# Odds of Churn

Victoria Espinola

2022-09-28

**Part 1: Research Question** A1. This project will focus on understanding the odds that a customer will churn based on customer profile comprised of tenure, usage, and survey responses.

A2. Since new customer aquisition is 10 times that of keeping an existing customer, this question is fitting for stakeholders to understand customer life cycles and key metrics for improving customer experience by including customer survey responses and bandwidth consumption.

B1. The assumptions of the logistic regression model are: (1) the response variable is binary, (2) the observations are independent of each other, (3) There is no multicollinearity between explanatory variables, (4) there are no extreme outliers in the data set, (5) there is a linear relationship between the response and explanatory variables, and (6) the sample size is suffiently large (statology).

B2. The benefits for using R is that R is a statistical programming language and inherently has many statistical applications build into it. According to Srinivasan from the Pluralsight video, Understanding and Applying Logistic Regression, R is the best tool to use for regression. (Logistic Regression).

B3. Logistic regression is an appropriate technique for this anaylsis because the response variable is binary, i.e., a customer only has two outcomes in behavior, they either churn, discontinue services or do not churn and retain services.

**Part III. Data Preparation** C1.To prepare the data, the cleaning function was used, this eliminates outlier variables and removes all missing values by imputing them with the mean and removing outliers from the data set. The preparation goals are to scale the observations for each variable by using log10 scale. This will help to ensure the outcomes are not being heavily influenced by large quantities. The goal of this manipulation is to improve the understanding of the logistic model. The churn variable was also converted to a numeric variable with outcomes of 0 and 1.

C2. The initial variable selection was based on customer profile attributes that could impact their likelihood of churning and further analyzed to include only variables that could be influenced by stakeholder actions. After the intial model (logit_1) was created, variable importance function in R was used. This function ranks the variables importance to churn and allows for the best variable selection that will help to not over fit the model. The final model will include churn, bandwidth_gb_year, tenure, monthly_charge, techie, item3, item 4, and item5. These variable were ranked highest in the variable importance and they are all statisticall signficant based on the p-value being less than 0.05.

