toc486762375)

[Confusion matrix or classification table 8](#_Toc486762376)

[ROC Curve: Receiver Operating Characteristic(ROC) 8](#_Toc486762377)

[Conclusion 9](#_Toc486762378)

[Appendix 11](#_Toc486762379)

[R-Code 11](#_Toc486762380)

CHURN analysis

DESCRIPTION.

Customer usage data foe wireless customers is given and we have to build the model to predict the Churn

The **churn** rate, also known as the rate of attrition, is the percentage of subscribers to a service who discontinue their subscriptions to that service within a given time period. For a company to expand its clientele, its growth rate, as measured by the number of new customers, must exceed its **churn** rate.

In the given data set, Churn rate is 17%. Our Model should help the organization to predict the customers who might Churn in future and apply respective treatment to avoid the situation. Our scope will be only to predict the Churning customers.

To do that we are provided with following variables.

|  |  |
| --- | --- |
| Variables |  |
| Churn (Y) | 1 if customer cancelled service, 0 if not |
| AccountWeeks | number of weeks customer has had active account |
| ContractRenewal | 1 if customer recently renewed contract, 0 if not |
| DataPlan | 1 if customer has data plan, 0 if not |
| DataUsage | gigabytes of monthly data usage |
| CustServCalls | number of calls into customer service |
| DayMins | average daytime minutes per month |
| DayCalls | average number of daytime calls |
| MonthlyCharge | average monthly bill |
| OverageFee | largest overage fee in last 12 months |
| RoamMins | average number of roaming minutes |

## Approach

Use LOGISTIC Regression.

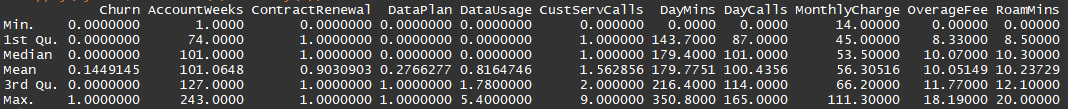
Because, Churn Data, which is to be predicted turns out to be dichotomous (Categorical) and All the independent variables or predictors are either continuous are categorical. We can Justify this by showing probability output from linear and why it is not best fit.

Step by Step approach explained:

1. Understand the data
2. Look for any existence of the correlations
3. Run Linear to check if the predicted probability is between the range of 0 to 1.
4. If Step 3 proves false. Perform Logistic regression
5. Split the data to Training data and Test data in (70:30) proportion.
6. Perform logits regression on training data.
7. Test the “Goodness of Fit”
8. Predict the Test Data using the Model developed.
9. Measure the accuracy of Prediction
10. Conclude.

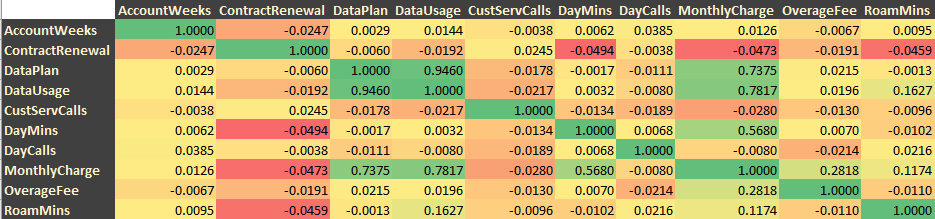
# Understanding the Data

## Exploratory Data Analysis of Raw Data:



sapply(mydata, function(x) summary(x))

## Check for Correlations



There seems to be Significant Correlations between the selective independent variables.

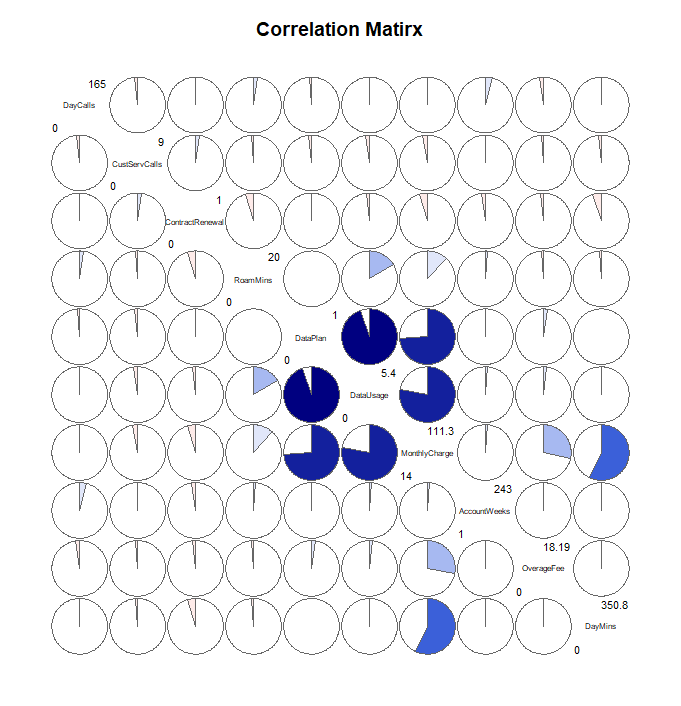
Like data plan and data usage, Monthly charges and Data plan etc. We can Confirm this by visually looking at the data. (Find “Correlation Matrix” image). It confirms the existence of strong correlation between,

