Logistic Regression – Performance Assessment

Steven Schindler

Predictive Modeling – D208

Straw, Eric; PhD

College of I.T., Western Governors University

Table of Contents

Part I: Research Question *3*

A1. ..... **3**

A2. ..... **3**

Part II: Method Justification *4*

B1. ..... **4**

B2. ..... **4**

B3. ..... **5**

Part III: Data Preparation *5*

C1. ..... **5**

C2. ..... **7**

C3. ..... **8**

..... **9**

..... **9**

..... **9**

..... **10**

..... **11**

..... **11**

..... **11**

..... **12**

C4. ..... **12**

C5. ..... **13**

Part IV: Model Comparison and Analysis *13*

D1. ..... **13**

D2. ..... **13**

D3. ..... **14**

..... **15**

E1. ..... **15**

E2. ..... **16**

E3. ..... **16**

Part V: Data Summary and Implications *20*

F1 ..... **20**

F2 ..... **21**

References *21*
H. 21
I. 21

Part I: Research Question

A1.

For my research question I am working with the Churn dataset. My research question is “What variables contribute to a customer that will churn?” Churn is a binomial categorical variable on whether the customer has discontinued service in the last month.

A2.

The goal of this analysis is to try to predict a reasonable cause of Churn using 11 explanatory variables. There are 11 ponteential variables that could influence churn 8 numeric variables and 3 categorical variables. The 11 explanatory variables are Outage_sec_perweek, Yearly_equip_failure, Income, Email, Contacts, Techie, Contract, Multiple, Tenure, MonthlyCharge, and Bandwidth_GB_Year.

For Outage_sec_perweek, which is the average number of seconds for system outages in the customers neighborhood, is there a correlation for high outages and Churn? Yearly_equip_failure is the number of times the customers equipment failed in the past year, is there a correlation for high number of equipment failure and Churn? Income, which is the annual salary of a customer per year, could there be a correlation between Income and Churn based on amounts?

Email, which is the amount of emails sent to the customer, does the amount emails affect churn? Number of times the customer contacted customer support is Contacts, is there a correlation? Techie is defined as whether the customer

considers themselves good with technology, does a techie stay with the company or not? Contract is the contract term made with the customer, does the type of contract affect churn? Does having Multiple lines lead to not churning? Does having a longer Tenure with the company correlate to not churning? MonthlyCharge and Bandwidth_GB_Year are the final two explanatory variables. MonthlyCharge is how much the customer is billed per month and Bandwidth_GB_Year is how much data they us in a year, is there a correlation between them and churn respectively?

Part II: Method Justification

B1.

An assumption is that logistic regression requires a large sample size. A second assumption is that the predicted values are nominal values such as ‘yes’ or ‘no’. The third assumption is there is little to no multicollinearity among independent variables. Finally, it assumes that independent variables are linearly related to the log odds.(Statistics Solutions. (n.d.))

B2.

I used both Python and R in this performance assessment. I used Python to clean the data then I used R for the regression model. Python has many useful libraries that make it ideal to clean data such as pandas, numpy and stats. These libraries make it easy to find missing values, replace trailing and leading spaces as well as easily calculate Z-scores and detect outliers. Python also has easy readability of it’s code.

I used R for the regression model because it too has many useful libraries such as pyler,dpyler,ggplot2 and car. R also has many built in functions that make it ideal working with data to get statistics. It provides summary statistics and with

ggplot2 graphs are made easily to visualize the data. It also provides simple functions such as the glm function to calculate logistic regression easier.

B3.

Churn is a binomial categorical variable consisting of values of yes and no therefore, logistical regression is not only appropriate but required. The research question also wants probability which is one of the assumptions for logistic regression.

Part III: Data Preparation

C1.

My data cleaning goals are to check for missing values, detect and treat outliers if appropriate and check for leading and trailing spaces in all the variables mentioned in A2. I used Python to clean the data and start by checking the number of missing values in the 12 variables plus CaseOrder as a unique ID variable. I see that each variable has zero missing values. I then strip leading and trailing spaces of all string variables. Finally, I display the max and min of each numeric type as well as the number of outliers using the Z-score. I believe all the outliers are legitimate data, so I do not remove any, I also want to see the effects of large and small numbers have on churn as another reason to not remove any. My Python code is as follows:

```
# import the numpy and pandas libraries
import numpy as np
import pandas as pd
import string as str
from scipy import stats

#read in the churn data set as a pandas dataframe
df = pd.read_csv('churn_clean.csv')
```

```
#new dataframe for the variables that are needed and the CaseOrder variable for identification.
new_df = pd.DataFrame().assign(CaseOrder=df['CaseOrder'],
                               Churn=df['Churn'],
                               Outage_sec_perweek=df['Outage_sec_perweek'],
                               Yearly_equip_failure=df['Yearly_equip_failure'],
                               Income=df['Income'],
                               Email=df['Email'],
                               Contacts=df['Contacts'],
                               Techie=df['Techie'],
                               Contract=df['Contract'],
                               Multiple=df['Multiple'],
                               Tenure=df['Tenure'],
                               MonthlyCharge=df['MonthlyCharge'],
                               Bandwidth_GB_Year=df['Bandwidth_GB_Year']
                              )

new_df.isna().sum()

#strips leading and trailing spaces
string_list = list(new_df.select_dtypes(include = {'object'}))
for i in string_list:
   new_df[i] = new_df[i].str.strip()

col_num_names =
['Outage_sec_perweek','MonthlyCharge','Yearly_equip_failure','Income',
            'Bandwidth_GB_Year','Tenure','Email','Contacts']


#Find min max of the numeric datatypes except CaseOrder as well as number of outliers.
for i in col_num_names:
   #check for outliers using zscore
   new_df_count = new_df[(np.abs(stats.zscore(new_df[i])) < 3)].count()
   print(new_df[i].name,"number of outliers is ",10000 - new_df_count[i], ",\n")
   print(new_df[i].name,"min is ", new_df[i].min(),"max is ",new_df[i].max(),"\n")

new_df.to_csv('/Users/stevenschindler/Documents/R/D208/clean_churn_file_log.csv', index=False)
```

