**PREDICTIVE MODELING GROUP ASSIGNMENT**

**LOGISTIC REGRESSION-CELLPHONE DATA**

Mugila.M,

MuthuVignesh.N,

Sindhuja Hariharan,

Valli Mohan

**You will have to build a logistic regression model and interpret the result. Make sure you partition the data set by allocating 70% -for training data and 30% -for validating the results.**

**Solution:**

1. **Import Data and Analyze the Structure:**

The data is read and the structure is analyzed.

data = read.csv("Dataset\_Cellphone.csv",header = TRUE)

str(data)

'data.frame': 3333 obs. of 11 variables:

$ Churn : int 0 0 0 0 0 0 0 0 0 0 ...

$ AccountWeeks : int 128 107 137 84 75 118 121 147 117 141 ...

$ ContractRenewal: int 1 1 1 0 0 0 1 0 1 0 ...

$ DataPlan : int 1 1 0 0 0 0 1 0 0 1 ...

$ DataUsage : num 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...

$ CustServCalls : int 1 1 0 2 3 0 3 0 1 0 ...

$ DayMins : num 265 162 243 299 167 ...

$ DayCalls : int 110 123 114 71 113 98 88 79 97 84 ...

$ MonthlyCharge : num 89 82 52 57 41 57 87.3 36 63.9 93.2 ...

$ OverageFee : num 9.87 9.78 6.06 3.1 7.42 ...

$ RoamMins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...

The categorical variables are converted into factor variables.

|  |
| --- |
| data$ContractRenewal<-factor(data$ContractRenewal,levels=c("0","1"),labels=c("Non-Renewed","Renewed"))  data$DataPlan<-factor(data$DataPlan,levels=c("0","1"),labels=c("No","Yes"))  str(data)  'data.frame': 3333 obs. of 11 variables:  $ Churn : int 0 0 0 0 0 0 0 0 0 0 ...  $ AccountWeeks : int 128 107 137 84 75 118 121 147 117 141 ...  $ ContractRenewal: Factor w/ 2 levels "Non-Renewed",..: NA NANANANANANANANANA ...  $ DataPlan : Factor w/ 2 levels "No","Yes": NA NANANANANANANANANA ...  $ DataUsage : num 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...  $ CustServCalls : int 1 1 0 2 3 0 3 0 1 0 ...  $ DayMins : num 265 162 243 299 167 ...  $ DayCalls : int 110 123 114 71 113 98 88 79 97 84 ...  $ MonthlyCharge : num 89 82 52 57 41 57 87.3 36 63.9 93.2 ...  $ OverageFee : num 9.87 9.78 6.06 3.1 7.42 ...  $ RoamMins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ... |
|  |
| |  | | --- | |  | |

1. **Splitting data into train and test data:**

The data is split into train and test data in the ratio of 70:30.

library("caTools")

set.seed(1234)

split<-sample.split(data$Churn, SplitRatio = 0.70)

train<-subset(data, split == TRUE)

test<-subset(data, split == FALSE)

table(train$Churn)

0 1

1817 304

table(test$Churn)

0 1

1033 179

1. **The Logistic Regression Model: Log Likelihood Ratio Test**

The logit model is executed.

The log likelihood ratio test is done.

logit = glm(Churn ~ AccountWeeks + ContractRenewal + DataPlan

+ DataUsage + CustServCalls + DayMins + DayCalls

+ MonthlyCharge + OverageFee + RoamMins , data=train,family=binomial)

library(lmtest)

lrtest(logit)

Likelihood ratio test

Model 1: Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +

CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee +

RoamMins

Model 2: Churn ~ 1

#Df LogLik Df ChisqPr(>Chisq)

1 11 -766.82

2 1 -965.21 -10 396.77 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Interpretation:**

* From the log likelihood ratio test, it is understood that the overall significance of the model based on the chisq test is **highly significant.**
* This denotes that the likelihood of churn depends on all other variables.
* It also implies that the null hypothesis of all betas are zero is rejected as the p value is less than 0.05.
* Also, we conclude that atleast one beta is nonzero.

