MIS 545 Lab 06 Logistic Regression

Part 1: Hypotheses

AccountWeeks: The higher the duration for which the user has been with the mobile phone plan, the lower is the risk of attrition. Thus, this will have an indirect relation.

RecentRenewal: If the user has renewed the plan recently, chances of attrition are lower. Thus, this will have an indirect relation.

DataPlan: If the user has a data plan, the chances of attrition are lower. Thus, this will have an indirect relation.

DataUsage: The higher the data usage by the user has been with the lower is the risk of attrition. Thus, this will have an indirect relation.

CustServCalls: The higher the number of customer service calls by the user, the risk of attrition as multiple calls is an indication of poor user experience/service. Thus, this will have a direct relation.

AvgCallMinsPerMonth: The higher the duration (no. of mins) for which the user is on call per month, the lower is the risk of attrition as the long duration is indicative of a good user experience/service to the user.

Thus, this will have an indirect relation.

AvgCallsPerMonth: This will have no relation with the service as the number of calls can be higher or lower based on the user's requirement and has nothing to do with the mobile plan.

MonthlyBill: The higher the monthly bill of the user, the higher is the risk of attrition as the user will try to find cheaper substitutes. Thus, this will have an direct relation.

MIS 545 Lab 06 Logistic Regression

OverageFee: The higher the overage fee paid by the user, the higher is the risk of attrition as the user will try to find cheaper substitutes. Thus, this will have a direct relation.

Part 2: R Code

```
# Prajakta Tambe, Varun Kapuria
# MIS 545 Section 01
# Lab06TambeKapuria.R
# The following code is to determine the chances of a user cancelling
# their mobile service based on several factors using logistic regression

# Install the tidyverse, corrplot, olsrr, and smotefamily packages
# install.packages("tidyverse")
# install.packages("corrplot")
# install.packages("olsrr")
# install.packages("smotefamily")

# Load the tidyverse, corrplot, olsrr, and smotefamily libraries
library("tidyverse")
library("corrplot")
library("olsrr")
library("smotefamily")

# Set the working directory to your Lab06 folder
setwd("C:/Users/ual-laptop/Desktop/MIS545/Week6/Lab06")
getwd()

# Read MobilePhoneSubscribers.csv into a tibble called mobilePhone
mobilePhone <- read_csv(file = "MobilePhoneSubscribers.csv",
                        col_types = "lillninnn",
                        col_names = TRUE)

# Display mobilePhone in the console
print(mobilePhone)

# Display the structure of mobilePhone in the console
str(mobilePhone)


# Display the summary of mobilePhone in the console
summary(mobilePhone)

# Recreate the displayAllHistograms() function as shown in a prior video
```

MIS 545 Lab 06 Logistic Regression

```
# demonstration
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()
}

# Call the displayAllHistograms() function, passing in mobilePhone as an
# argument
displayAllHistograms(mobilePhone)

# Display a correlation matrix of mobilePhonerounded to two decimal places
mobilePhoneRounded <- round(cor(mobilePhone),2)

# Display a correlation plot using the "number" method and limit output to the
# bottom left
corrplot(cor(mobilePhone),
         method = "number",
         type = "lower")

# The correlation plot should reveal three pairwise correlations that are above
# the threshold of 0.7. Remove the data plan and data usage variables from the
# tibble
mobilePhone <- mobilePhone %>%
  select(-c(DataUsage,DataPlan))

# Randomly split the dataset into mobilePhoneTraining (75% of records)
# and mobilePhoneTesting (25% of records) using 203 as the random seed
set.seed (203)
sampleMobilePhoneSet <- sample(nrow(mobilePhone),
                               round(nrow(mobilePhone)*0.75),
                               replace = FALSE)
mobilePhoneTraining <- mobilePhone[sampleMobilePhoneSet, ]
mobilePhoneTesting <- mobilePhone[-sampleMobilePhoneSet, ]
summary(mobilePhoneTraining)

# Check if we have a class imbalance issue in CancelledService
summary(mobilePhoneTraining$CancelledService)
classImbalanceMagnitude <- 1253 / 360

# Deal with class imbalance using the SMOTE technique
```