C3. Below is the code with annotations showing the data preparation explained in C1 and summary statistic for each variable in the final data set as required for C2. To prepare the data, the names were standardized and have all capital letter changed to lower case and spaces replaced with underscores. Next, the variables have been selected and have had a log base ten plus one applied to adjusted all the numeric values. Finally, the categorical variable churn was encoded with zeros and ones for no, yes outcomes.

```
#Bringing in the data, initial cleaning with cleaning source function
data <- read.csv("churn_clean.csv", header = TRUE) %>%
  janitor::clean_names() %>%
  churn_cleaning()
```

```
#scaling the data with log 10 + 1, converting churn to numeric binary outcomes of 0's and 1's.
reduced_scaled <- data %>%
                select(churn, marital, outage_sec_perweek,
                       yearly_equip_failure, bandwidth_gb_year,
                       tenure, monthly_charge, techie, item1:item8) %>% #Select variables for model
                mutate(across(.col = c(-churn, -marital, -techie), log1p))%>%  #Change to log10+1 scale
                mutate(churn = as.numeric(ifelse(data$churn == 'Yes',1,0))) #Encode churn as 0's and 1'
```

```
#Summary Statistics
summary(reduced_scaled)
```

```
##      churn                marital     outage_sec_perweek yearly_equip_failure
##  Min.   :0.0000   Divorced     :1866   Min.   :0.763      Min.   :0.0000
##  1st Qu.:0.0000   Married      :1699   1st Qu.:2.201      1st Qu.:0.0000
##  Median :0.0000   Never Married:1753   Median :2.399      Median :0.0000
##  Mean   :0.2666   Separated    :1804   Mean   :2.359      Mean   :0.2446
##  3rd Qu.:1.0000   Widowed      :1828   3rd Qu.:2.562      3rd Qu.:0.6931
##  Max.   :1.0000                       Max.   :2.988      Max.   :1.0986
##  bandwidth_gb_year     tenure       monthly_charge  techie         item1
##  Min.   :5.053     Min.   :0.6957   Min.   :4.394   No :7452   Min.   :0.6931
##  1st Qu.:7.114     1st Qu.:2.1852   1st Qu.:4.949   Yes:1498   1st Qu.:1.3863
##  Median :8.046     Median :3.4266   Median :5.127              Median :1.3863
##  Mean   :7.832     Mean   :3.1107   Mean   :5.127              Mean   :1.4702
##  3rd Qu.:8.627     3rd Qu.:4.1334   3rd Qu.:5.315              3rd Qu.:1.6094
##  Max.   :8.876     Max.   :4.2904   Max.   :5.674              Max.   :1.9459
##      item2            item3            item4            item5
##  Min.   :0.6931   Min.   :0.6931   Min.   :0.6931   Min.   :0.6931
##  1st Qu.:1.3863   1st Qu.:1.3863   1st Qu.:1.3863   1st Qu.:1.3863
##  Median :1.3863   Median :1.3863   Median :1.3863   Median :1.3863
##  Mean   :1.4741   Mean   :1.4700   Mean   :1.4744   Mean   :1.4743
##  3rd Qu.:1.6094   3rd Qu.:1.6094   3rd Qu.:1.6094   3rd Qu.:1.6094
##  Max.   :1.9459   Max.   :1.9459   Max.   :1.9459   Max.   :1.9459
##      item6            item7            item8
##  Min.   :0.6931   Min.   :0.6931   Min.   :0.6931
##  1st Qu.:1.3863   1st Qu.:1.3863   1st Qu.:1.3863
##  Median :1.3863   Median :1.3863   Median :1.3863
##  Mean   :1.4726   Mean   :1.4758   Mean   :1.4724
##  3rd Qu.:1.6094   3rd Qu.:1.6094   3rd Qu.:1.6094
##  Max.   :1.9459   Max.   :1.9459   Max.   :1.9459
```

C3. This code shows the first model (logit_1), its performance and variable importance table results.

```
#Initial model with variable importance to determine which variables need to be included in final data
logit_1 <- glm(churn ~ ., data = reduced_scaled, family = "binomial")
summary(logit_1)
```

```
##
## Call:
## glm(formula = churn ~ ., family = "binomial", data = reduced_scaled)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.9787  -0.5408  -0.2582   0.2614   3.2521
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -31.578411   1.193241 -26.464  < 2e-16 ***
## maritalMarried         -0.020620   0.102408  -0.201    0.840
## maritalNever Married   -0.083178   0.100157  -0.830    0.406
## maritalSeparated        0.074676   0.099811   0.748    0.454
## maritalWidowed          0.098916   0.099112   0.998    0.318
## outage_sec_perweek     -0.030407   0.108316  -0.281    0.779
## yearly_equip_failure    0.001226   0.087455   0.014    0.989
## bandwidth_gb_year       1.498774   0.158639   9.448  < 2e-16 ***
## tenure                 -2.576468   0.123487 -20.864  < 2e-16 ***
## monthly_charge          5.262699   0.171892  30.616  < 2e-16 ***
## techieYes               0.529101   0.081794   6.469 9.89e-11 ***
## item1                  -0.011187   0.189249  -0.059    0.953
## item2                  -0.055177   0.179690  -0.307    0.759
## item3                  -0.205587   0.161633  -1.272    0.203
## item4                  -0.155414   0.145666  -1.067    0.286
## item5                  -0.169259   0.151517  -1.117    0.264
## item6                   0.038159   0.156186   0.244    0.807
## item7                   0.073926   0.148517   0.498    0.619
## item8                  -0.091078   0.140988  -0.646    0.518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10379.1  on 8949  degrees of freedom
## Residual deviance:  6296.7  on 8931  degrees of freedom
## AIC: 6334.7
##
## Number of Fisher Scoring iterations: 6
```

```
var_imp <- caret::varImp(logit_1, scale = TRUE)
var_imp #scale = TRUE complete the normalization step.
```

```
##                          Overall
## maritalMarried        0.20134898
## maritalNever Married  0.83047062
## maritalSeparated      0.74816940
## maritalWidowed        0.99802243
## outage_sec_perweek    0.28072784
## yearly_equip_failure  0.01401720
## bandwidth_gb_year     9.44771822
## tenure               20.86420638
## monthly_charge       30.61631942
## techieYes             6.46867172
## item1                 0.05911262
## item2                 0.30706840
## item3                 1.27194040
## item4                 1.06692078
## item5                 1.11709314
## item6                 0.24431968
```

