Lab06Group05TummuruGriffin

1. Hypotheses:

- Accountweeks has no relation with CancelledService as logically services can be cancelled due to any reasons like price increase.

- RecentRenewal has a direct negative relationship with CancelledService as the customers who recently renew accounts would presumably be less likely to cancel the services.

- DataPlan has a direct negative relationship with CancelledService as the customers who have a data plan should be more invested in the service and less likely to cancel.

- DataUsage has a direct positive relationship with CancelledService as customers with higher data usage are more likely to cancel the services due to overage fees.

- CustServCalls has a direct positive relationship with CancelledService as customers who make more customer service calls are more likely to have problems with the service, and thus are more likely to cancel.

- AvgCallMinsPerMonth has a direct negative relationship with CancelledService as customers who use the service more are more likely to depend on the service and are thus less likely to cancel the service.

- AvgCallsPerMonth has a direct negative relationship with CancelledService as customers who are using the service more are presumably more satisfied with the service and thus less likely to cancel.

- MonthlyBill has a direct positive relationship with CancelledService as customers with higher monthly bills are more likely to cancel their services due to an inability to keep up with the costs.

- OverageFee has a direct positive relationship with CancelledService as customers who receive more overage fees are more likely to cancel their service and move to a network with fewer overage fees.

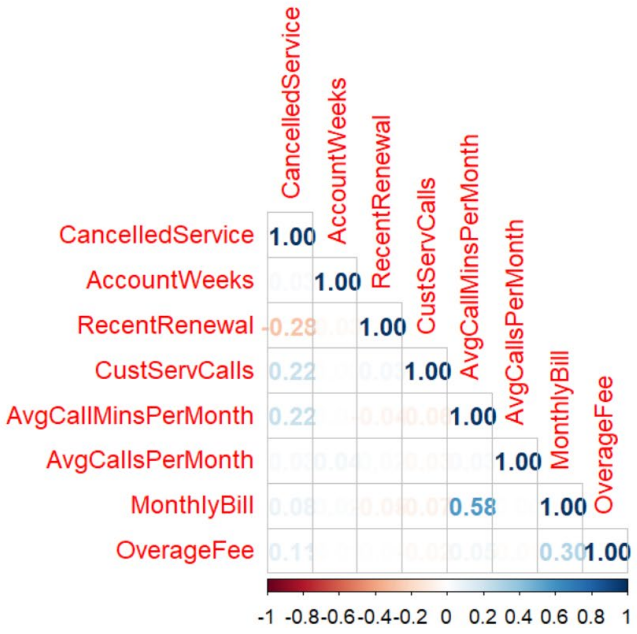
2. R Code:

```
# Vihari Reddy Tummuru & Chace Griffin
# MIS 545-001
# Lab06Group05TummuruGriffin.R
# This code reads and summarizes a CSV file, creates and displays a histogram
# for the data, produces a correlation matrix and plot, tests the data for class
# imbalance, uses the SMOTE technique to deal with class imbalance, generates
# a logistic regression model and odds ratios, predicts outcomes in the testing
# dataset, generates a confusion matrix, and calculates confusion model accuracy,
# false positive, and false negative rates

# Install tidyverse package
install.packages("tidyverse")

# Load tidyverse package
library(tidyverse)

# Install corrplot package
install.packages("corrplot")

# Load corrplot package
library(corrplot)

# Install olsrr package
install.packages("olsrr")

# Load olsrr package
library(olsrr)

# Install smotefamily package
install.packages("smotefamily")

# Load smotefamily package
library(smotefamily)

# Set working directory to Lab06 folder
setwd("C:/Users/ual-laptop/Documents/MIS545/Lab06")

# Read MobilePhoneSubscribers.csv file into a tibble and defining column types
mobilePhone <- read_csv(file = "MobilePhoneSubscribers.csv",
                        col_types = "lillnininn",
                        col_names = TRUE)

# Display the mobilePhone tibble in the console
print(mobilePhone)

# Display the structure of the tibble
print(str(mobilePhone))

# Display the summary of the mobilephone tibble in the console
print(summary(mobilePhone))

# Converts all columns to numeric and creates a histogram for the data
displayAllHistograms <- function(tibbleDataset) {
  tibbleDataset %>%
    keep(is.numeric) %>%
    gather() %>%
    ggplot() + geom_histogram(mapping = aes(x=value,fill=key),
                              color = "black")+
    facet_wrap( ~ key,scales= "free")+
    theme_minimal()
}
```