MIS 545 Lab 06 Logistic Regression

```r
# using a duplicate size of 3. Save the result into a
# new tibble called mobilePhoneTrainingSmoted
mobilePhoneSmoted <-
  tibble(SMOTE(X = data.frame(mobilePhoneTraining),
               target = mobilePhoneTraining$CancelledService,
               dup_size = 3)$data)
summary(mobilePhoneSmoted)

# Convert CancelledService and RecentRenewal back into logical types
mobilePhoneSmoted <- mobilePhoneSmoted %>%
  mutate(CancelledService = as.logical(CancelledService),
         RecentRenewal = as.logical(RecentRenewal))

# Get rid of the "class" column in the tibble
mobilePhoneSmoted <- mobilePhoneSmoted %>%
  select(-class)
summary(mobilePhoneSmoted)

# Generate the logistic regression model using CancelledService as the
# binary dependent variable and save it in an object called mobilePhoneModel
mobilePhoneModel <- glm(data = mobilePhoneSmoted,
                        family = binomial,
                        formula = CancelledService ~ .)

# Display the logistic regression model results using the summary() function
summary(mobilePhoneModel)

# Calculate the odds ratios for each of the 7 independent variable coefficients
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
# Treating anything below or equal to 0.5 as a 0, anything above 0.5 as a 1.
mobilePhonePrediction <- predict(mobilePhoneModel,
                                 mobilePhoneTesting,
                                 type = "response")
print(mobilePhonePrediction)
mobilePhonePrediction <-
  ifelse(mobilePhonePrediction >= 0.5, 1, 0)
```

```
# Generate a confusion matrix of predictions
mobilePhoneConfusionMatrix <- table(mobilePhoneTesting$CancelledService,
                                    mobilePhonePrediction)
print(mobilePhoneConfusionMatrix)

# Calculate the false positive rate
mobilePhoneConfusionMatrix[1,2] /
  (mobilePhoneConfusionMatrix[1,2] +
     mobilePhoneConfusionMatrix[1,1])

# Calculate the false negative rate
mobilePhoneConfusionMatrix[2,1] /
  (mobilePhoneConfusionMatrix[2,1] +
     mobilePhoneConfusionMatrix[2,2])

# Calculate the model prediction accuracy
sum(diag(mobilePhoneConfusionMatrix))/nrow(mobilePhoneTesting)
```

Part 3: RapidMiner Screenshots

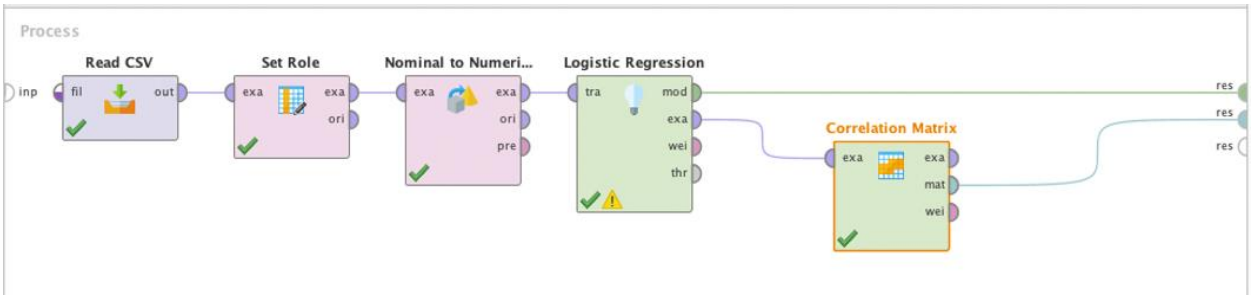

Figure 1: Screenshot of Process

MIS 545 Lab 06 Logistic Regression

Attribu...	Recent...	DataPl...	Accoun...	DataUs...	CustSe...	AvgCal...	AvgCal...	Monthl...	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
DataUs...	-0.025	0.945	0.021	1	-0.035	-0.026	-0.015	0.757	0.012	-0.108
CustSer...	0.033	-0.033	-0.003	-0.035	1	-0.060	-0.026	-0.069	-0.022	0.224
AvgCall...	-0.042	-0.033	0.009	-0.026	-0.060	1	0.029	0.579	0.049	0.223
AvgCall...	0.018	-0.009	0.042	-0.015	-0.026	0.029	1	0.003	-0.010	0.026
Monthly...	-0.045	0.710	0.018	0.757	-0.069	0.579	0.003	1	0.299	0.075
Overag...	-0.006	0.010	-0.009	0.012	-0.022	0.049	-0.010	0.299	1	0.109
Cancell...	-0.284	-0.130	0.025	-0.108	0.224	0.223	0.026	0.075	0.109	1

