GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI BABI-Jan-2017 | Group 2

CHURN ANALYSIS

7/2/2017

Predictive Modeling

Group 2

Shruthi

Meena

Krishna

Manish

Santhosh

Contents

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

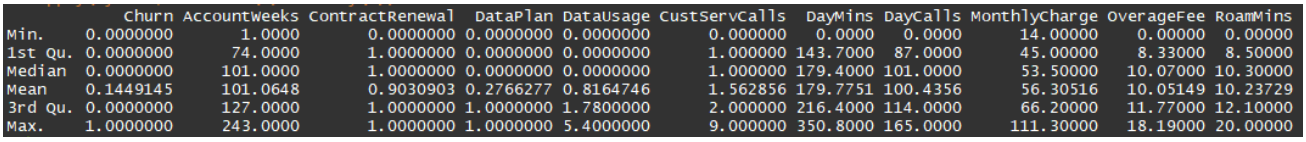
Use LOGISTIC Regression.
Because, Churn Data, which is to be predicted turns out to be dichotomous (Categorical) and All the independent variables or predictors are either continuous are categorical. We can Justify this by showing probability output from linear and why it is not best fit.

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

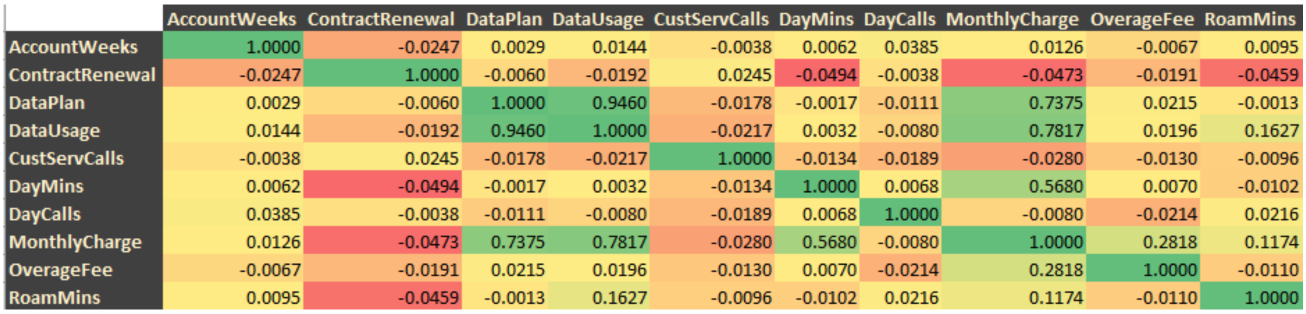
Exploratory Data Analysis of Raw Data:



```
          Churn AccountWeeks ContractRenewal  DataPlan DataUsage CustServCalls  DayMins  DayCalls MonthlyCharge OverageFee RoamMins
Min.    0.0000000       1.0000       0.0000000 0.0000000 0.0000000      0.000000   0.0000    0.0000      14.00000    0.00000  0.00000
1st Qu. 0.0000000      74.0000       1.0000000 0.0000000 0.0000000      1.000000 143.7000   87.0000      45.00000    8.33000  8.50000
Median  0.0000000     101.0000       1.0000000 0.0000000 0.0000000      1.000000 179.4000  101.0000      53.50000   10.07000 10.30000
Mean    0.1449145     101.0648       0.9030903 0.2766277 0.8164746      1.562856 179.7751  100.4356      56.30516   10.05149 10.23729
3rd Qu. 0.0000000     127.0000       1.0000000 1.0000000 1.7800000      2.000000 216.4000  114.0000      66.20000   11.77000 12.10000
Max.    1.0000000     243.0000       1.0000000 1.0000000 5.4000000      9.000000 350.8000  165.0000     111.30000   18.19000 20.00000
```