1. **Mcfadden R Square:**

The Mcfadden R Square is obtained.

library(pscl)

pR2(logit)

llhllhNull G2 McFadden r2ML r2CU

-766.8221376 -965.2094900 396.7747047 0.2055381 0.1563947 0.2778705

**Interpretation:**

Based on Mcfadden R Square, we conclude that **20.55%**of the uncertainty of the inetercept only model has been explained by the full model.

Thus the goodness of fit is **good and is accepted.**

1. **The significance of the individual co-efficients:**

The Significance of the individual co-efficients can be assessed using the summary function.

summary(logit)

Call:

glm(formula = Churn ~ AccountWeeks + ContractRenewal + DataPlan +

DataUsage + CustServCalls + DayMins + DayCalls + MonthlyCharge +

OverageFee + RoamMins, family = binomial, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0417 -0.5097 -0.3504 -0.2088 3.0459

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.1567684 0.6674321 -9.225 < 2e-16 \*\*\*

AccountWeeks -0.0001507 0.0016498 -0.091 0.92722

ContractRenewalRenewed -1.9669347 0.1762725 -11.158 < 2e-16 \*\*\*

DataPlanYes -0.6793909 0.6385085 -1.064 0.28732

DataUsage 1.3441189 2.3006497 0.584 0.55906

CustServCalls 0.5529267 0.0468420 11.804 < 2e-16 \*\*\*

DayMins 0.0360104 0.0387869 0.928 0.35319

DayCalls 0.0054809 0.0032975 1.662 0.09648 .

MonthlyCharge -0.1399872 0.2280180 -0.614 0.53926

OverageFee 0.4001170 0.3896952 1.027 0.30454

RoamMins 0.0715722 0.0265797 2.693 0.00709 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1930.4 on 2332 degrees of freedom

Residual deviance: 1533.6 on 2322 degrees of freedom

AIC: 1555.6

Number of Fisher Scoring iterations: 5

**Interpretation:**

It is seen that the variables which is significant in explaining the customer churn are **ContractRenewal, CustServCalls and the RoamMins.**

1. **Explanatory power of the odds and the probabilities:**

The explanatory power of the odds and the probabilities help in understanding the practical significance of the variables in addition to the statistical significance.

**Odds:**

exp(coef(logit))

(Intercept) AccountWeeksContractRenewalRenewed

0.00211909 0.99984932 0.13988499

DataPlanYes **DataUsage CustServCalls**

0.50692565 **3.83480618 1.73833314**

**DayMins DayCalls** MonthlyCharge

**1.03666667 1.00549594** 0.86936938

**OverageFee RoamMins**

**1.49199932 1.07419566**

**Probabilities:**

exp(coef(logit))/(1+exp(coef(logit)))

(Intercept) AccountWeeksContractRenewalRenewed

0.002114609 0.499962327 0.122718509

DataPlanYes **DataUsage CustServCalls**

0.336397253 **0.793166476 0.634814338**

**DayMins DayCalls** MonthlyCharge

**0.509001637 0.501370221** 0.465060245

**OverageFee RoamMins**

**0.598715781 0.517885406**

**Interpretation:**

* From the odds and the probabilities,it is clearly seen that the variable **Data Usage** is more significant than all other variables.
* The variables, **RoamMins and CustServCalls**which are **statistically significant** are also significant in terms of odds and probabilities.
* The other variables which are significant are **DayMins, OverageFee, and DayCalls.**
* It is also important to note that the variable **ContractRenewal** which was **statistically important** was **not important** in terms of odds and probabilities.

1. **Confusion Matrix:**

The accuracy of the training data is obtained.

Prediction<-predict(logit,type="response")

cutoff<-floor(Prediction+0.5)

table(Actual=train$Churn,Predicted=cutoff)

Predicted

Actual 0 1

0 1781 36

1 254 50

##Accuracy

(1781+50)/2121

[1] 0.86327204

The accuracy of the model is **86.33%.**

1. **Prediction using test data:**

The test data is used to check if the model is able to provide the same accuracy.

Prediction1<-predict(logit,type="response",newdata = test)

cutoff1<-floor(Prediction1+0.5)

table(Actual=test$Churn,Predicted=cutoff1)

Predicted

Actual 0 1

0 999 34

1 140 39

##Accuracy

(999+39)/1212

[1] 0.85653465

The accuracy obtained through the test data is **85.65%**.