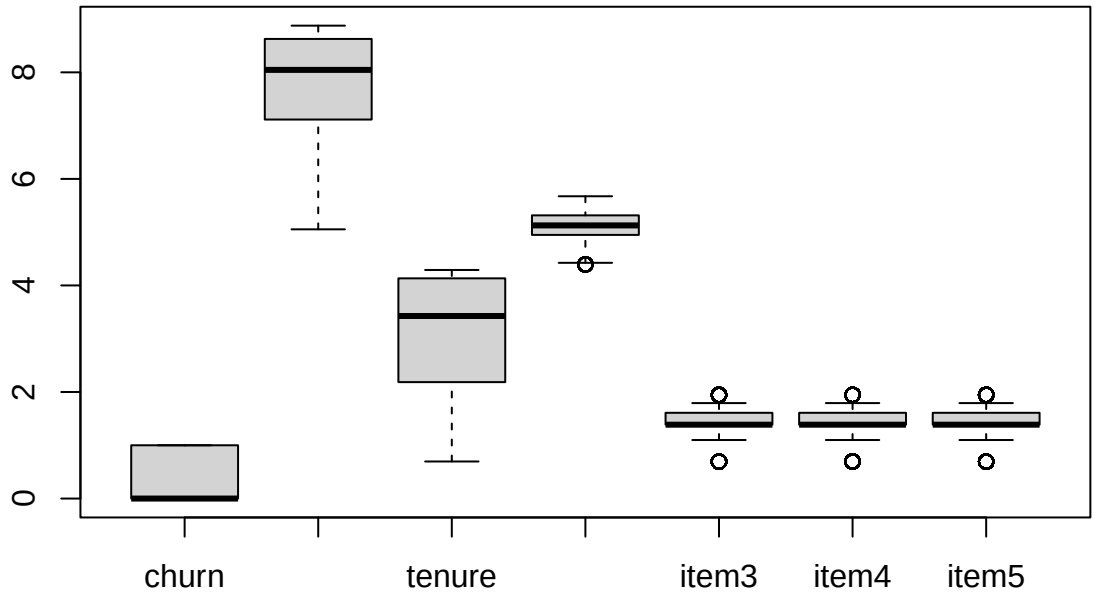
```
## item7                0.49776262
## item8                0.64599671
```

C3. This is the final variable selection after reviewing the second model.