C2.

```
> summary(df$Churn)
   Length     Class      Mode
    10000 character character
> summary(df$Outage_sec_perweek)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.09975  8.01821 10.01856 10.00185 11.96949 21.20723
> summary(df$Yearly_equip_failure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.398   1.000   6.000
> summary(df$Income)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   348.7  19224.7  33170.6  39806.9  53246.2 258900.7
> summary(df$Email)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   10.00   12.00   12.02   14.00   23.00
> summary(df$Contacts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.9942  2.0000  7.0000
> summary(df$Techie)
   Length     Class      Mode
    10000 character character
> summary(df$Contract)
   Length     Class      Mode
    10000 character character
> summary(df$Multiple)
   Length     Class      Mode
    10000 character character
> summary(df$Tenure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   7.918  35.431  34.526  61.480  71.999
> summary(df$MonthlyCharge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  79.98  139.98  167.48  172.62  200.73  290.16
> summary(df$Bandwidth_GB_Year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  155.5  1236.5  3279.5  3392.3  5586.1  7159.0
```

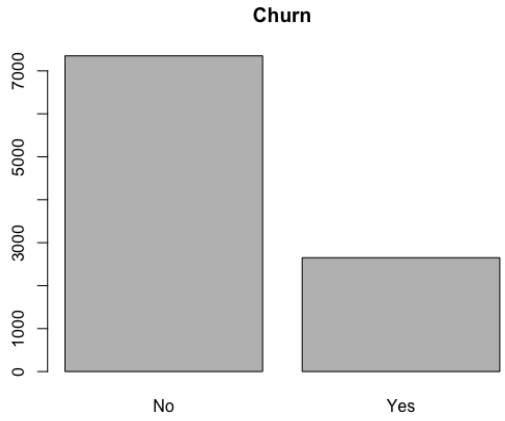
Above is summary output of both the dependent variable and all explanatory variables. Churn, Techie, Contract and Multiple are all string type variables so their summary only consists of their class and frequency. Outage_sec_perweek has

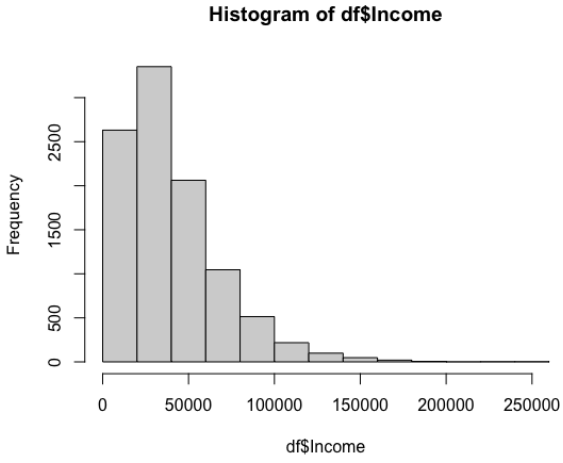
a median of 10.018 and a mean of 10.0018. Yearly_equip_failure has a median of 0 and a mean of .398. Income has a median of 33,170.6 and a mean of 39,806.9.

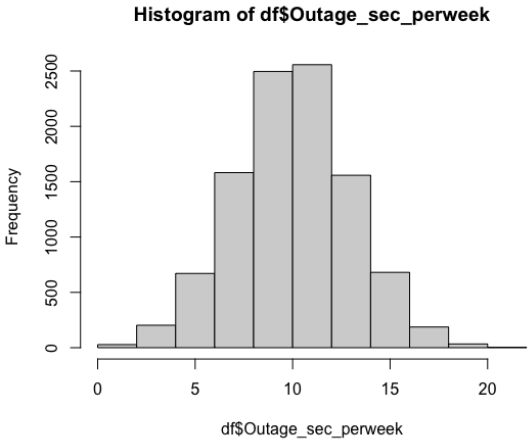
Email has a median of 12 and a mean of 12.02. Contacts has a median of 1 and a mean of 0.9942. Tenure has a median of 35.431 and a mean of 34.526. MonthlyCharge has a median of 167.48 and a mean of 172.62. Finally, Bandwidth_GB_year has a median of 3279.5 and a mean of 3392.3.

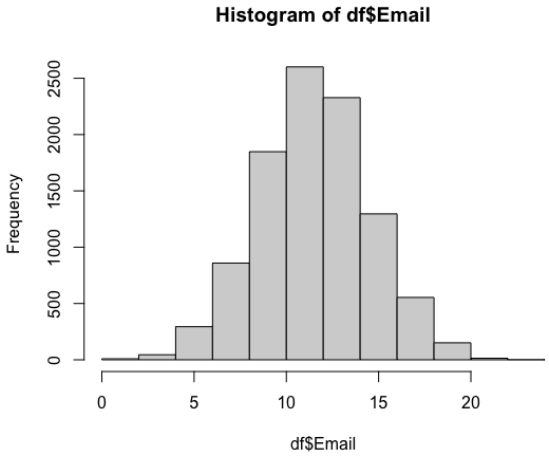
C3.

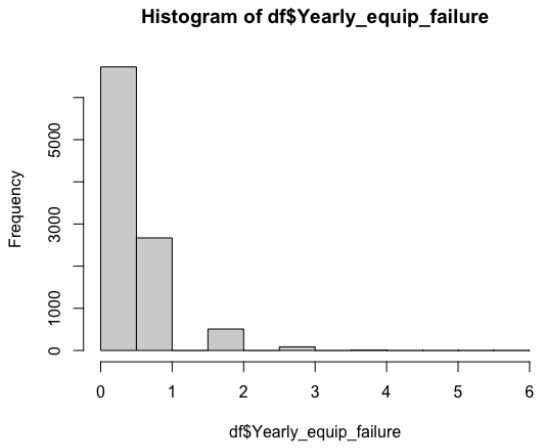
Below are the 12 univariate visualizations histograms for the numeric variables and bar graphs for the categorical generated using R.