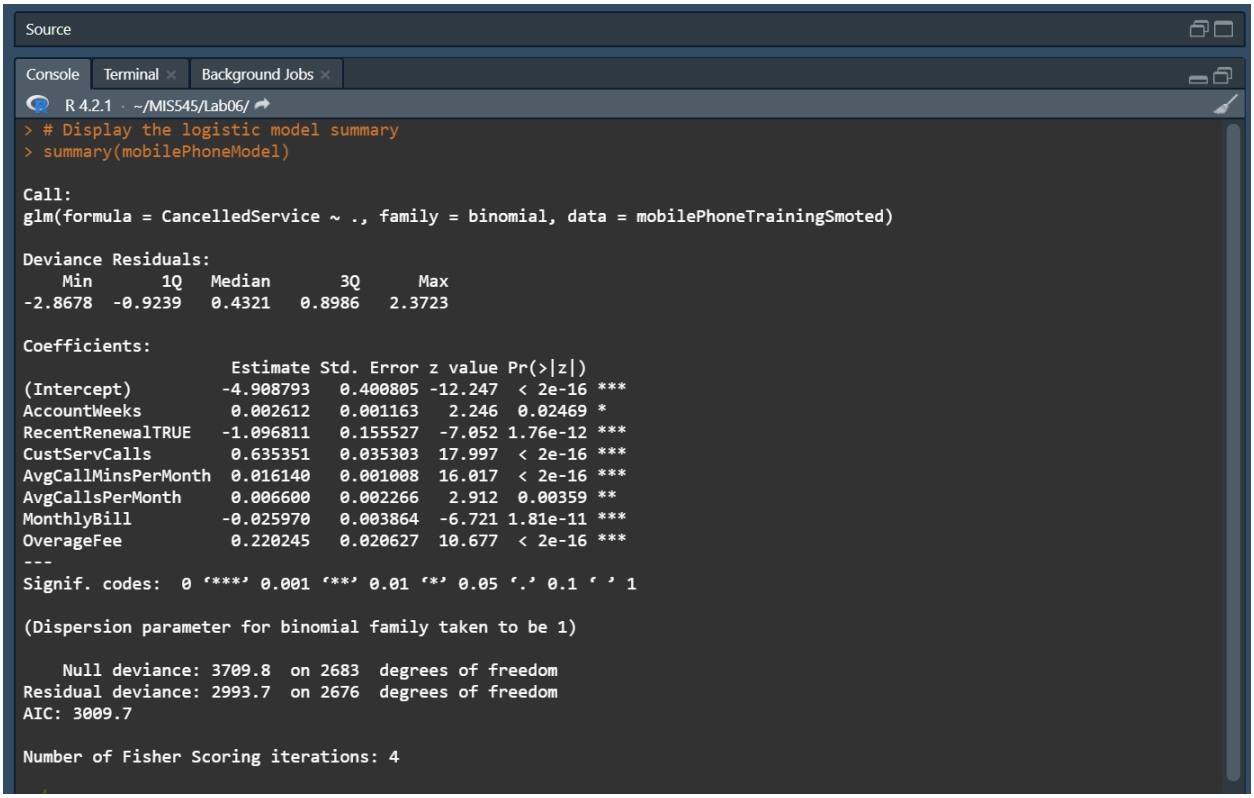
```
# Display the histogram of the mobilephone tibble
displayAllHistograms(mobilePhone)

# Display a correlation matrix of the mobilePhone tibble and round the coefficient
# to 2 digits after decimal
round(cor(mobilePhone),digits = 2)

# Display of a correlation plot of mobilePhone with number
corrplot(cor(mobilePhone),
         method = "number",
         type = "lower")

# Remove data plan and data usage columns from the mobilePhone data
mobilePhone<- subset(mobilePhone, select = -c(DataPlan,DataUsage))

# Splitting mobilephone data into training dataset and test data set based on IQR
set.seed(203)
sampleSet <- sample(nrow(mobilePhone),
                    round(nrow(mobilePhone)*.75),
                    replace = FALSE)

mobilePhoneTraining <- mobilePhone[sampleSet, ]
mobilePhoneTest <- mobilePhone[-sampleSet, ]

# Checking for class imbalance
summary(mobilePhoneTraining$CancelledService)

# SMOTE
mobilePhoneTrainingSmoted <- tibble(SMOTE(
  X=data.frame(mobilePhoneTraining),
  target = mobilePhoneTraining$CancelledService,
  dup_size = 3)$data)

summary(mobilePhoneTrainingSmoted)

# Convert cancelled service and recent renewal to logical
mobilePhoneTrainingSmoted <- mobilePhoneTrainingSmoted %>%
  mutate(CancelledService = as.logical(CancelledService),
         RecentRenewal = as.logical(RecentRenewal))

summary(mobilePhoneTrainingSmoted)

# Removing "class" column in tibble
mobilePhoneTrainingSmoted <- mobilePhoneTrainingSmoted %>%
  select(-class)

# Check for class imbalance in the training set
summary(mobilePhoneTrainingSmoted)