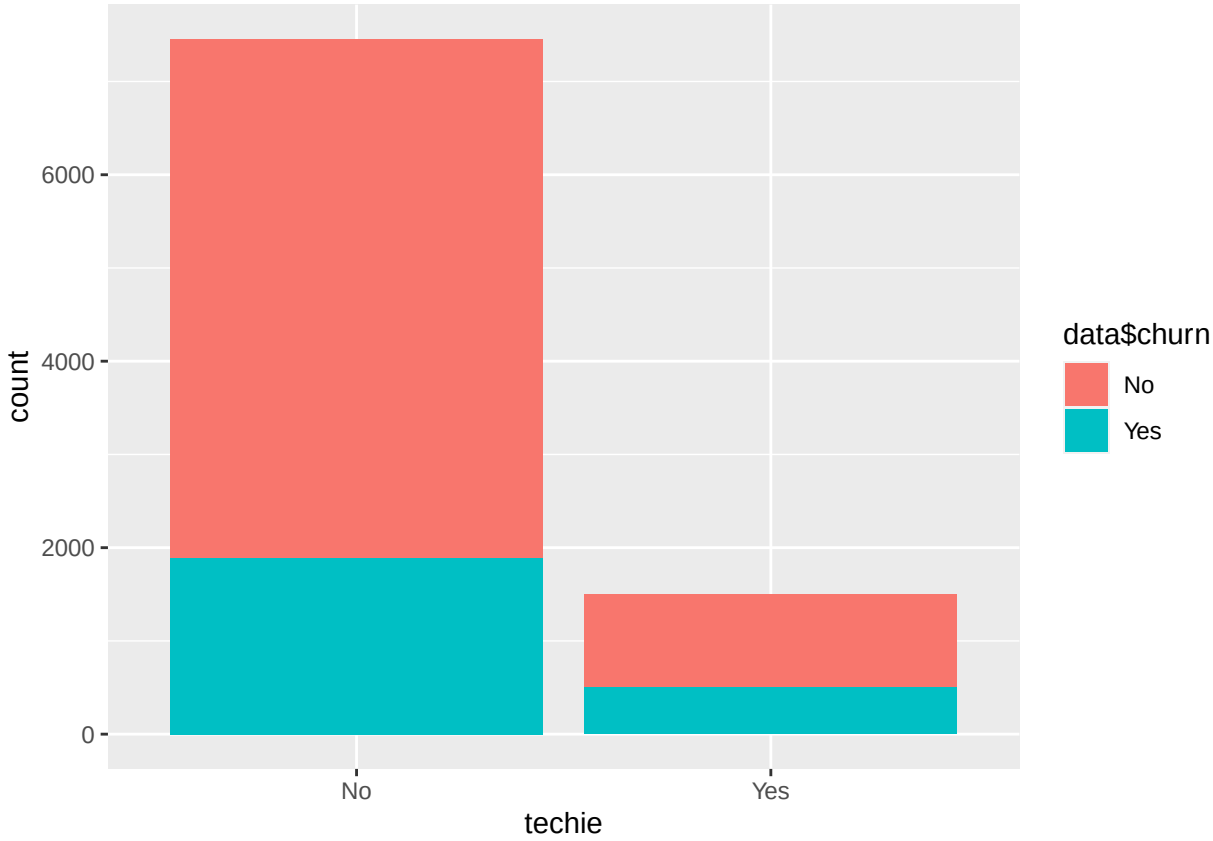
```
#Final data set, cleaned and ready for logistic model. The variable were chosen based on their importan
reduction <- reduced_scaled %>%
                select(churn, bandwidth_gb_year, tenure, monthly_charge, techie, item3:item5)
```

C4. Univariate visualizations with distrubution for all variables in the final data set.

