D208 Task 2 v2

# A. Business Question

### What is the relationship between churn and other customer demographic variables?

As market competition increases in the telecommunications industry from technological advances spurring the arrival of new players in streaming and information sharing, organizations must find ways to operate at finer margins and cut costs wherever possible.

As the cost to on-board a new customer is already known to be 10 times that of retaining an existing customer, one way to cut costs is by raising retention rates among the existing customer base. Therefore, it would be a valuable undertaking for an organization to explore causal relationships between churn and various customer demographics. Armed with this knowledge, new strategic initiatives could be developed as needed to target the retention of highly valued existing customers.

The data set used for this analysis will be the churn\_clean.csv file.

# B. Logistic Regression Model

Using a logistic regression model we can attempt to predict churn based on customer variables. The logistic regression model is chosen as we have a binary response variable (Churn: Yes / No) for which to explore the explanatory relationships.

With a logistic regression model, we have an assumption that the response variable is binary. Additionally, we have an assumption that the independent variables are not highly correlated individually but may be correlated to the dependent variable of churn. Whereas with a linear regression model residuals follow a normal distribution, in a logistic regression model we have an assumption that the residuals will follow a binomial distribution. This is because the logistic model attempts to predict the probability of a binary outcome.

R will be used for this analysis. R is open source software that was specifically made for statistical analysis. Using R, we can ingest the raw data set, and leveraging an extensive library of data manipulation and visualization packages, clean and investigate the data. More information can be found on the R project website (<https://www.r-project.org/>).

# C. Data Preparation

To prepare the data, first a check is run for missing values in the data set. No action is needed as no missing values were detected.

df<-read.csv("c:/users/shua/documents/Predictive Modeling\_D208/churn\_clean.csv")  
sapply(df, function(x) sum(is.na(x)))

## CaseOrder Customer\_id Interaction   
## 0 0 0   
## UID City State   
## 0 0 0   
## County Zip Lat   
## 0 0 0   
## Lng Population Area   
## 0 0 0   
## TimeZone Job Children   
## 0 0 0   
## Age Income Marital   
## 0 0 0   
## Gender Churn Outage\_sec\_perweek   
## 0 0 0   
## Email Contacts Yearly\_equip\_failure   
## 0 0 0   
## Techie Contract Port\_modem   
## 0 0 0   
## Tablet InternetService Phone   
## 0 0 0   
## Multiple OnlineSecurity OnlineBackup   
## 0 0 0   
## DeviceProtection TechSupport StreamingTV   
## 0 0 0   
## StreamingMovies PaperlessBilling PaymentMethod   
## 0 0 0   
## Tenure MonthlyCharge Bandwidth\_GB\_Year   
## 0 0 0   
## Item1 Item2 Item3   
## 0 0 0   
## Item4 Item5 Item6   
## 0 0 0   
## Item7 Item8   
## 0 0

Only select variables will be used to streamline the data set. To begin, the data set will be filtered to only include the following variables: Population, Children, Age, Churn, Outage seconds per week, Contacts, Tenure, Monthly Charge, and Bandwidth GB per Year.

The churn variable will be converted to a dummy variable by transforming “Yes”/“No” values to 1/0 respectively. To do this we will use the case\_when statement from the “dplyr” library. Dplyr is a data manipulation package within R. More information can be found on the tidyverse website (<https://dplyr.tidyverse.org/>).

colnames(df)