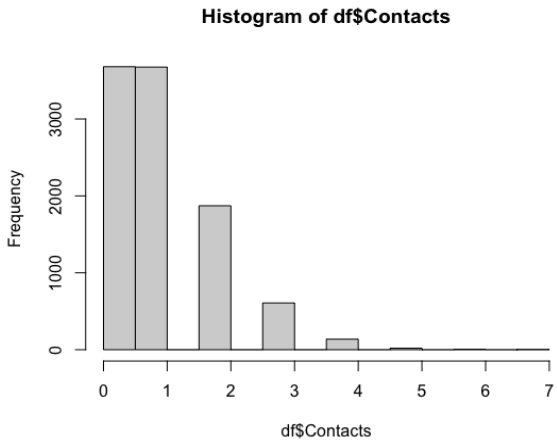


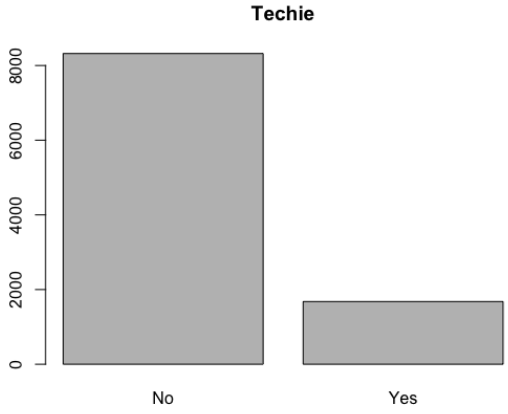


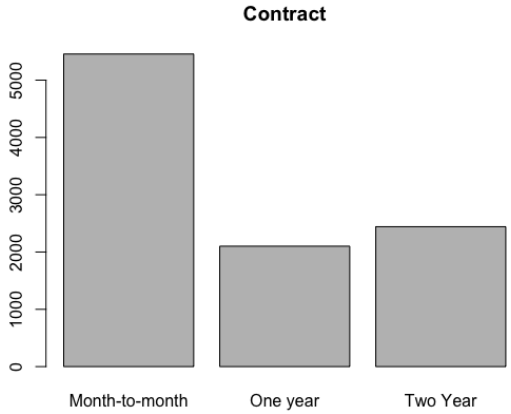


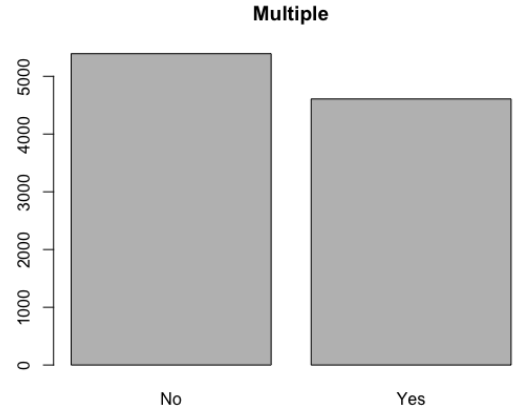


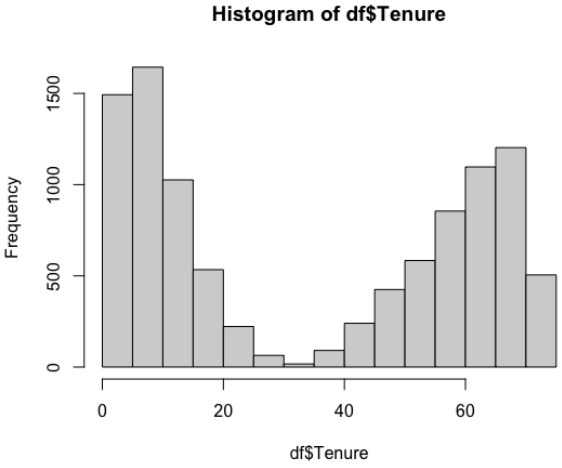

Histogram of df$Yearly_equip_failure
Frequency
0
1000
3000
5000
0
1
2
3
4
5
6
df$Yearly_equip_failure