Data Plan & Data Usage

Data Plan & Monthly Charge

Monthly Charge & DayMins

Monthly Charge & OverageMins



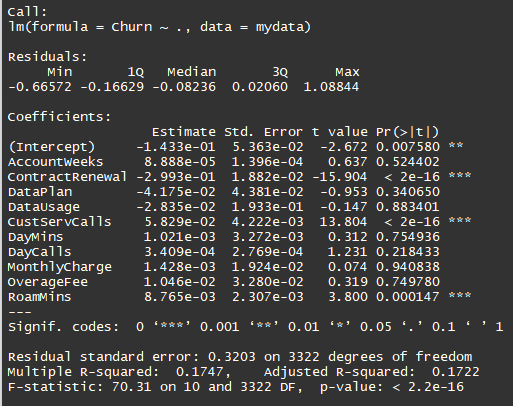
corData = cor(mydata[,-1],method = "pearson")

corrgram(mydata[,-1], order = TRUE, lower.panel = panel.pie, upper.panel = panel.pie, text.panel = panel.txt, main="Correlation Matirx", diag.panel = panel.minmax)

## 

## Can the Linear Regression (OLS) solve the problem?

regression = lm(Churn~., data = mydata)



After Predicting based on OLS, there are 474 observations, whose probability is beyond the realistic range of 0 to 1. Hence, we are using Logistic regression, instead of Linear regression. So, we cannot use it.

## Perform Logistic Regression

* Divide the data set into Training and Testing. We will train the model using training data and test it in Testing data.

set.seed(45)

splitPropotion = sample(nrow(mydata),nrow(mydata)\*.7)

mydata\_train = mydata[splitPropotion,]

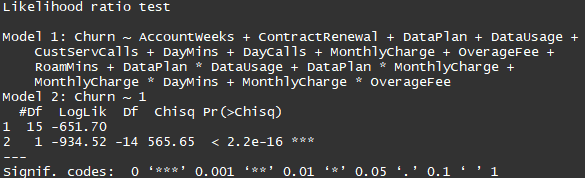
mydata\_test = mydata[-splitPropotion,]

* As we know there are 4 variables that are highly correlated in the following order, we will use them as interaction effect while defining the equation.
  1. Data Plan X Data Usage
  2. Data Plan X Monthly Charge
  3. Monthly Charge X DayMins
  4. Monthly Charge X OverageMins

logitModel = glm(Churn~AccountWeeks+ContractRenewal+DataPlan+DataUsage+CustServCalls+DayMins+DayCalls+MonthlyCharge+OverageFee+RoamMins+DataPlan\*DataUsage+DataPlan\*MonthlyCharge+MonthlyCharge\*DayMins+MonthlyCharge\*OverageFee,data = mydata\_train, family = binomial)

## Testing “Goodness of Fit”

### Log likelihood Ratio Test



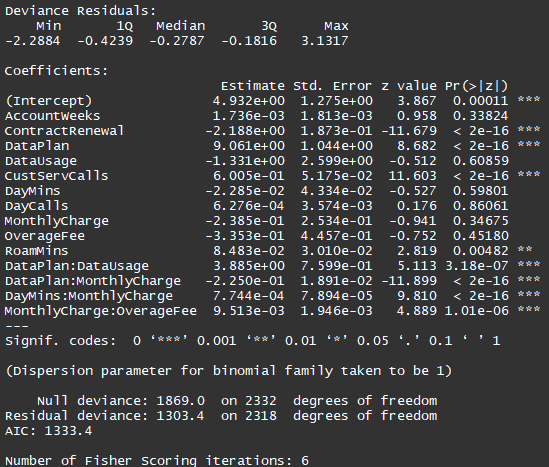
## The overall test of the model is significant based on the Chisq test and it is highly significant. This tells that likelihood of any customer churning out is heavily dependent on all the variables or at least one variable.