## [1] "CaseOrder" "Customer\_id" "Interaction"   
## [4] "UID" "City" "State"   
## [7] "County" "Zip" "Lat"   
## [10] "Lng" "Population" "Area"   
## [13] "TimeZone" "Job" "Children"   
## [16] "Age" "Income" "Marital"   
## [19] "Gender" "Churn" "Outage\_sec\_perweek"   
## [22] "Email" "Contacts" "Yearly\_equip\_failure"  
## [25] "Techie" "Contract" "Port\_modem"   
## [28] "Tablet" "InternetService" "Phone"   
## [31] "Multiple" "OnlineSecurity" "OnlineBackup"   
## [34] "DeviceProtection" "TechSupport" "StreamingTV"   
## [37] "StreamingMovies" "PaperlessBilling" "PaymentMethod"   
## [40] "Tenure" "MonthlyCharge" "Bandwidth\_GB\_Year"   
## [43] "Item1" "Item2" "Item3"   
## [46] "Item4" "Item5" "Item6"   
## [49] "Item7" "Item8"

df<-df[,c(11,15,16,20, 21,23,40,41,42)]  
colnames(df)

## [1] "Population" "Children" "Age"   
## [4] "Churn" "Outage\_sec\_perweek" "Contacts"   
## [7] "Tenure" "MonthlyCharge" "Bandwidth\_GB\_Year"

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

df$dummy\_churn<-case\_when(  
 df$Churn == "Yes" ~ 1,  
 TRUE ~ 0  
)  
colnames(df)

## [1] "Population" "Children" "Age"   
## [4] "Churn" "Outage\_sec\_perweek" "Contacts"   
## [7] "Tenure" "MonthlyCharge" "Bandwidth\_GB\_Year"   
## [10] "dummy\_churn"

unique(df$dummy\_churn)

## [1] 0 1

Next, summary statistics will be generated for each of the numeric variables to gain an understanding of variation and central tendency. No further action is needed as none of the values appeared unreasonable.

# Summary statistics for each variable  
colnames(df)

## [1] "Population" "Children" "Age"   
## [4] "Churn" "Outage\_sec\_perweek" "Contacts"   
## [7] "Tenure" "MonthlyCharge" "Bandwidth\_GB\_Year"   
## [10] "dummy\_churn"

summary(df$Population)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 738 2910 9757 13168 111850

summary(df$Children)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 1.000 2.088 3.000 10.000

summary(df$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 35.00 53.00 53.08 71.00 89.00

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$Contacts)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 1.0000 0.9942 2.0000 7.0000

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)

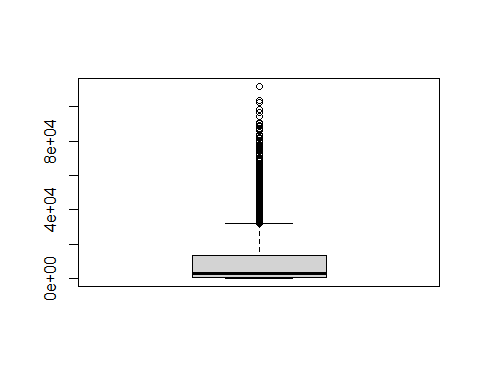
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 79.98 139.98 167.48 172.62 200.73 290.16

summary(df$Bandwidth\_GB\_Year)

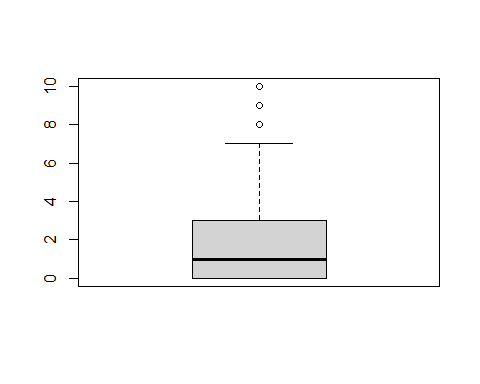
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 155.5 1236.5 3279.5 3392.3 5586.1 7159.0

Next, visualizations are produced to check the data preparedness and to inspect for anomalies. First, univariate visualizations of select variables did not show anything requiring attention.