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

Check for Correlations



	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
AccountWeeks	1.0000	-0.0247	0.0029	0.0144	-0.0038	0.0062	0.0385	0.0126	-0.0067	0.0095
ContractRenewal	-0.0247	1.0000	-0.0060	-0.0192	0.0245	-0.0494	-0.0038	-0.0473	-0.0191	-0.0459
DataPlan	0.0029	-0.0060	1.0000	0.9460	-0.0178	-0.0017	-0.0111	0.7375	0.0215	-0.0013
DataUsage	0.0144	-0.0192	0.9460	1.0000	-0.0217	0.0032	-0.0080	0.7817	0.0196	0.1627
CustServCalls	-0.0038	0.0245	-0.0178	-0.0217	1.0000	-0.0134	-0.0189	-0.0280	-0.0130	-0.0096
DayMins	0.0062	-0.0494	-0.0017	0.0032	-0.0134	1.0000	0.0068	0.5680	0.0070	-0.0102
DayCalls	0.0385	-0.0038	-0.0111	-0.0080	-0.0189	0.0068	1.0000	-0.0080	-0.0214	0.0216
MonthlyCharge	0.0126	-0.0473	0.7375	0.7817	-0.0280	0.5680	-0.0080	1.0000	0.2818	0.1174
OverageFee	-0.0067	-0.0191	0.0215	0.0196	-0.0130	0.0070	-0.0214	0.2818	1.0000	-0.0110
RoamMins	0.0095	-0.0459	-0.0013	0.1627	-0.0096	-0.0102	0.0216	0.1174	-0.0110	1.0000

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

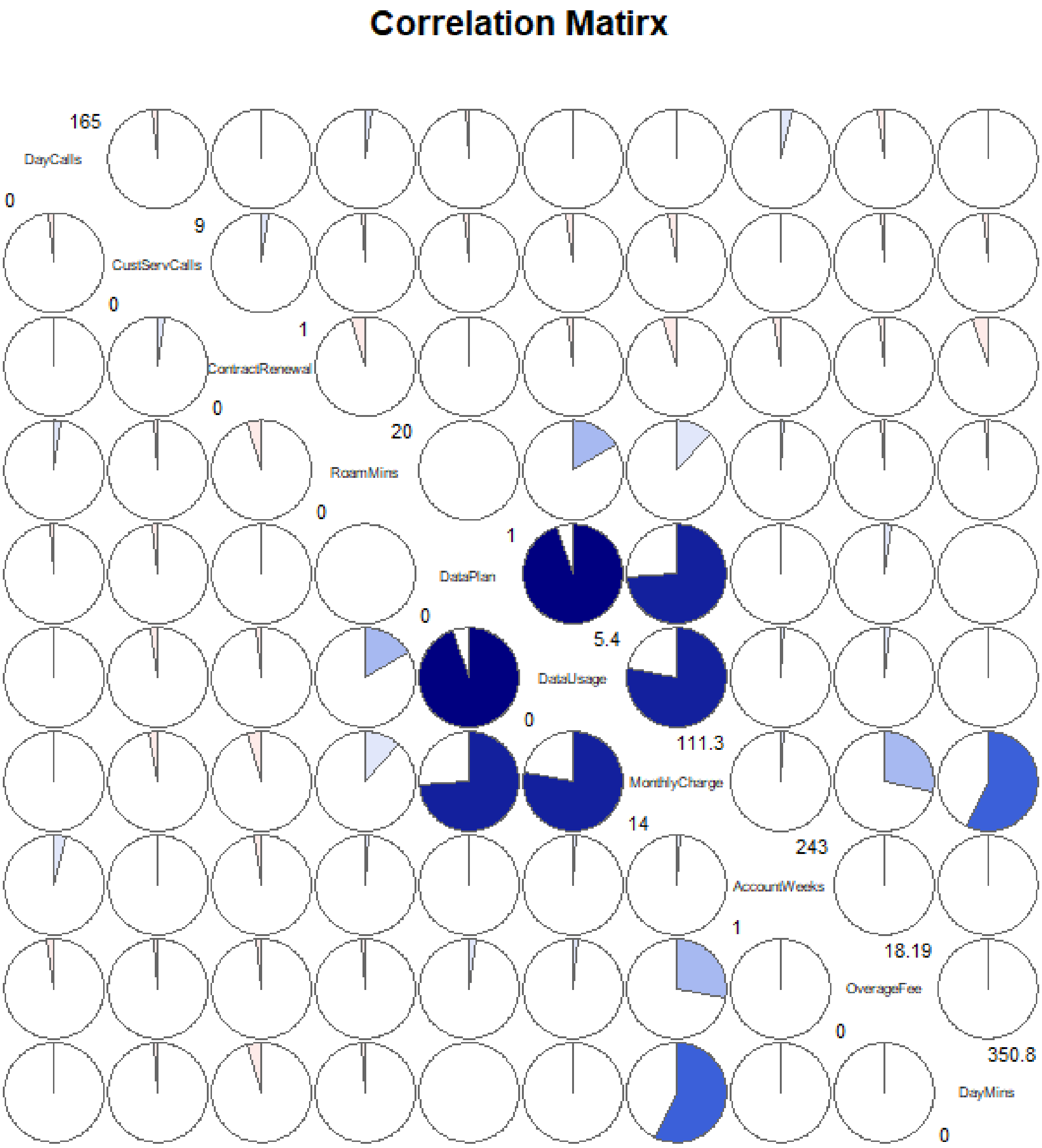
Like data plan and data usage, Monthly charges and Data plan etc. We can Confirm this by visually looking at the data. (Find "Correlation Matrix" image). It confirms the existence of strong correlation between,