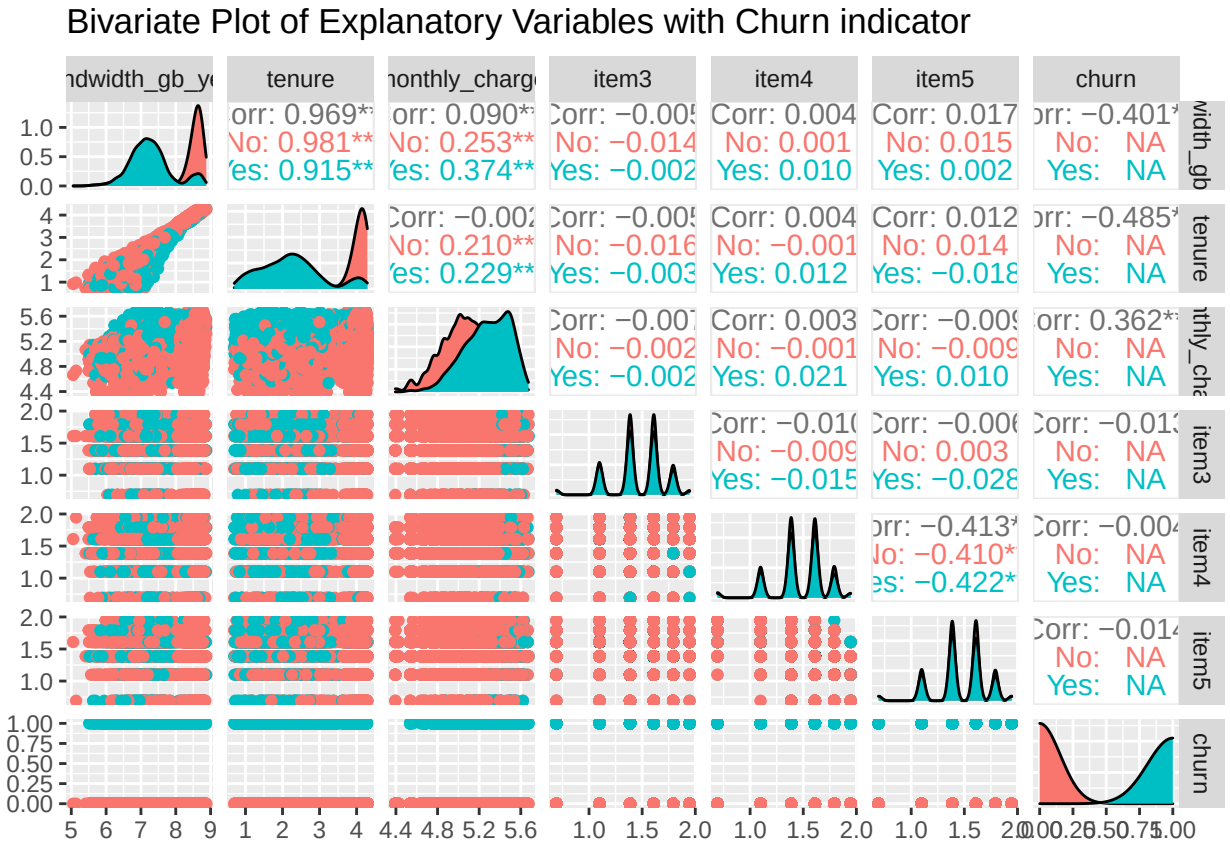
```
#C4- Univariate Visualization, numeric variables
boxplot(reduction %>% select_if(is.numeric))
```




```
#C4- Univariate Visualization, categorical variables
ggplot(data = reduction, aes(x = techie, fill = data$churn)) +
  geom_bar()
```




```
#C4- Bivariate plots using ggpairs (R-Bloggers)
ggpairs(reduction, columns = c("bandwidth_gb_year", "tenure", "monthly_charge", "item3", "item4", "item
```




**Part IV: Model Comparison and Anaylysis** D1. The following shows the next model, which includes the response variable churn, with all explanatory variables kept based on the table of variable importance: bandwidth_gb_year, tenure, monthly_charge, techie, item3, item4, and item5.

```
#D1- initial model with all predictors.
logit_2 <- glm(churn ~ ., data = reduction, family = "binomial")
summary(logit_2)
```

```
##
## Call:
## glm(formula = churn ~ ., family = "binomial", data = reduction)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9837  -0.5433  -0.2582   0.2594   3.2247
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -31.63933    1.13136 -27.966  < 2e-16 ***
## bandwidth_gb_year   1.49827    0.15828   9.466  < 2e-16 ***
## tenure             -2.57483    0.12322 -20.896  < 2e-16 ***
## monthly_charge      5.26112    0.17173  30.637  < 2e-16 ***
## techieYes           0.53393    0.08165   6.539 6.17e-11 ***
## item3              -0.22445    0.13097  -1.714   0.0866 .
## item4              -0.16119    0.14313  -1.126   0.2601
```

```
## item5              -0.18249    0.14397  -1.268   0.2050
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10379.1  on 8949  degrees of freedom
## Residual deviance:  6301.9  on 8942  degrees of freedom
## AIC: 6317.9
##
## Number of Fisher Scoring iterations: 6
```

D2. Based on this initial model, item3, item4, and item5 are not statistically significant and will therefore be dropped from the next model.