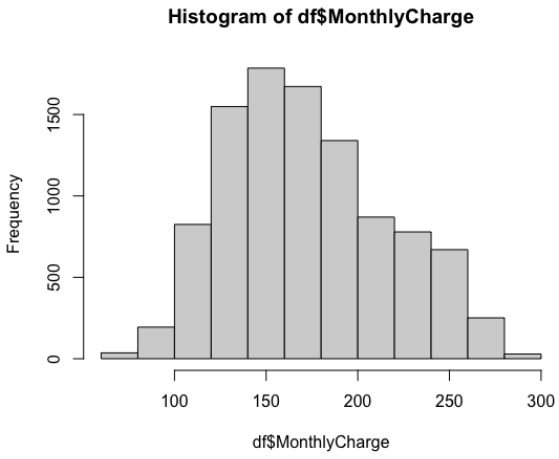

Histogram of df$Contacts
Frequency
0
1000
2000
3000
0
1
2
3
4
5
6
7
df$Contacts

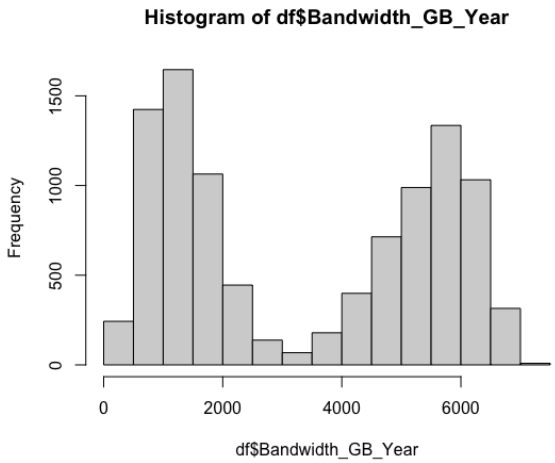

Techie
0
2000
4000
6000
8000
No
Yes


Contract
0
1000
2000
3000
4000
5000
Month-to-month
One year
Two Year

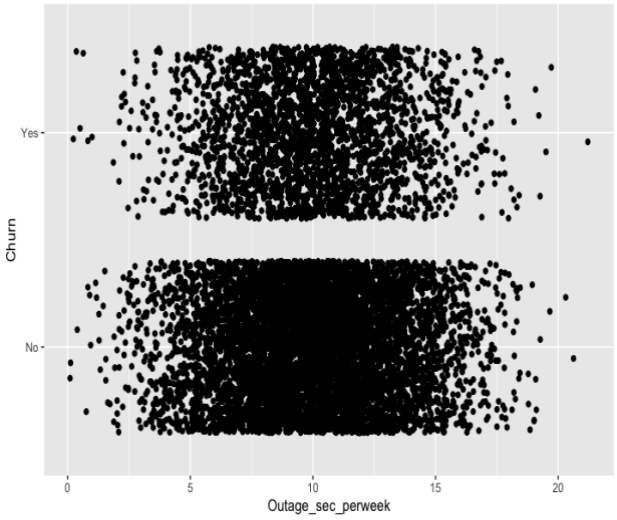

Multiple
0
1000
2000
3000
4000
5000
No
Yes

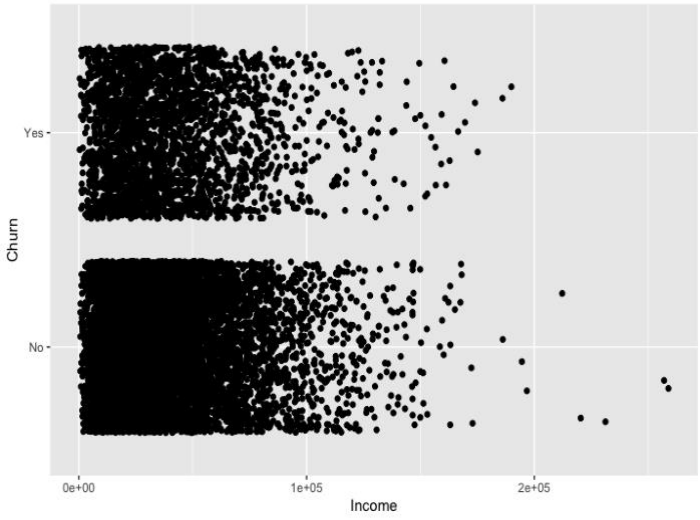

Histogram of df$Tenure
Frequency
0
500
1000
1500
0
20
40
60
df$Tenure

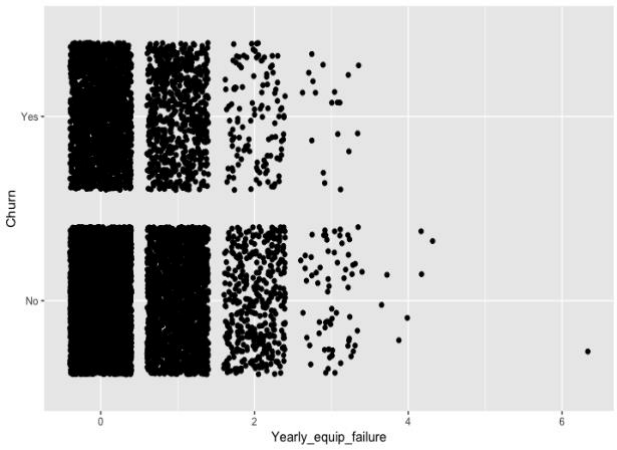


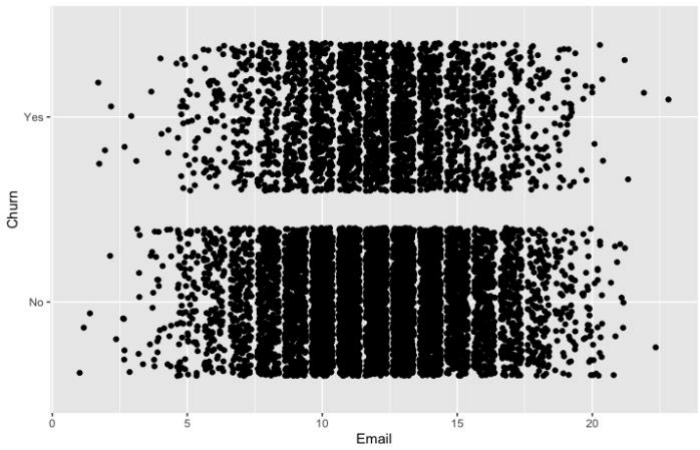


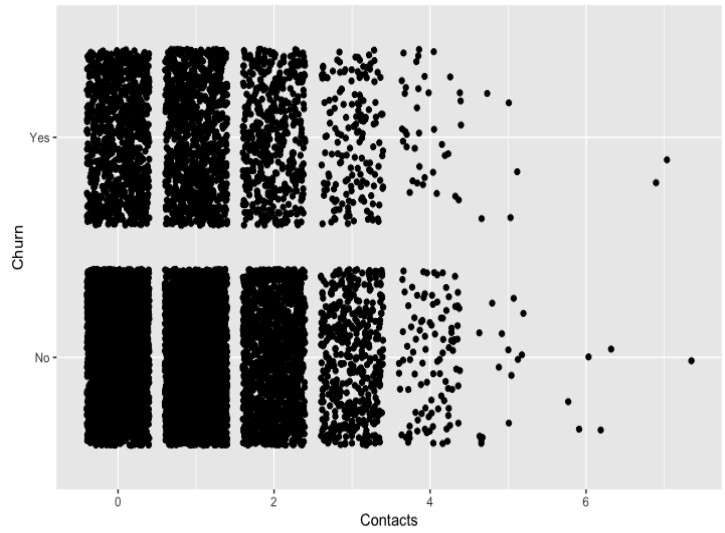
The 11 bivariate visualizations were created using R, the numeric variables use a scatterplot while the categorical use a bar graph. Churn is the dependent variable for all of them.