### Testing for Pseudo R Square

Screen Clipping

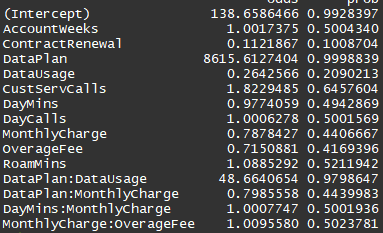
McFadden RSquare is 30%, which means that 30% of uncertainty of intercept only model (model2 in likelihood ration test) has been explained by full model (model1). **Goodness of fit is reasonable**

### Interpreting Summary from logistic regression model



* Customers who has recently renewed the contract, with significant usage of data during the day, customers’ whose monthly charges are more, and who has paid significantly large amount in past 12 months is negatively impacting the churn, which means they are churning out. Especially recent contact renewals.
* Customers having data services and also users of roaming services are positively impacting the churn. Especially customers with Data as they are Statistically Significant.
* Customers who frequently call the call centers seems to be well informed and feel more loyal, which is what reflecting in Positive impact with high Significance.
* All the correlated variables are statistically very significant and also most of them are positive.
* Deviance has dropped sharply from Null to Residual, that signifies that Independent variables are impacting the churn strongly
* Company has to look in the existing contact renewal model to control churn, and also revisit and improvise its data plan so that we can bring in more customers to data and control churn.
* There is a strong correlation between Data Plan and Usage and Charges, also, significantly impacting the churn. Company might want to look into pricing structure as well.

### Odds and Probability



If data plan changes by 1 unit, the odds for churn is impacted by 8616 times, compared to the loyal customers.

Similar trend we can see for DataPlan with DataUsage

### Results from Goodness of fit.

All the above 4 tests suggest that Model developed is significantly good enough to be used for prediction. Let us procced with prediction.

## Predict using the Model

logitPredcit = predict(logitModel,newdata = mydata\_test, type = "response")

#### Can the logistic regression solve the problem?

sqldf('select \* from df\_logitPredcit where logitPredcit > 1 or logitPredcit < 0')

Since we have converted them to Probability based on Exponentials, we do not see negative probability or probability more than 1. This way we have brought all the likelihood with the realistic range.

We have to set the Cut-off value for Prediction. By default, it is .5.

ChurnPredicted = as.data.frame(ifelse(logitPredcit>.5,1,0))

Now That we have predicted, we have test the performance or accuracy of prediction.

## Test for Accuracy

### Confusion matrix or classification table

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Prediction** |  |
|  |  | No Churn | Churn |
| **Actual** | No Churn | 824 | 14 |
|  | Churn | 104 | 58 |

* **Model is 88% Accurate ((TP+TN)/(TP+FN+FP+TN))**
* **35% of time model predicted churning customers correctly. (Specificity = TN/(TN+FP))**
* **98% of times model predicted non-churning customers correctly. (Sensitivity =TP/(TP+FN))**

### ROC Curve: Receiver Operating Characteristic(ROC)

Model’s performance by evaluating the tradeoffs between true positive rate (sensitivity) and false positive rate(1- specificity).

With assumption of cut of point at .5

.90-1 = excellent (A)

.80-.90 = good (B)

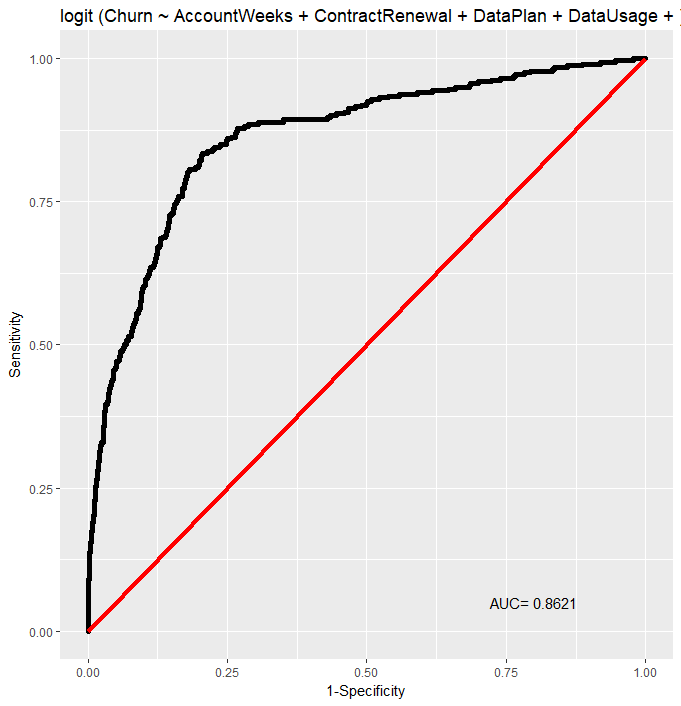
.70-.80 = fair (C)

.60-.70 = poor (D)

.50-.60 = fail (F)

(Refer the Image Provide in next page)

**Here we have AUC - 0.86, which is a good predictive model**



## Conclusion

Using the equation "AccountWeeks+ContractRenewal+DataPlan+DataUsage+CustServCalls+DayMins+DayCalls+MonthlyCharge+OverageFee+RoamMins+DataPlan\*DataUsage+DataPlan\*MonthlyCharge+MonthlyCharge\*DayMins+MonthlyCharge\*OverageFee

We should be able to predict the churning customer 86% of time accurately.