D3. The final model was produced using the follow code.

```
#Reducing the model to include only variables with significant p-values
logit_final <- glm(churn ~ bandwidth_gb_year + tenure + monthly_charge + techie, data = reduction, famil
summary(logit_final)
```

```
##
## Call:
## glm(formula = churn ~ bandwidth_gb_year + tenure + monthly_charge +
##     techie, family = "binomial", data = reduction)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9862  -0.5425  -0.2581   0.2568   3.2037
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -32.45029    1.06290 -30.530  < 2e-16 ***
## bandwidth_gb_year   1.49356    0.15817   9.443  < 2e-16 ***
## tenure             -2.57036    0.12315 -20.872  < 2e-16 ***
## monthly_charge      5.26078    0.17158  30.660  < 2e-16 ***
## techieYes           0.53802    0.08157   6.596 4.24e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10379.1  on 8949  degrees of freedom
## Residual deviance:  6306.8  on 8945  degrees of freedom
## AIC: 6316.8
##
## Number of Fisher Scoring iterations: 6
```

E. The analysis began with an inquire on customer churn odds based largely on stakeholders ability to influence customer outcomes; these included tenure, usage, and survey responses. Upon first glance analysis, using the variable importance table, all survey response items except item 3, 4, and 5 were dropped. The first reduced model showed that item 3, 4, and 5 from survey responses were statistically insignificant, having values greater than 0.05, and were thus dropped from the next and final model. The final model included bandwidth_gb_year, tenure, montly_charge and techie. All are statistically significant according to the

inital model. To determine the performance of each model the ROC curve and AUC values will used in the performance analysis as seen in the code below.

```
#ROC curves and AUC for comparison
glm.fit = glm(reduction$churn ~ reduction$monthly_charge, family = binomial)
par(pty = "s")
roc(reduction$churn, glm.fit$fitted.values, plot = TRUE, legacy.axes = TRUE, percent = TRUE,
    xlab = "False Positive Percent", ylab = "True Positive Percent", print.auc = TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

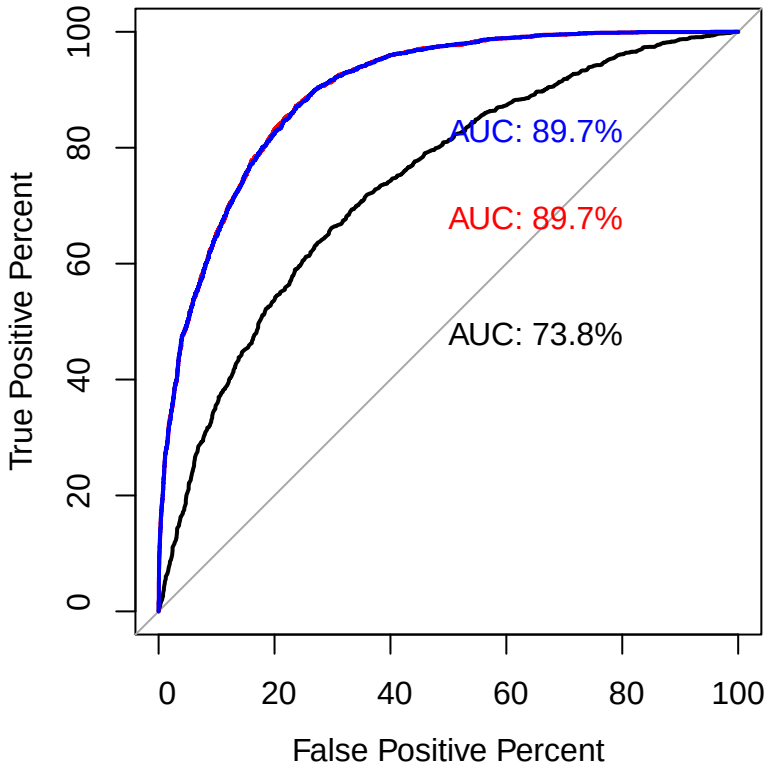
```
##
## Call:
## roc.default(response = reduction$churn, predictor = glm.fit$fitted.values,     percent = TRUE, plot =
##
## Data: glm.fit$fitted.values in 6564 controls (reduction$churn 0) < 2386 cases (reduction$churn 1).
## Area under the curve: 73.84%
```