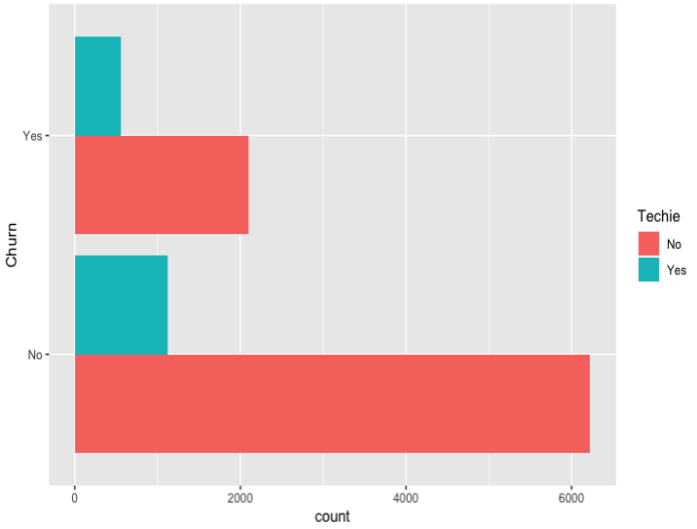


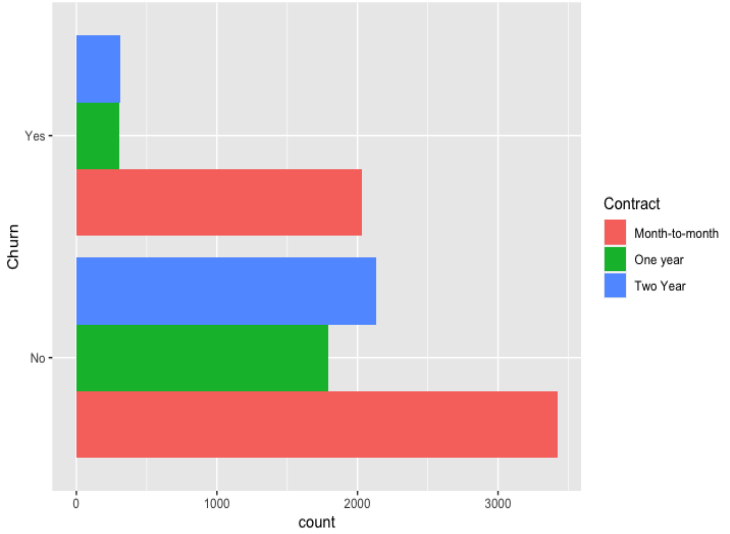


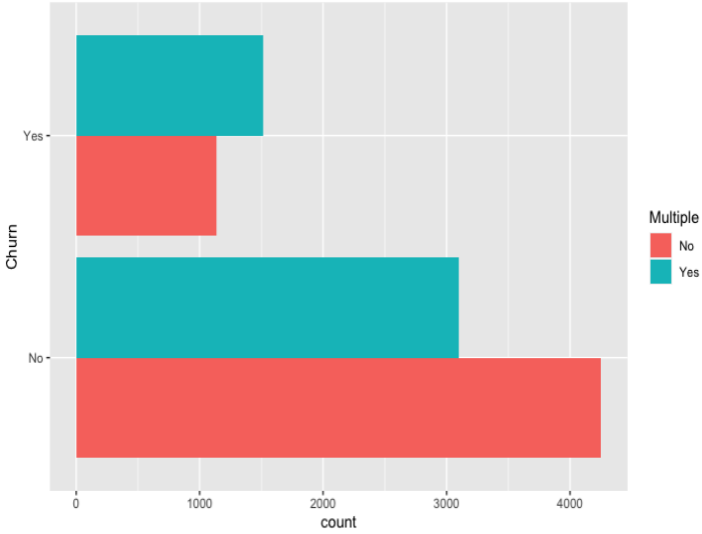


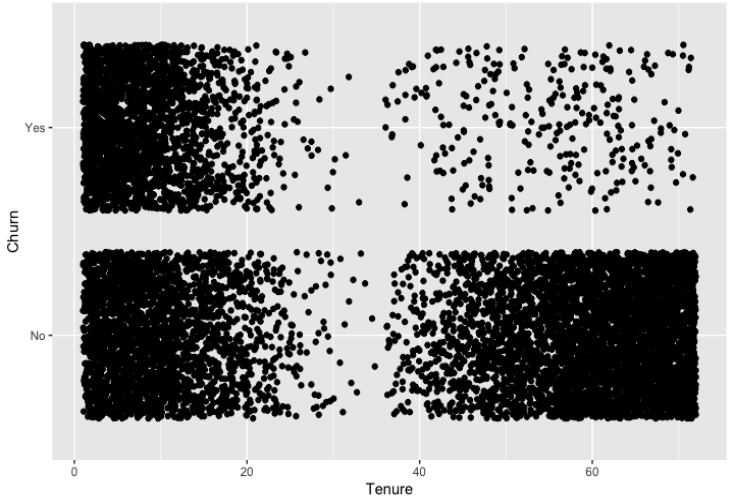


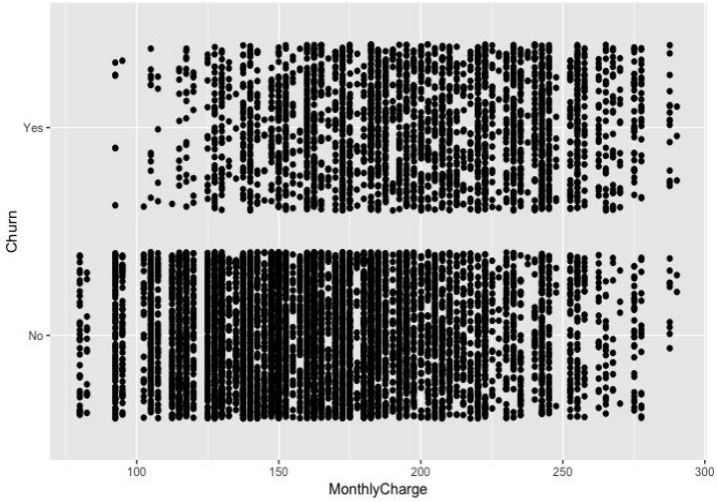

Churn
Yes
No
0
2
4
6
Contacts

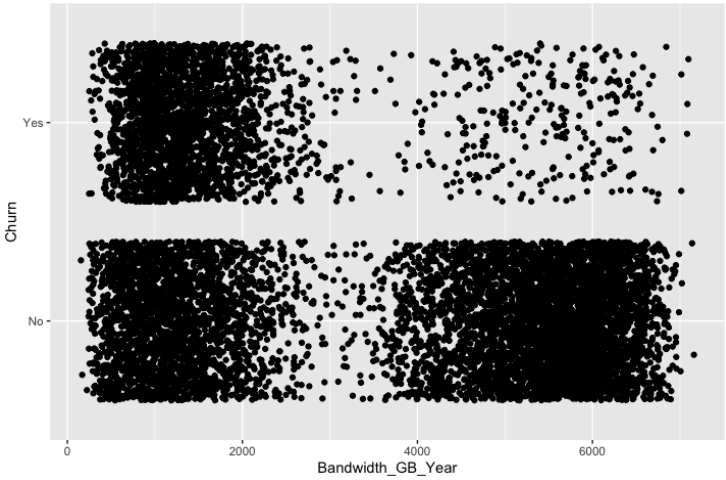

Churn
Yes
No
0
2000
4000
6000
count
Techie
No
Yes


Churn
Yes
No
0
1000
2000
3000
count
Contract
Month-to-month
One year
Two Year


Churn
Yes
No
0
1000
2000
3000
4000
count
Multiple
No
Yes


Churn
Yes
No
0
20
40
60
Tenure


Churn
Yes
No
100
150
200
250
300
MonthlyCharge



C4.

My goals for data transformation are to transform the categorical variables into numeric representations. Churn, Techie, and Multiple are all binomial so they can be transformed into 1 for “Yes” and 0 for “No”. Contract is a nominal value so we can use one hot encoding to transform the three values into columns, Contract_month, Contract_One_year and Contract_Two year respectively. A 1 in the respective columns signals the customer has that particular contract and a 0 means they do not have that contract.

```
#factor Contract
df$Contract <- as.factor(df$Contract)
Contract <- table(df$Contract)
df <-  df %>% mutate(Churn = revalue(Churn, c("Yes" = 1, "No" = 0)))
df$Churn <- as.numeric(df$Churn)
df <- df %>% mutate(Techie = revalue(Techie, c("Yes" = 1, "No" = 0)))
df$Techie <- as.numeric(df$Techie)
df <-   df %>%  mutate(Multiple = revalue(Multiple, c("Yes" = 1, "No" = 0)))
df$Multiple <- as.numeric(df$Multiple)


new_df <- one_hot(as.data.table(df))
```

C5.

The new data file is called new_churn_clean_log.csv.

Part IV: Model Comparison and Analysis

D1.