Data Plan & Data Usage

Data Plan & Monthly Charge

Monthly Charge & DayMins

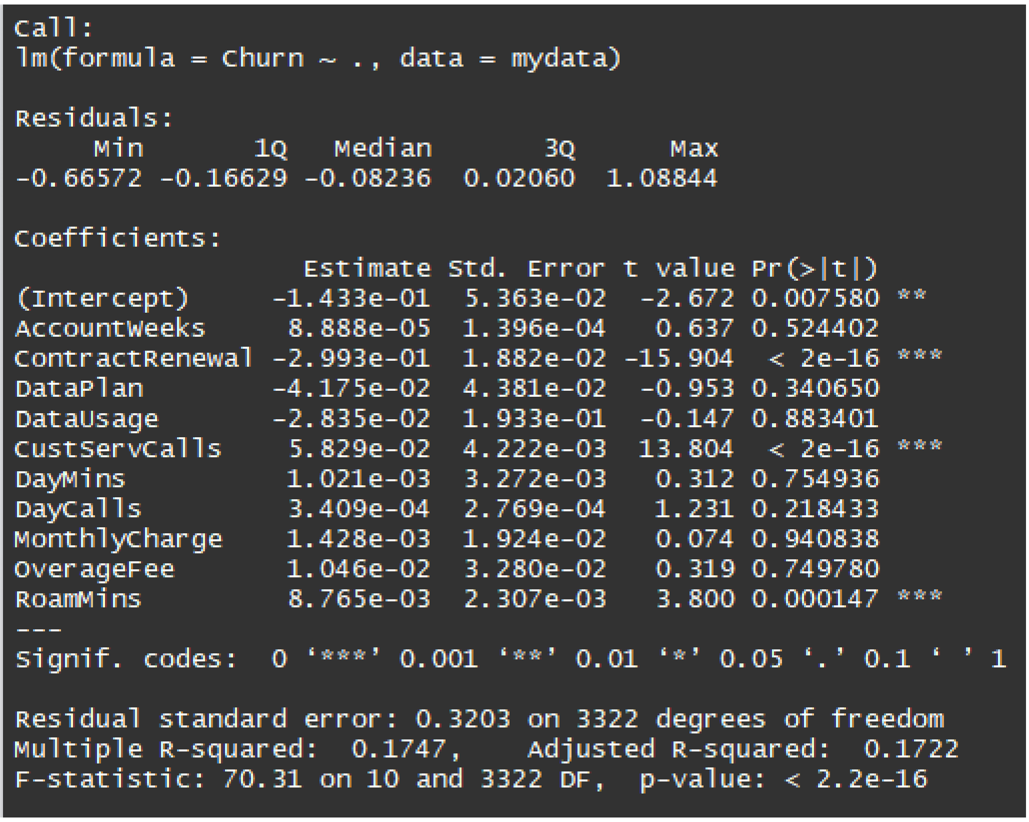
Monthly Charge & OverageMins



```
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)