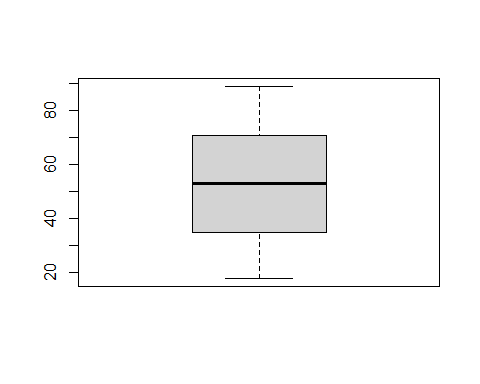
# Boxplots for each select variable  
boxplot(df$Population)



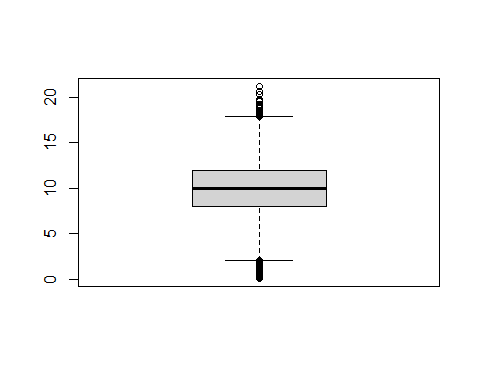
boxplot(df$Children)



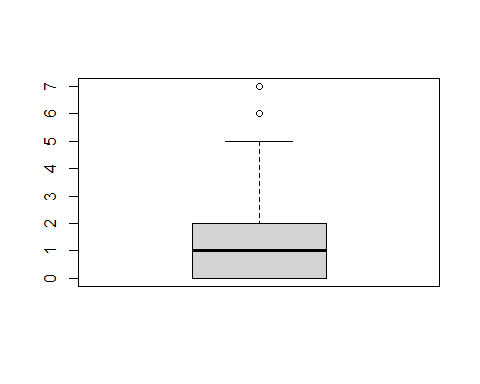
boxplot(df$Age)



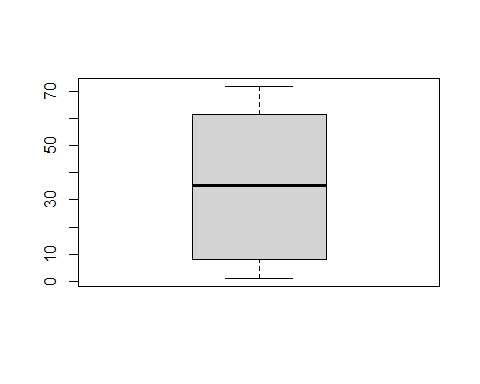
boxplot(df$Outage\_sec\_perweek)



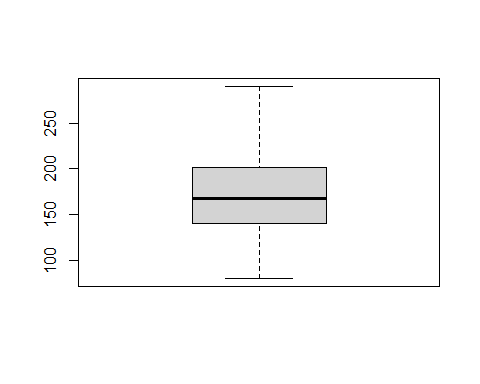
boxplot(df$Contacts)



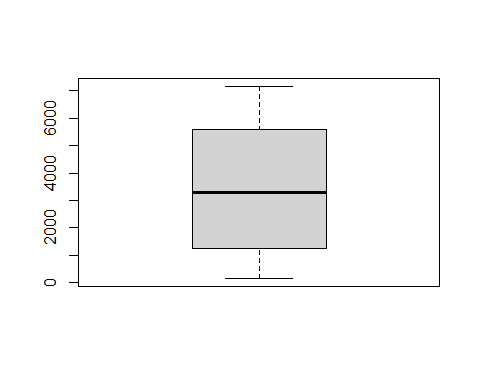
boxplot(df$Tenure)



boxplot(df$MonthlyCharge)