My initial model is in the form of ln(p/p-1) = Y = $B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6 + B_7X_7 + B_8X_8 + B_9X_9 + B_{10}X_{10} + B_{11}X_{11} + B_{12}X_{12.}$ Y represents churn and B_0 is the Y intercept. $B_1 – B_{12}$ are the coefficients for the following X-values representations: X_1 is Outage_sec_perweek, X_2 is Yearly_equip_failure, X_3 is Income, X_4 is Email, X_5 is Contacts, X_6 is Techie, X_7 is Contract_Month, X_8 is Contract_One_year, X_9 is Multiple, X_{10} is Tenure, X_{11} is MonthlyCharge and X_{12} is Bandwidth_GB_Year.

Contract_Two year is left out due to k-1 to reduce multicollinearity. Then running the following model in R:

model <- glm(Churn ~ Outage_sec_perweek + Yearly_equip_failure + Income + Email + Contacts + Techie + Contract_Month + Contract_One_year Multiple + Tenure + MonthlyCharge + Bandwidth_GB_Year ,data = new_df, family = "binomial")

We get the coefficients thus giving us the initial model of Y = -9.506 - $2.994e\text{-}03X_1$ - $1.690e\text{-}02X_2$ + $1.850e\text{-}07X_3$ - $6.372e\text{-}03X_4$ + $6.253e\text{-}02X_5$ + $9.613e\text{-}01X_6$ + $3.206X_7$ + $8.573e\text{-}02X_8$ + $8.691e\text{-}02X_9$ - $3.963e\text{-}01X_{10}$ + $3.654e\text{-}02X_{11}$ + $3.554e\text{-}03X_{12.}$

D2.

Running a variance inflation factor we see that Tenure has a vif of 66 and bandwidth has a vif of 65 thus giving evidence that these two variables are correlated. Then running backwards stepwise regression on the model using the criteria of P-values; if it is greater than .05 it is removed and less than .05 it is kept.

7 variables are removed Income, Outage_sec_perweek, Yearly_equip_failure, Email, Contract_One_year, Multiple, Contacts.