```
Call:
lm(formula = Churn ~ ., data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.66572 -0.16629 -0.08236  0.02060  1.08844 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.433e-01  5.363e-02  -2.672 0.007580 ** 
AccountWeeks     8.888e-05  1.396e-04   0.637 0.524402    
ContractRenewal -2.993e-01  1.882e-02 -15.904  < 2e-16 ***
DataPlan        -4.175e-02  4.381e-02  -0.953 0.340650    
DataUsage       -2.835e-02  1.933e-01  -0.147 0.883401    
CustServCalls    5.829e-02  4.222e-03  13.804  < 2e-16 ***
DayMins          1.021e-03  3.272e-03   0.312 0.754936    
DayCalls         3.409e-04  2.769e-04   1.231 0.218433    
MonthlyCharge    1.428e-03  1.924e-02   0.074 0.940838    
OverageFee       1.046e-02  3.280e-02   0.319 0.749780    
RoamMins         8.765e-03  2.307e-03   3.800 0.000147 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3203 on 3322 degrees of freedom
Multiple R-squared:  0.1747,	Adjusted R-squared:  0.1722 
F-statistic: 70.31 on 10 and 3322 DF,  p-value: < 2.2e-16
```

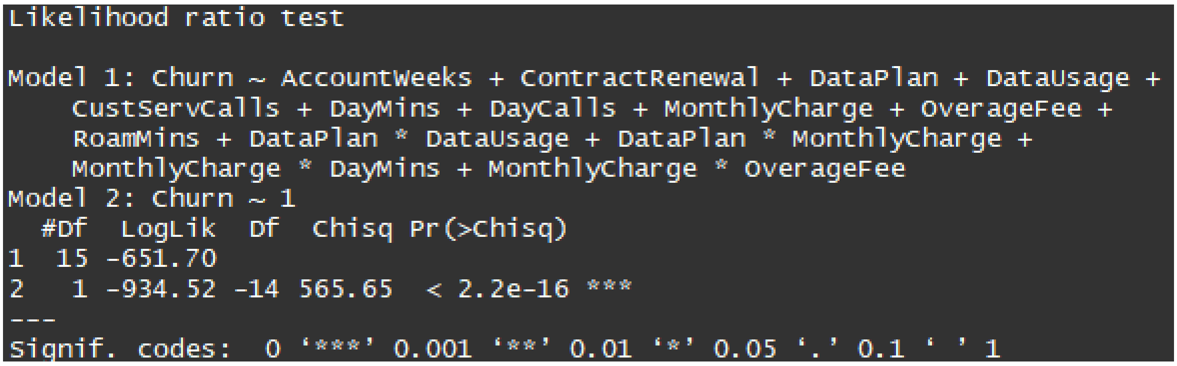
After Predicting based on OLS, there are 474 observations, whose probability is beyond the realistic range of 0 to 1. Hence, we are using Logistic regression, instead of Linear regression. So, we cannot use it.

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.
  a. Data Plan X Data Usage
  b. Data Plan X Monthly Charge
  c. Monthly Charge X DayMins
  d. 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"

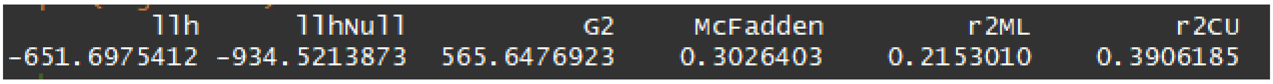
Log likelihood Ratio Test



```
Likelihood ratio test

Model 1: Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +
    CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee +
    RoamMins + DataPlan * DataUsage + DataPlan * MonthlyCharge +
    MonthlyCharge * DayMins + MonthlyCharge * OverageFee
Model 2: Churn ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  15 -651.70
2   1 -934.52 -14 565.65  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## 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.