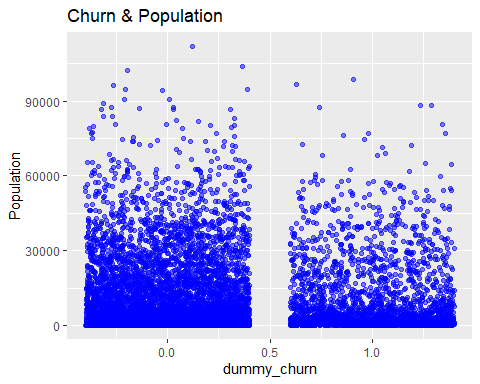


boxplot(df$Bandwidth\_GB\_Year)

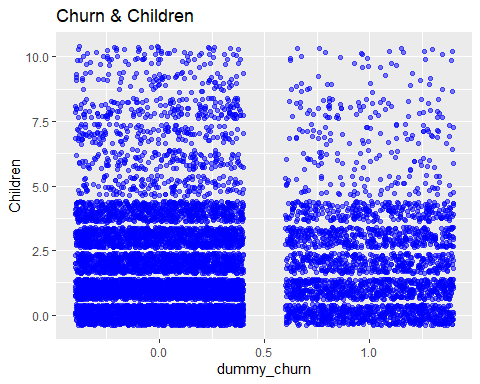


Finally, bivariate visualizations between the independent variables and the dependent variable (churn) are produced. For this the library “ggplot2” is used which is useful for plotting data within R. More information on the library ggplot2 can be found on the tidyverse website (<https://ggplot2.tidyverse.org/>). After review of the bivariate visualizations, the data preparedness was determined to be sufficient.

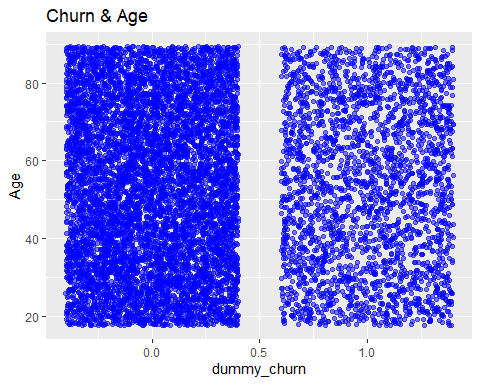
# Scatter plots for each variable  
library(ggplot2)  
ggplot(df, aes(x=dummy\_churn, y=Population))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Churn & Population")



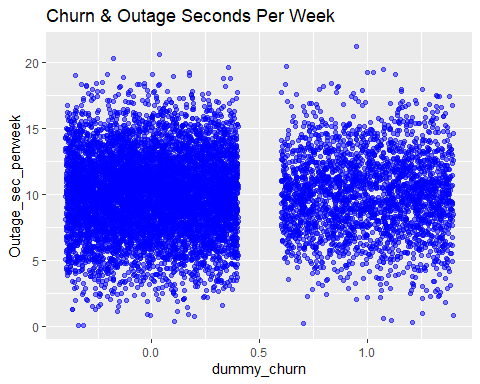
ggplot(df, aes(x=dummy\_churn, y=Children))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Churn & Children")



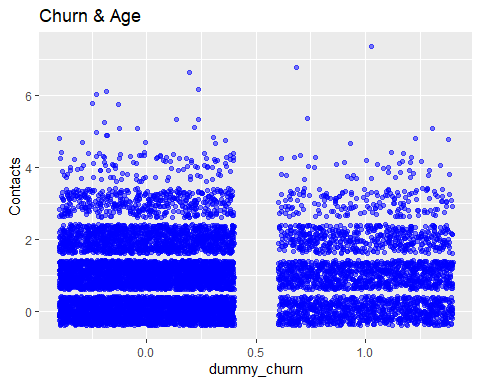
ggplot(df, aes(x=dummy\_churn, y=Age))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Churn & Age")