D3.

Using the above D2 to remove variables we are left with a reduced model in the form of $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4$. With X_1 now Techie, X_2 is Contract_month, X_3 is MonthlyCharge and X_4 is Bandwidth_GB_Year. In R the new model looks like:

new_model <- glm(Churn ~ Techie + Contract_Month + MonthlyCharge + Bandwidth_GB_Year,data = new_df, family = "binomial")

Running this gives the coefficients of $Y = -7.933 + 0.7485X_1 + 2.553X_2 + 4.314e\text{-}02X_3 - 9.907e\text{-}04X_4$. Below is a screenshot first of my initial model then my reduced model:

```
Call:
glm(formula = Churn ~ Outage_sec_perweek + Yearly_equip_failure +
    Income + Email + Contacts + Techie + Contract_Month + Contract_One_year +
    Multiple + Tenure + MonthlyCharge + Bandwidth_GB_Year, family = "binomial",
    data = new_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8217  -0.3031  -0.0717   0.1018   3.3844

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -9.506e+00  3.251e-01 -29.241   <2e-16 ***
Outage_sec_perweek   -2.994e-03  1.235e-02  -0.242   0.8084
Yearly_equip_failure -1.690e-02  5.755e-02  -0.294   0.7690
Income                1.850e-07  1.290e-06   0.143   0.8860
Email                -6.372e-03  1.208e-02  -0.527   0.5979
Contacts              6.253e-02  3.652e-02   1.712   0.0869 .
Techie                9.613e-01  9.646e-02   9.966   <2e-16 ***
Contract_Month        3.206e+00  1.153e-01  27.804   <2e-16 ***
Contract_One_year     8.573e-02  1.282e-01   0.668   0.5038
Multiple              8.691e-02  7.736e-02   1.123   0.2613
Tenure               -3.963e-01  1.466e-02 -27.039   <2e-16 ***
MonthlyCharge         3.654e-02  1.245e-03  29.342   <2e-16 ***
Bandwidth_GB_Year     3.554e-03  1.636e-04  21.727   <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: 11564.4  on 9999  degrees of freedom
Residual deviance:  4806.6  on 9987  degrees of freedom
AIC: 4832.6

Number of Fisher Scoring iterations: 7
```

```
> summary(new_model)

Call:
glm(formula = Churn ~ Techie + Contract_Month + MonthlyCharge +
    Bandwidth_GB_Year, family = "binomial", data = new_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7917  -0.4071  -0.1317   0.1856   3.2353

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.933e+00  2.047e-01  -38.76   <2e-16 ***
Techie             7.486e-01  8.735e-02    8.57   <2e-16 ***
Contract_Month     2.553e+00  8.128e-02   31.41   <2e-16 ***
MonthlyCharge      4.314e-02  1.094e-03   39.42   <2e-16 ***
Bandwidth_GB_Year -9.907e-04  2.392e-05  -41.42   <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: 11564.4  on 9999  degrees of freedom
Residual deviance:  5771.2  on 9995  degrees of freedom
AIC: 5781.2

Number of Fisher Scoring iterations: 6
```

E1.

Calculating the p-values for each model we follow the steps outlined in (Statology. (n.d.).)By taking Null deviance – Residual deviance than using that number and the degrees of freedom 12 for the first model and 4 for the second model. Then we can plug that into a p-values calculator(Statology. (n.d.).) we get .000000 for each model. The first model values for Null – Residual would be 11564.4 - 4806.6 = 6757.8. The new model values are 11564.4 - 5771.2 = 5793.2.

So, both models are statistically significant as they are both below a significance level of .05. However, using the AIC values given from the summaries in D3 of 4832.6 for the first model and 5781.2 for the new model, this implies that the first model is a better fit for predicting Churn.

E2.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6778  572
         1  785 1865

               Accuracy : 0.8643
```

Above is a screenshot of the confusion matrix and the accuracy for my reduced model, using the confusionMatrix function in R. It has 6778 true positives, 572 false positives, 785 false negatives and 1865 true negatives. The model also has an 86.43% accuracy.

E3.

This is the code I used for logistic regression model in R:

```
library(plyr)
library(dplyr)
library(ggplot2)
library(car)
library(mltools)
library(data.table)
library(DescTools)
library(glmtoolbox)
library(caret)
getwd()
setwd('/Users/stevenschindler/Documents/R/D208')
df <- read.csv('clean_churn_file_log.csv', header = TRUE)
head(df)
summary(df$Churn)
summary(df$Outage_sec_perweek)
summary(df$Yearly_equip_failure)
summary(df$Income)
summary(df$Email)
summary(df$Contacts)
summary(df$Techie)
```

```
summary(df$Contract)
summary(df$Multiple)
summary(df$Tenure)
summary(df$MonthlyCharge)
summary(df$Bandwidth_GB_Year)



#factor Contract
df$Contract <- as.factor(df$Contract)
Contract <- table(df$Contract)

#convert Yes,No to 1,0
df <-
  df %>%
  mutate(Churn = revalue(Churn, c("Yes" = 1, "No" = 0)))
df$Churn <- as.numeric(df$Churn)

df <-
  df %>%
  mutate(Techie = revalue(Techie, c("Yes" = 1, "No" = 0)))
df$Techie <- as.numeric(df$Techie)

df <-
  df %>%
  mutate(Multiple = revalue(Multiple, c("Yes" = 1, "No" = 0)))
df$Multiple <- as.numeric(df$Multiple)

#onehot encoding
new_df <- one_hot(as.data.table(df))


new_df <- new_df %>% rename("Contract_Month" = "Contract_Month-to-
month")
new_df <- new_df %>% rename("Contract_One_year" = "Contract_One year")
write.csv(new_df, "new_churn_clean_log.csv", row.names=FALSE)
```