Figure 2: Correlation Matrix

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	0.0417029108493529	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

Figure 3: Logistic Regression Model

MIS 545 Lab 06 Logistic Regression

Part 4: From R code

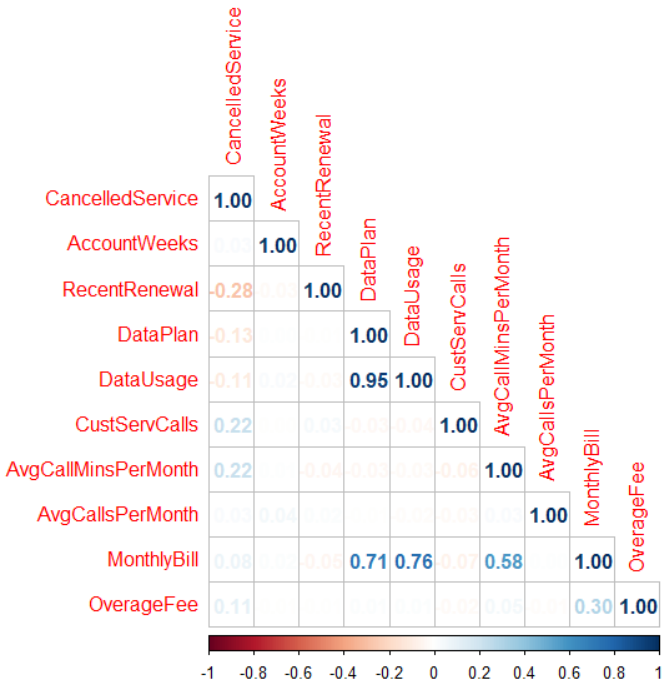

Figure 4: Correlation Plot

```
> summary(mobilePhoneModel)

Call:
glm(formula = CancelledService ~ ., family = binomial, data = mobilePhoneSmoted)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8678  -0.9239   0.4321   0.8986   2.3723

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.908793   0.400805 -12.247  < 2e-16 ***
AccountWeeks        0.002612   0.001163   2.246  0.02469 *
RecentRenewalTRUE  -1.096811   0.155527  -7.052 1.76e-12 ***
CustServCalls       0.635351   0.035303  17.997  < 2e-16 ***
AvgCallMinsPerMonth 0.016140   0.001008  16.017  < 2e-16 ***
AvgCallsPerMonth    0.006600   0.002266   2.912  0.00359 **
MonthlyBill        -0.025970   0.003864  -6.721 1.81e-11 ***
OverageFee          0.220245   0.020627  10.677  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3709.8  on 2683  degrees of freedom
Residual deviance: 2993.7  on 2676  degrees of freedom
AIC: 3009.7

Number of Fisher Scoring iterations: 4
```

Figure 5: Summary of the Logistic Model

MIS 545 Lab 06 Logistic Regression

Part 5: Answers after model creation

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

AvgCallsPerMonth was one of the predictions that were incorrect. We originally thought that this would have no significant relation in the model but according to the correlation matrix it has a direct impact with 0.03. Which means, higher number of calls indicate that the customer is happy with the network and would not cancel their service. AccountWeeks and AvgCallMinsPerMonth is also directly correlated to CancelledServiceTRUE, We're not sure why that must be.

Why is DataPlan highly correlated with DataUsage?

Users with higher the DataPlan would mean higher the DataUsage.

Why is MonthlyBill highly correlated with DataPlan and DataUsage?

Usually, higher data plans are priced more than lower data plans which means that monthly bill would be higher. Thus, higher DataPlan means their DataUsage is more which also means MonthlyBill would be higher.