# Generate logistic regression Model
mobilePhoneModel<- glm(data=mobilePhoneTrainingSmoted, family=binomial,
                       formula=CancelledService ~ .)

# Display the logistic model summary
summary(mobilePhoneModel)

# Odds ratios of the independent variables
exp(coef(mobilePhoneModel)["AccountWeeks"])
exp(coef(mobilePhoneModel)["RecentRenewalTRUE"])
exp(coef(mobilePhoneModel)["CustServCalls"])
exp(coef(mobilePhoneModel)["AvgCallMinsPerMonth"])
exp(coef(mobilePhoneModel)["AvgCallsPerMonth"])
```

```
exp(coef(mobilePhoneModel)["MonthlyBill"])
exp(coef(mobilePhoneModel)["OverageFee"])

# Use the model to predict outcomes in the testing dataset
mobilePhonePrediction <- predict(mobilePhoneModel,
                                 mobilePhoneTest,
                                 type='response')

# Display the test model
print(mobilePhonePrediction)

# Converting less than 0.5 as 0 and greater than 0.5 as 1 using IF Else
mobilePhonePrediction <-
  ifelse(mobilePhonePrediction >= 0.5,1,0)

# Generating a mobile phone confusion matrix
mobilePhoneConfusionMatrix <- table(mobilePhoneTest$CancelledService,
                                    mobilePhonePrediction)

# Display confusion matrix (mobilePhoneConfusionMatrix)
print(mobilePhoneConfusionMatrix)

# Calculating false positive for confusion matrix
mobilePhoneConfusionMatrix[1,2]/
  (mobilePhoneConfusionMatrix[1,2]+mobilePhoneConfusionMatrix[1,1])

# Calculating false negative for confusion matrix
mobilePhoneConfusionMatrix[2,1]/
  (mobilePhoneConfusionMatrix[2,1]+mobilePhoneConfusionMatrix[2,2])

# Calculating Model Prediction Accuracy
sum(diag(mobilePhoneConfusionMatrix))/ nrow(mobilePhoneTest)
```