```
model <- glm(Churn ~ Outage_sec_perweek + Yearly_equip_failure + Income +
Email
          + Contacts + Techie + Contract_Month + Contract_One_year +
          Multiple + Tenure + MonthlyCharge + Bandwidth_GB_Year ,data =
new_df, family = "binomial")


summary(model)
vif(model)

stepCriterion(model, direction="backward", criterion="p-value") (
```
R Core Team. (2021))

```
#null deviance - residual deviance for model
 11564.4 - 4806.6

new_model <- glm(Churn ~ Techie + Contract_Month + MonthlyCharge +
Bandwidth_GB_Year,data = new_df,
              family = "binomial")

summary(new_model)

#null deviance - residual deviance for new_model
 11564.4 - 5771.2

 #confusion matrix
pred <- predict(new_model,type="response" )
pred1 <-ifelse(pred > 0.5,1,0)

Churn <- new_df$Churn

confusionMatrix(as.factor(Churn),as.factor(pred1))
```
(Statology. (n.d.).)

```
#factor Churn,Techie,Multiple
df$Churn <- as.factor(df$Churn)
df$Techie <- as.factor(df$Techie)
df$Multiple <- as.factor(df$Multiple)
```

```
Churn <- table(df$Churn)
Churn
Techie <- table(df$Techie)
Multiple <- table(df$Multiple)



#histograms
hist(df$Outage_sec_perweek)
hist(df$Yearly_equip_failure)
hist(df$Income)
hist(df$Email)
hist(df$Bandwidth_GB_Year)
hist(df$Tenure)
hist(df$Contacts)
hist(df$MonthlyCharge)
#barplot
barplot(Churn,main ="Churn")
barplot(Techie,main="Techie")
barplot(Contract,main="Contract")
barplot(Multiple,main="Multiple")


ggplot(df, aes(x = Outage_sec_perweek, y = Churn)) + geom_jitter()
ggplot(df, aes(x = Yearly_equip_failure, y = Churn)) + geom_jitter()
ggplot(df, aes(x = Income , y = Churn)) + geom_jitter()
ggplot(df, aes(x = Email, y = Churn)) + geom_jitter()
ggplot(df, aes(x = Bandwidth_GB_Year, y = Churn)) + geom_jitter()
ggplot(df, aes(x = Tenure, y = Churn)) + geom_jitter()
ggplot(df, aes(x = Contacts, y = Churn)) + geom_jitter()
ggplot(df, aes(x = MonthlyCharge, y = Churn)) + geom_jitter()

ggplot(df, aes(y = Churn, fill = Techie)) + geom_bar(position = "dodge")
ggplot(df, aes(y = Churn, fill = Contract)) + geom_bar(position = "dodge")
ggplot(df, aes(y = Churn, fill = Multiple)) + geom_bar(position = "dodge")
```

Part V: Data Summary and Implications

F1.

From D3 we know that a regression equation for a reduced model is ln(p/1-p) = -7.933 + 0.7485X_1 + 2.553X_2 + 4.314e-02X_3 - 9.907e-04X_4. The Y intercept is -7.933 meaning when all other variables are zero the log odds for churn are -7.933, using this calculator(Lowry, R. (2021).) gives a probability of 0.0004 for a customer to churn. The coefficient for Techie(X_1) is 0.7485 and using the same calculator it gives a probability of 0.67 of a customer churn if they consider themselves a techie, all other things being equal.

The coefficient for Contract_month(X_2) is 2.553 which gives a probability of 0.9278 that a customer will churn if they have a month-to-month contract. MonthlyCharge(X_3) has a coefficient of 4.314 *10^-02 which gives a probability of 0.5108 that a customer will churn. Probability of churning increases as MonthlyCharge increases. Finally, Bandwidth_GB_Year(X_4) has a negative coefficient of -9.907*10^-04 which gives a probability of 0.4997 of the customer churning, the probability of churning decreases as bandwidth increases.

From the summary of new_model we see that the p-value is 2 *10^-16 for each variable. The assumed significance level is .05 so we can conclude that my reduced model is statistically significant. The model also has 86% accuracy so could be of practical use in predicting if a customer will churn or not. So, the model is practically significant however, there is still a 14% chance of error to be cautioned of.

A limitation of my analysis is that regression is not resistant to outliers, and I did not remove any outliers on my dataset. A second limitation is correlation does not equal causation, so even though Contract_month has a high probability of .92 the type of contract is not necessary the reason for the churn. Along the same lines

being an individual techie does not automatically mean they have a higher chance of churning.

F2.

The four variables that will contribute to a customer that will churn are Techie, Contract_month, MonthlyCharge and Bandwidth_GB_year. The longer a customer stays with the company the more GB they use per year the less likely it is that they will churn. Along the same lines a contract of only a month seems to indicate a high churn. A recommended course of action would be to remove the monthly contract and offer only one-to-two-year contracts. Churn also increases as MonthlyCharge increases so the contracts need to be of a reasonable price which can be calculated in another analysis.

References

H.

R Core Team. (2021). stepCriterion.glm function. GLM Toolbox Package Documentation. Retrieved April 12, 2023, from https://search.r-project.org/CRAN/refmans/glmtoolbox/html/stepCriterion.glm.html

Statology. (n.d.). How to create a confusion matrix in R. Retrieved April 12, 2023, from https://www.statology.org/confusion-matrix-in-r/

I.

Statistics Solutions. (n.d.). Assumptions of Logistic Regression. Retrieved April 14, 2023, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

Statology. (n.d.). How to calculate the null residual deviance in R. Retrieved April 12, 2023, from https://www.statology.org/null-residual-deviance/

Statology. (n.d.). Chi-square p-value calculator. Retrieved April 12, 2023, from https://www.statology.org/chi-square-p-value-calculator/

Lowry, R. (2021). Odds ratio calculator. VassarStats Website. Retrieved April 12, 2023, from http://vassarstats.net/tabs_odds.html