```
plot.roc(reduction$churn, logit_1$fitted.values, percent = TRUE, col = "red", print.auc = TRUE, add = TR
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot.roc(reduction$churn, logit_2$fitted.values, percent = TRUE, col = "blue", print.auc = TRUE, add = T
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




###### Using the ROC curve and AUC values shown in the graph, the reduced model performed better than the first initial models since its curve is closer to the line y = x and has a lower AUC values of 73.8%. It was noteworthy that the first two inital models had not change even though they had different explanatory variables.

E2. The follow code was used to produce the confusion matrix and its visualization will follow.

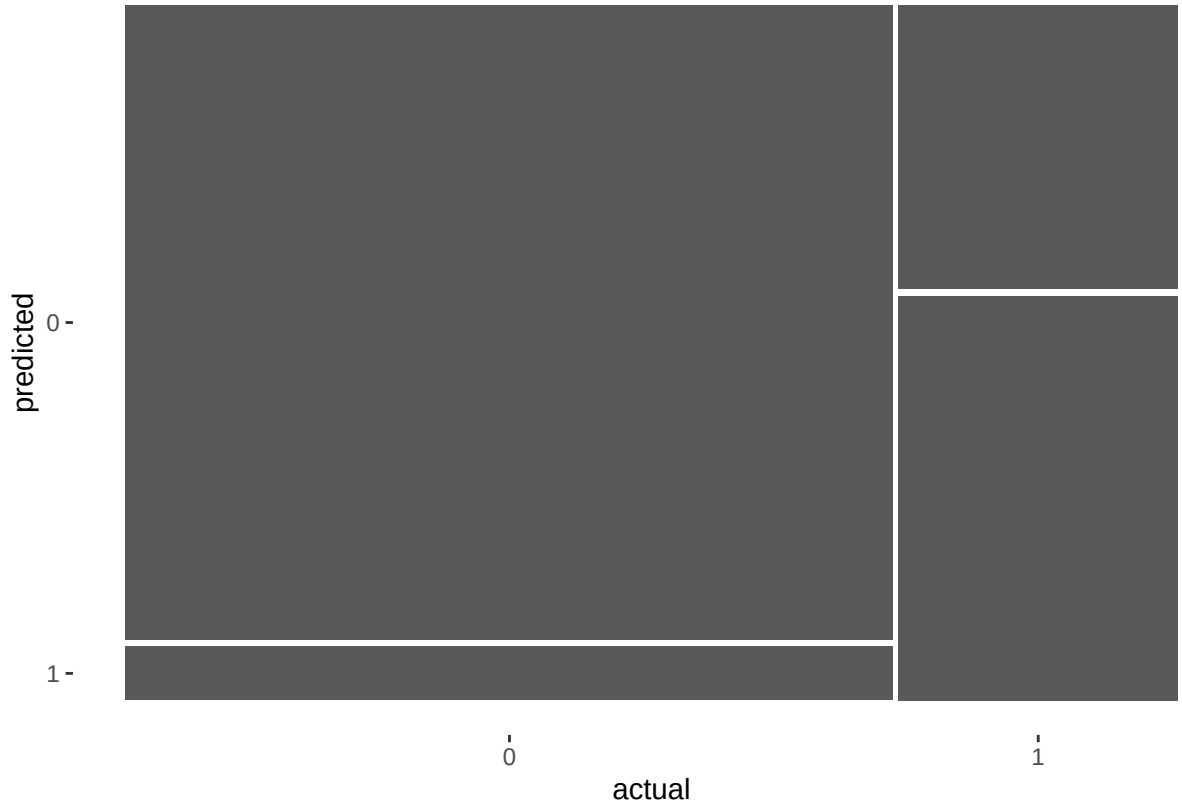
```
#confusion matrix
pred_data <-  reduction %>%
  mutate(
    has_churned = predict(logit_final, reduced_scaled, type = "response")
  )

actual <- reduced_scaled$churn
predicted <- round(fitted(logit_final))

outcomes <- table(predicted, actual)
outcomes
```

```
##          actual
## predicted    0    1
##         0 6053  984
##         1  511 1402
```

```
#Visualizing Confusion Matrix
confusion <- yardstick::conf_mat(outcomes)
autoplot(confusion, )
```