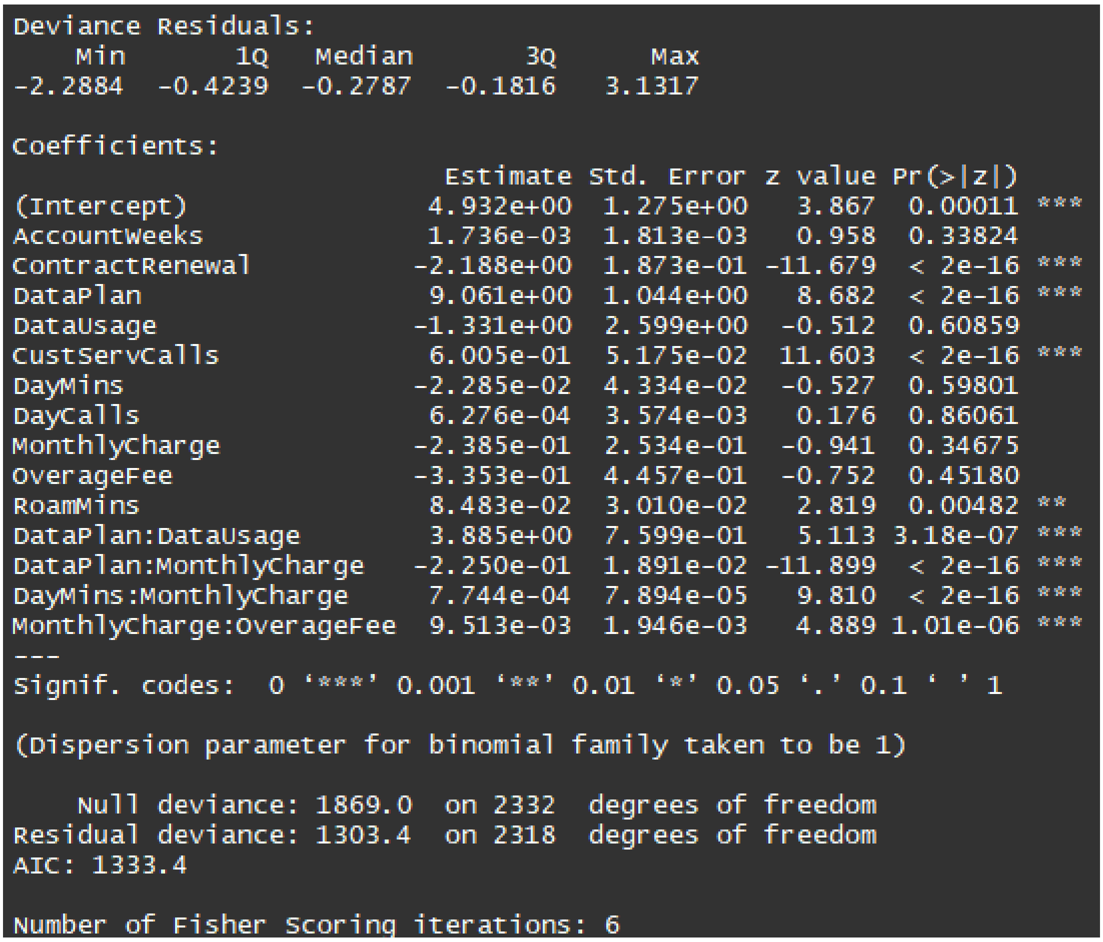
Testing for Pseudo R Square



```
         llh       llhNull            G2      McFadden          r2ML          r2CU
-651.6975412 -934.5213873  565.6476923     0.3026403     0.2153010     0.3906185
```