1. **ROC Curve:**

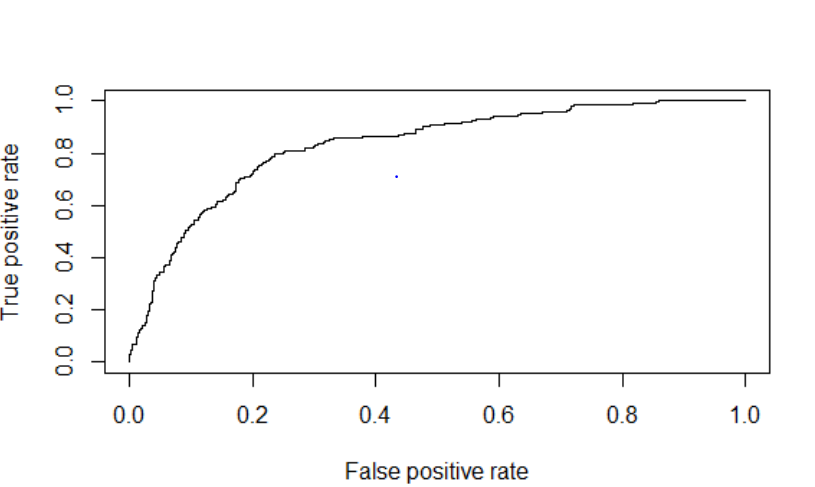
library(ROCR)

roc <- predict(model, newdata=test, type="response")

pr<- prediction(roc, test$Churn)

prf<- performance(pr, measure = "tpr", x.measure = "fpr")

plot(prf)



auc<- performance(pr, measure = "auc")

auc<- auc@y.values[[1]]

auc

[1] 0.8312503

1. **R Code for the problem:**

setwd("C:\\Users\\user\\Desktop\\R Programming")

getwd()

data = read.csv("Dataset\_Cellphone.csv",header = TRUE)

str(data)

names(data)

table(data$Churn)

##Convert the factor variables

data$ContractRenewal<-factor(data$ContractRenewal,levels=c("0","1"),labels=c("Non-Renewed","Renewed"))

data$DataPlan<-factor(data$DataPlan,levels=c("0","1"),labels=c("No","Yes"))

##Splitting data into training and testing data

##install.packages("caret")

##install.packages("lattice")

##install.packages("ggplot2")

##install.packages("caTools")

library("caTools")

library("caret")

library("lattice")

library("ggplot2")

set.seed(1234)

split<-sample.split(data$Churn, SplitRatio = 0.70)

train<-subset(data, split == TRUE)

test<-subset(data, split == FALSE)

table(train$Churn)

table(test$Churn)

c(nrow(train),nrow(test))

nrow(data)

str(train)

##logit model

logit = glm(Churn ~ AccountWeeks + ContractRenewal + DataPlan

+ DataUsage + CustServCalls + DayMins + DayCalls

+ MonthlyCharge + OverageFee + RoamMins , data=train,family=binomial)

##Step 1:Overall validity of the model

library(lmtest)

lrtest(logit)

##Step 2:McFadden R Squared

library(pscl)

pR2(logit)

##Step 3:Individual coefficients significance

summary(logit)

##Step 4:Explanatory Power of odds and probabilities

##odds

exp(coef(logit))

##prob

exp(coef(logit))/(1+exp(coef(logit)))

##Step 5:Confusion Matrix

Prediction<-predict(logit,type="response")

cutoff<-floor(Prediction+0.5)

table(Actual=train$Churn,Predicted=cutoff)

##Accuracy

(1781+50)/2121

##test using the test data

Prediction1<-predict(logit,type="response",newdata = test)

cutoff1<-floor(Prediction1+0.5)

table(Actual=test$Churn,Predicted=cutoff1)

##Accuracy

(999+39)/1212

##Step 6:ROC Curve

library(ROCR)

roc <- predict(model, newdata=test, type="response")

pr<- prediction(roc, test$Churn)

# TPR = sensitivity, FPR=specificity

prf<- performance(pr, measure = "tpr", x.measure = "fpr")

plot(prf)

auc<- performance(pr, measure = "auc")

auc<- auc@y.values[[1]]

auc