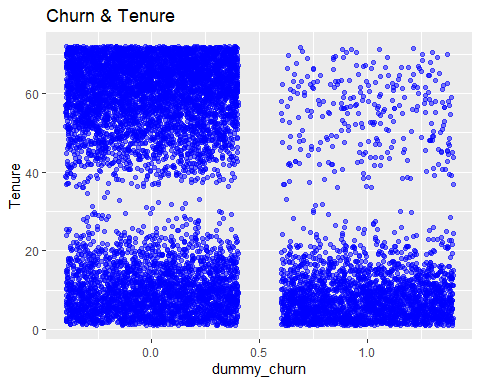
ggplot(df, aes(x=dummy\_churn, y=Outage\_sec\_perweek))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Churn & Outage Seconds Per Week")



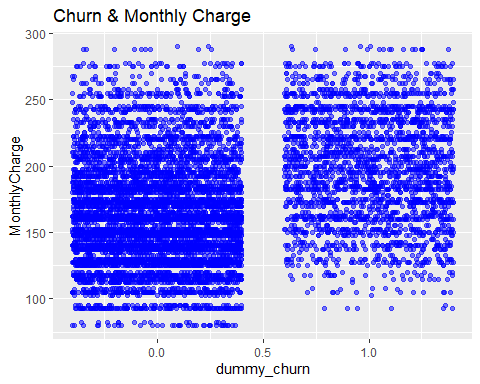
ggplot(df, aes(x=dummy\_churn, y=Contacts))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Churn & Age")



ggplot(df, aes(x=dummy\_churn, y=Tenure))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Churn & Tenure")



ggplot(df, aes(x=dummy\_churn, y=MonthlyCharge))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Churn & Monthly Charge")



The following prepared data set will be used for the analysis going forward: “prepared\_churn\_clean.csv”.

write.csv(df, "c:/users/shua/documents/Predictive Modeling\_D208/prepared\_churn\_clean.csv")

# D. Model Comparison and analysis

An initial logistic regression model is prepared using all of the variables identified during the data preparation phase.

The results of the initial logistic regression model show an overall accuracy at predicting the churn outcome of 83%! This is very good. The sensitivity of the model which is the proportion of correctly predicting a positive churn result, however, is only 60%. The model’s specificity, which is the proportion of correctly predicting negative churn results, is 92%. Since our model is performing 30 points higher at predicting a negative results than it is at predicting a positive churn result, we will attempt to reduce the variables used in the model to achieve a higher sensitivity score while maintaining the specificity score and therefore the overall accuracy of the model as well.

We observe in the summary of the model that only the variables of Tenure and Monthly Charge were notated as statistically significant. Therefore these will be our only two explanatory variables for the reduced model.

# Initial Model construction  
initial\_glm<-glm(dummy\_churn ~ Population + Children + Age + Outage\_sec\_perweek + Contacts + Tenure + MonthlyCharge, data=df, family = binomial)  
#Result of initial logistic regression model  
initial\_glm

##   
## Call: glm(formula = dummy\_churn ~ Population + Children + Age + Outage\_sec\_perweek +   
## Contacts + Tenure + MonthlyCharge, family = binomial, data = df)  
##   
## Coefficients:  
## (Intercept) Population Children Age   
## -5.213e+00 -2.552e-06 -7.994e-03 1.626e-03   
## Outage\_sec\_perweek Contacts Tenure MonthlyCharge   
## -1.280e-03 2.861e-02 -7.390e-02 3.318e-02   
##   
## Degrees of Freedom: 9999 Total (i.e. Null); 9992 Residual  
## Null Deviance: 11560   
## Residual Deviance: 6838 AIC: 6854