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**

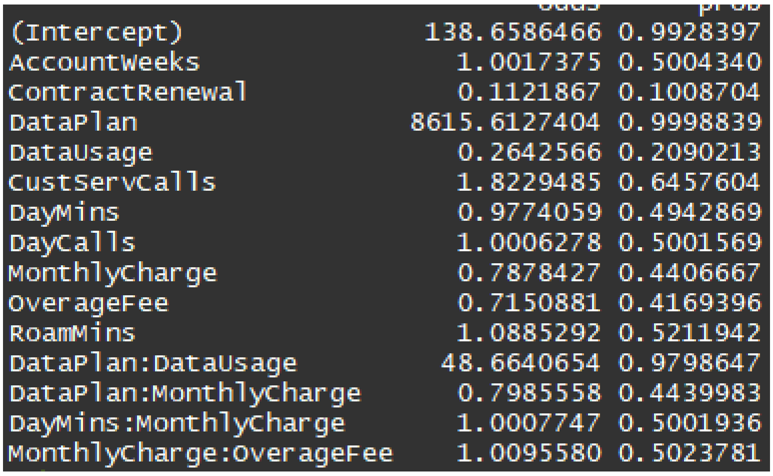
Interpreting Summary from logistic regression model



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2884  -0.4239  -0.2787  -0.1816   3.1317

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               4.932e+00  1.275e+00   3.867  0.00011 ***
AccountWeeks              1.736e-03  1.813e-03   0.958  0.33824
ContractRenewal          -2.188e+00  1.873e-01 -11.679  < 2e-16 ***
DataPlan                  9.061e+00  1.044e+00   8.682  < 2e-16 ***
DataUsage                -1.331e+00  2.599e+00  -0.512  0.60859
CustServCalls             6.005e-01  5.175e-02  11.603  < 2e-16 ***
DayMins                  -2.285e-02  4.334e-02  -0.527  0.59801
DayCalls                  6.276e-04  3.574e-03   0.176  0.86061
MonthlyCharge            -2.385e-01  2.534e-01  -0.941  0.34675
OverageFee               -3.353e-01  4.457e-01  -0.752  0.45180
RoamMins                  8.483e-02  3.010e-02   2.819  0.00482 **
DataPlan:DataUsage        3.885e+00  7.599e-01   5.113 3.18e-07 ***
DataPlan:MonthlyCharge   -2.250e-01  1.891e-02 -11.899  < 2e-16 ***
DayMins:MonthlyCharge     7.744e-04  7.894e-05   9.810  < 2e-16 ***
MonthlyCharge:OverageFee  9.513e-03  1.946e-03   4.889 1.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1869.0  on 2332  degrees of freedom
Residual deviance: 1303.4  on 2318  degrees of freedom
AIC: 1333.4

Number of Fisher Scoring iterations: 6
```

- 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