**Part V. Data Summary and Implications** F1. The regression equation for the final model that includes the response variable, churn, and explanatory variables, bandwidth_gb_year, tenure, monthly_charge, and techie is y =

```
#F1
log_coeffs <- coef(logit_final)
coeffs <- exp(coef(logit_final)) +1
log_coeffs
```

```
##       (Intercept) bandwidth_gb_year            tenure    monthly_charge
##        -32.450294          1.493564         -2.570360          5.260783
##         techieYes
##          0.538019
```

```
coeffs
```

```
##       (Intercept) bandwidth_gb_year            tenure    monthly_charge
##          1.000000          5.452937          1.076508        193.632304
##         techieYes
##          2.712611
```

```
reg_eq = 1 + 5.45*data$bandwidth_gb_year + 1.07*data$tenure +193.63*data$monthly_charge +2.71*data$techi
```

F1. According to the final model, for every 5.45gb of bandwidth increase the odds of a customer churning increase by a factor of 1.49, for every 1.07 months of tenure increase, the odds of a customer churning decrease by a factor of 2.57, for every monthly charge greater than 193.63, the odds of a customer churning increase by a factor of 5.26, for every 3 techies the odds of a customer churning increase by a factor of 0.54. The analysis was limited in several areas, one is that some of the variables, namely tenure with both bandwidth_gb_year and monthly_charge were correlated with each other. However, using a combination of variable elimination from the final model did not provide a better ROC curve or AUC value, so all variables were kept in the final model due to their importance value and p-value. Most importantly, the model was limited in the scope of the question, as the variables chosen were included or excluded based on their potenial for stakeholder influence.

F2. From a statistical sense, the model did not provide much insight into a customer's likelihood of churning since the final model only slightly improved when compared to the initial model. However, from a practical sense, it can be determined that a longer term customer is less likely to discontinue services with their current provider holding all other things constant. Furthermore, prices are the biggest influence on a customers decision to churn. Therefore, prices should be keep below $193.63 for customers and those who exceed this amount should be monitored more closely to educate them on how to best use their additional services.

**H. Sources**

1. The 6 Assumptions of Logistic Regression (With Examples). Zach. https://www.statology.org/assumptions-of-logistic-regression/

2. Understanding and Applying Linear Regression. https://app.pluralsight.com/player?course=understanding-applying-linear-regression&author=vitthal-srinivasan&name=understanding-applying-linear-regression-m0&clip=0&mode=live

3. ggpairs in R- A Brief Introduction to ggpairs: R-bloggers. Finnstats. https://www.r-bloggers.com/2021/06/ggpairs-in-r-a-brief-introduction-to-ggpairs/

4. DataCamp. D208. https://app.datacamp.com/learn