# Summary of initial logistic regression model  
summary(initial\_glm)

##   
## Call:  
## glm(formula = dummy\_churn ~ Population + Children + Age + Outage\_sec\_perweek +   
## Contacts + Tenure + MonthlyCharge, family = binomial, data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6329 -0.5617 -0.1789 0.3338 3.2007   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.213e+00 2.024e-01 -25.755 <2e-16 \*\*\*  
## Population -2.552e-06 2.120e-06 -1.204 0.229   
## Children -7.994e-03 1.432e-02 -0.558 0.577   
## Age 1.626e-03 1.467e-03 1.109 0.267   
## Outage\_sec\_perweek -1.280e-03 1.017e-02 -0.126 0.900   
## Contacts 2.861e-02 3.052e-02 0.937 0.349   
## Tenure -7.390e-02 1.768e-03 -41.799 <2e-16 \*\*\*  
## MonthlyCharge 3.318e-02 8.943e-04 37.102 <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: 6837.8 on 9992 degrees of freedom  
## AIC: 6853.8  
##   
## Number of Fisher Scoring iterations: 6

# Predicted responses  
pred\_response<-round(fitted(initial\_glm))  
# Actual responses  
act\_response<-df$dummy\_churn  
# Confusion Matrix  
outcomes<-table(pred\_response, act\_response)  
outcomes

## act\_response  
## pred\_response 0 1  
## 0 6734 1046  
## 1 616 1604

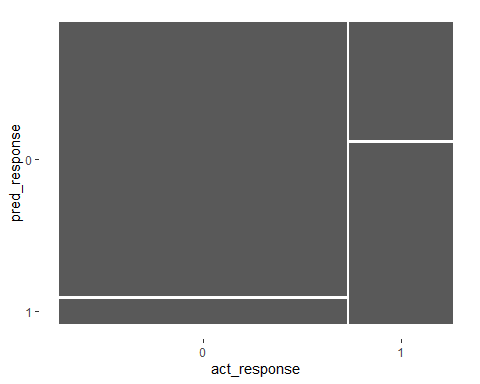
summary(outcomes)

## Number of cases in table: 10000   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 3066.7, df = 1, p-value = 0

# Yardstick confusion matrix  
#install.packages("yardstick")  
library(yardstick)

## For binary classification, the first factor level is assumed to be the event.  
## Use the argument `event\_level = "second"` to alter this as needed.

library(ggplot2)  
confusion<-conf\_mat(outcomes)  
autoplot(confusion)



summary(confusion, event\_level="second")

## # A tibble: 13 × 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 accuracy binary 0.834  
## 2 kap binary 0.550  
## 3 sens binary 0.605  
## 4 spec binary 0.916  
## 5 ppv binary 0.723  
## 6 npv binary 0.866  
## 7 mcc binary 0.554  
## 8 j\_index binary 0.521  
## 9 bal\_accuracy binary 0.761  
## 10 detection\_prevalence binary 0.222  
## 11 precision binary 0.723  
## 12 recall binary 0.605  
## 13 f\_meas binary 0.659

Reducing the variables, we construct a new model as well as a new confusion matrix of the actual vs predicted responses of the new model. However, in viewing the results we do not observe any improvement in the overall accuracy of the model. We also do not observe improvement greater than a percentage point in either the sensitivity or specificity in the new model.

# Reduced Model construction  
reduced\_glm<-glm(dummy\_churn ~ Tenure + MonthlyCharge, data=df, family = binomial)  
#Result of initial logistic regression model  
reduced\_glm

##   
## Call: glm(formula = dummy\_churn ~ Tenure + MonthlyCharge, family = binomial,   
## data = df)  
##   
## Coefficients:  
## (Intercept) Tenure MonthlyCharge   
## -5.15250 -0.07385 0.03317   
##   
## Degrees of Freedom: 9999 Total (i.e. Null); 9997 Residual  
## Null Deviance: 11560   
## Residual Deviance: 6842 AIC: 6848