```
                                 odds      prob
(Intercept)                 138.6586466 0.9928397
AccountWeeks                  1.0017375 0.5004340
ContractRenewal               0.1121867 0.1008704
DataPlan                   8615.6127404 0.9998839
DataUsage                     0.2642566 0.2090213
CustServCalls                 1.8229485 0.6457604
DayMins                       0.9774059 0.4942869
DayCalls                      1.0006278 0.5001569
MonthlyCharge                 0.7878427 0.4406667
OverageFee                    0.7150881 0.4169396
RoamMins                      1.0885292 0.5211942
DataPlan:DataUsage           48.6640654 0.9798647
DataPlan:MonthlyCharge        0.7985558 0.4439983
DayMins:MonthlyCharge         1.0007747 0.5001936
MonthlyCharge:OverageFee      1.0095580 0.5023781
```

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)

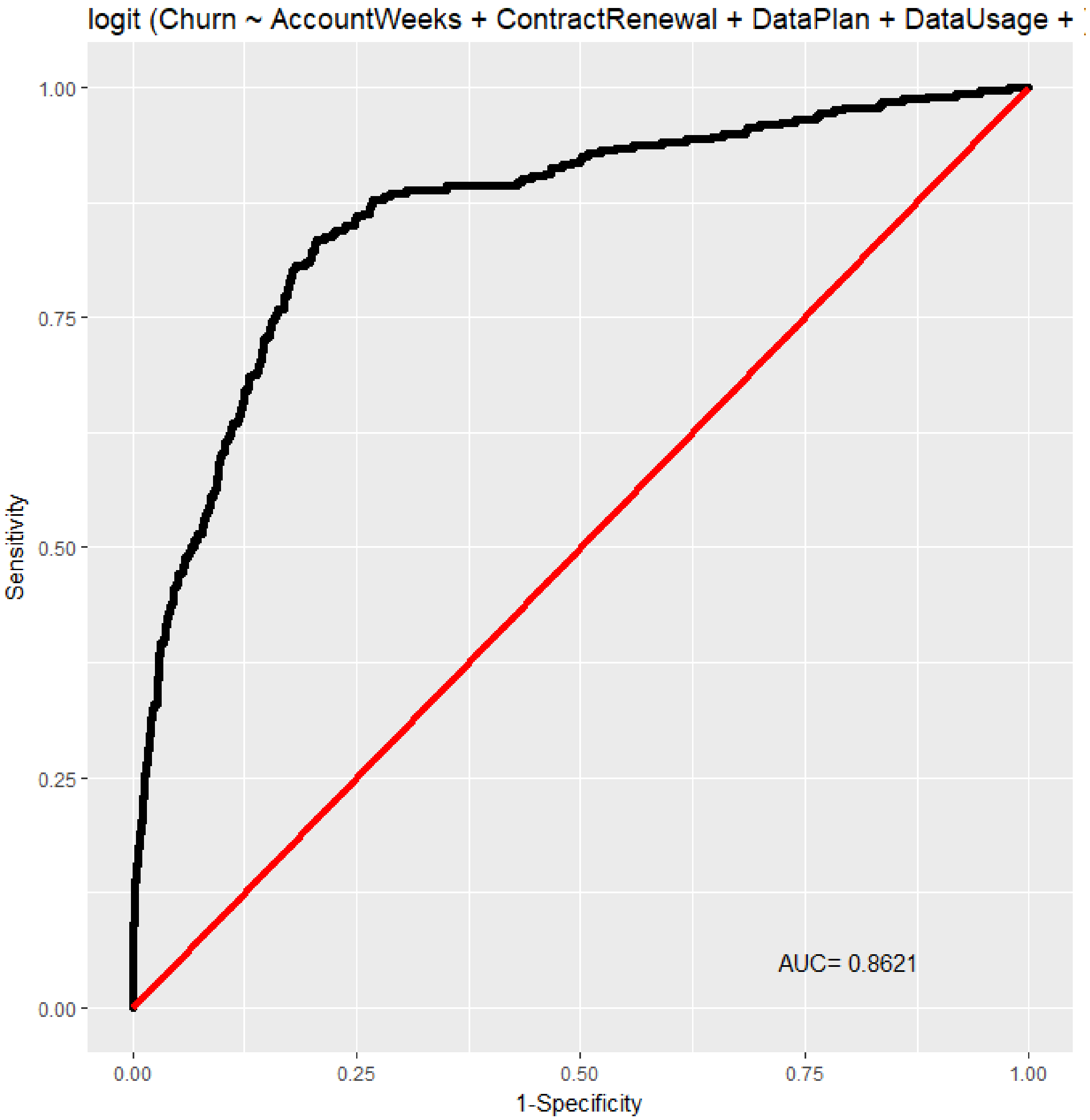
.70-.80 = fair (C)

.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(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:/___/")
mydata = read.table(file = " 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)
```