Organization must look into Data Plan , Usage and Charges to control the churn.

# Appendix

## R-Code

#Required Libraries

library(googleVis)

library(corrgram)

library(Deducer) #ROC curve

library(sqldf)

library(lmtest) #log likelihood

library(pscl) #Pseudo r squared

#Read the Data from CSV File

getwd()

setwd("C:/Home/Work/Data Science/models-and-methods")

mydata = read.table(file = "Logit/Data/CellPhone.csv", header = T, sep = ",")

names(mydata)[1] = "Churn"

sapply(mydata, function(x) summary(x))

#Check for Correlation

corData = cor(mydata[,-1],method = "pearson")

corData

corrgram(mydata[,-1], order = TRUE, lower.panel = panel.pie, upper.panel = panel.pie, text.panel = panel.txt, main="Correlation Matirx", diag.panel = panel.minmax)

#Performing Linear Regression

regression = lm(Churn~., data = mydata)

summary(regression)

lrPredict = predict(regression)

df\_lrPredict = as.data.frame(lrPredict)

sqldf('select count(1) from df\_lrPredict where lrPredict>1 or lrPredict <0 ')

#Predicting the Churn using Logistic Regression

#Splitting the data into 70:30

set.seed(45)

splitPropotion = sample(nrow(mydata),nrow(mydata)\*.7)

mydata\_train = mydata[splitPropotion,]

mydata\_test = mydata[-splitPropotion,]

#Model Equation and Model

logitModel = glm(Churn~AccountWeeks+ContractRenewal+DataPlan+DataUsage+CustServCalls+DayMins+DayCalls+MonthlyCharge+OverageFee+RoamMins+DataPlan\*DataUsage+DataPlan\*MonthlyCharge+MonthlyCharge\*DayMins+MonthlyCharge\*OverageFee,data = mydata\_train, family = binomial)

#Testing for Goodness of fit

### Log likelihood Ratio Test

lrtest(logitModel)

###Pseudo R Square

pR2(logitModel)

#Summary of the model

summary(logitModel)

#Odds and Probability

odds = exp(coef(logitModel))

prob = odds/(odds + 1)

odds\_and\_prob = cbind(as.data.frame(odds),as.data.frame(prob)[,1])

names(odds\_and\_prob)[2] = "prob"

odds\_and\_prob

confint(logitModel)

# Lets us now Predict the Test data using the model

logitPredcit = predict(logitModel,newdata = mydata\_test, type = "response")

df\_logitPredcit = as.data.frame(logitPredcit)

sqldf('select \* from df\_logitPredcit where logitPredcit > 1 or logitPredcit < 0')

# Setting the cut-off value.

ChurnPredicted = as.data.frame(ifelse(logitPredcit>.5,1,0))

names(ChurnPredicted)[1] = "ChurnPredicted"

# Testing the Performance or Accuraracy of Fit

## Confusion matrix or classification table

glm\_CM=table(mydata\_test$Churn, ChurnPredicted$ChurnPredicted)

glm\_CM = as.matrix(glm\_CM)

glm\_TP = glm\_CM[1,1]

glm\_FN = glm\_CM[1,2]

glm\_FP = glm\_CM[2,1]

glm\_TN = glm\_CM[2,2]

# accuracy = (TP+TN)/(TP+FN+FP+TN)

glm\_accuracy = (glm\_TP + glm\_TN)/(glm\_TP+glm\_TN+glm\_FP+glm\_FN)

glm\_accuracy

# Specificity = TN/(TN+FP)

glm\_Specificity = glm\_TN/(glm\_TN+glm\_FP)

glm\_Specificity

# Sensitivity =TP/(TP+FN)

glm\_Sensitivity = glm\_TP/(glm\_TP+glm\_FN)

glm\_Sensitivity

##ROC Curve: Receiver Operating Characteristic(ROC)

# model’s performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity).

# with assumption of cut of point at .5

#.90-1 = excellent (A)

#.80-.90 = good (B)

#.70-.80 = fair (C)

#.60-.70 = poor (D)

#.50-.60 = fail (F)

rocplot(logitModel)