Correlation Plot:

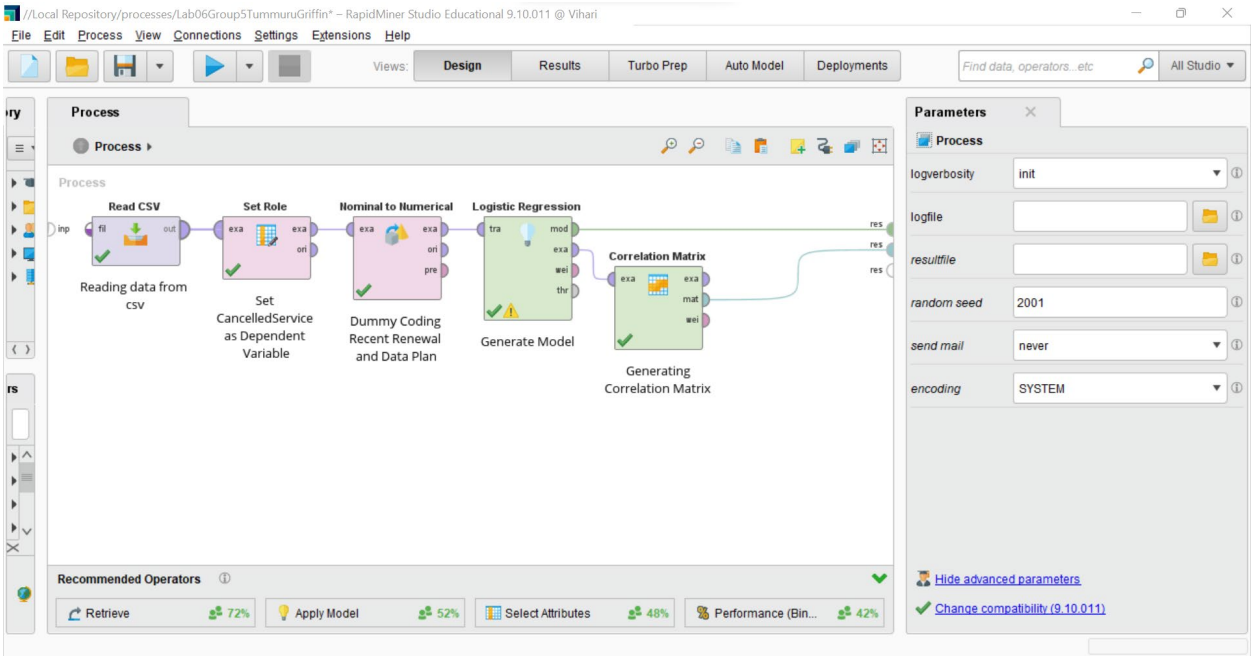


Logistic Model Summary:

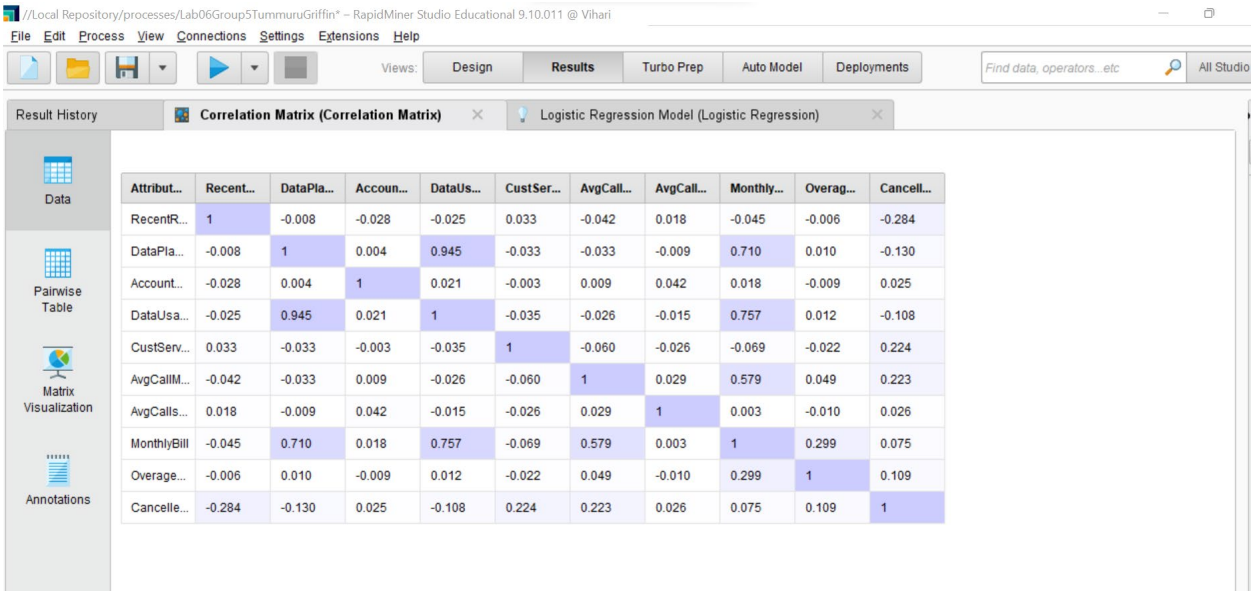


3. Rapid Miner:

Process:

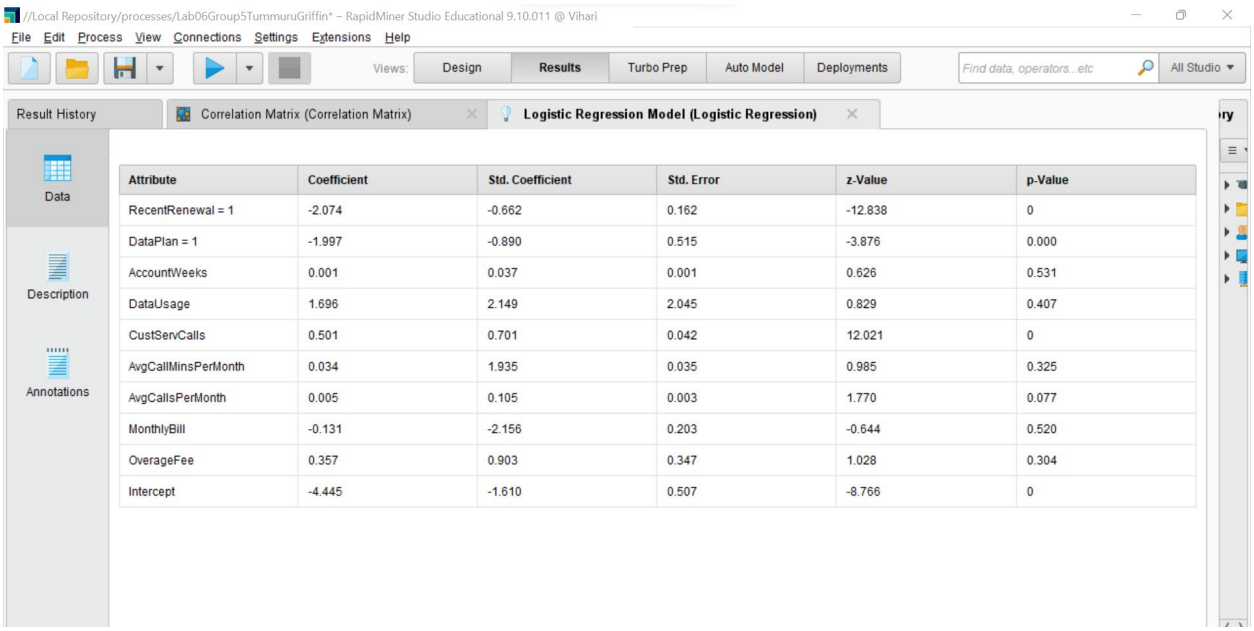


Correlation Matrix:



Attribut...	Recent...	DataPla...	Accoun...	DataUs...	CustSer...	AvgCall...	AvgCall...	Monthly...	Overag...	Cancell...
RecentR...	1	-0.008	-0.028	-0.025	0.033	-0.042	0.018	-0.045	-0.006	-0.284
DataPla...	-0.008	1	0.004	0.945	-0.033	-0.033	-0.009	0.710	0.010	-0.130
Account...	-0.028	0.004	1	0.021	-0.003	0.009	0.042	0.018	-0.009	0.025
DataUsa...	-0.025	0.945	0.021	1	-0.035	-0.026	-0.015	0.757	0.012	-0.108
CustServ...	0.033	-0.033	-0.003	-0.035	1	-0.060	-0.026	-0.069	-0.022	0.224
AvgCallM...	-0.042	-0.033	0.009	-0.026	-0.060	1	0.029	0.579	0.049	0.223
AvgCalls...	0.018	-0.009	0.042	-0.015	-0.026	0.029	1	0.003	-0.010	0.026
MonthlyBill	-0.045	0.710	0.018	0.757	-0.069	0.579	0.003	1	0.299	0.075
Overage...	-0.006	0.010	-0.009	0.012	-0.022	0.049	-0.010	0.299	1	0.109
Cancelle...	-0.284	-0.130	0.025	-0.108	0.224	0.223	0.026	0.075	0.109	1

Model Results:



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
RecentRenewal = 1	-2.074	-0.662	0.162	-12.838	0
DataPlan = 1	-1.997	-0.890	0.515	-3.876	0.000
AccountWeeks	0.001	0.037	0.001	0.626	0.531
DataUsage	1.696	2.149	2.045	0.829	0.407
CustServCalls	0.501	0.701	0.042	12.021	0
AvgCallMinsPerMonth	0.034	1.935	0.035	0.985	0.325
AvgCallsPerMonth	0.005	0.105	0.003	1.770	0.077
MonthlyBill	-0.131	-2.156	0.203	-0.644	0.520
OverageFee	0.357	0.903	0.347	1.028	0.304
Intercept	-4.445	-1.610	0.507	-8.766	0

4. Which, if any, of your predictions were incorrect. Explain why this might be the case.

Our prediction that MobileBill would have a direct positive relationship with cancelled services was incorrect as the relationship between the two variables is negative. This may be because people with lower bills are more likely to have lower incomes, and thus are actually more likely to have to cancel services due to financial troubles.

5. Why is DataPlan highly correlated with DataUsage? Answer the question *logically*, not by simply stating that they have a high pairwise correlation.

This is because people with data plans may be restricted to only using mobile data, and thus would use large quantities of data. Conversely, people without data plans would use lower amounts of data because it might result in overage fees.

6. Why is MonthlyBill highly correlated with DataPlan and DataUsage?

Because there is always a higher cost associated with having a data plan and higher data usage, these two variables would usually result in a higher monthly bill. In contrast, not having a data plan and not using data would typically result in a lower monthly bill.