# Summary of initial logistic regression model  
summary(reduced\_glm)

##   
## Call:  
## glm(formula = dummy\_churn ~ Tenure + MonthlyCharge, family = binomial,   
## data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6690 -0.5620 -0.1788 0.3351 3.1891   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.1524988 0.1502928 -34.28 <2e-16 \*\*\*  
## Tenure -0.0738459 0.0017671 -41.79 <2e-16 \*\*\*  
## MonthlyCharge 0.0331743 0.0008937 37.12 <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: 6841.7 on 9997 degrees of freedom  
## AIC: 6847.7  
##   
## Number of Fisher Scoring iterations: 6

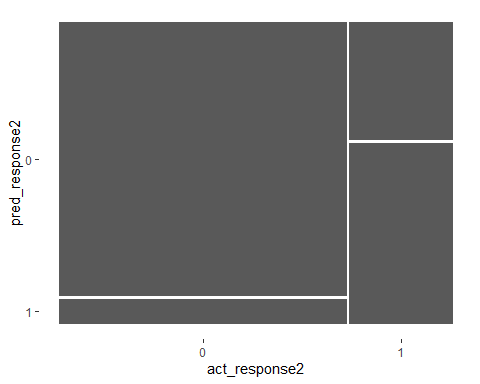
# Predicted responses  
pred\_response2<-round(fitted(reduced\_glm))  
# Actual responses  
act\_response2<-df$dummy\_churn  
# Confusion Matrix  
outcomes2<-table(pred\_response2, act\_response2)  
outcomes2

## act\_response2  
## pred\_response2 0 1  
## 0 6738 1050  
## 1 612 1600

summary(outcomes2)

## Number of cases in table: 10000   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 3063.2, df = 1, p-value = 0

# Yardstick confusion matrix  
#install.packages("yardstick")  
library(yardstick)  
library(ggplot2)  
confusion2<-conf\_mat(outcomes2)  
autoplot(confusion2)



summary(confusion2, event\_level="second")

## # A tibble: 13 × 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 accuracy binary 0.834  
## 2 kap binary 0.550  
## 3 sens binary 0.604  
## 4 spec binary 0.917  
## 5 ppv binary 0.723  
## 6 npv binary 0.865  
## 7 mcc binary 0.553  
## 8 j\_index binary 0.521  
## 9 bal\_accuracy binary 0.760  
## 10 detection\_prevalence binary 0.221  
## 11 precision binary 0.723  
## 12 recall binary 0.604  
## 13 f\_meas binary 0.658

# E. Analysis

The initial logistic regression model attempted to predict a response variable of customer churn based the following explanatory variables:

1. Population
2. Children
3. Age
4. Outage seconds per week
5. Contacts
6. Tenure
7. Monthly Charge

The results of the initial model predicted the churn outcome with 83.38% accuracy. The initial model was better at predicting a negative churn outcome demonstrating 91.61 specificity, while the sensitivity metric was only 60.52% of predicting a positive churn outcome.

The summary of the initial logistic model only indicated two of the original 7 explanatory variables as being statistically significant.

1. Tenure
2. Monthly Charge

The model therefore was reduced to only two explanatory variables to predict the response variable of churn. This reduction however did not result in any meaningful improvement. The sensitivity actually decreased slightly from 60.52% to 60.37%. The specificity conversely had a small increase from 91.61% to 91.67%. Overall the accuracy of the model remained exactly the same at 83.38%.

# F. Summary

The reduced model having an accuracy of 83.38% can be used as a good predictor of customer churn, with the understanding that it is much better at predicting when a customer will not churn than when they will.

The reduced model showed regression equation of logit(y)=-5.15 + -.07*Tenure + .03*Monthly Charge. An intercept of -5.15 showed that when all other explanatory variables are zero, the log(odds of churn) were -5.14. A negative .07 coefficient value for Tenure shows that for every month of Tenure the log(odds of churn) decreased by .07. A positive .03 coefficient value for Monthly charge indicates that for every one dollar of monthly charge the log(odds of churn) increased by .03.

The model is of course limited to the data set provided. It would be both interesting and beneficial to understand additional factors such as whether the customer joined the provider on any type of offertory rate. Since the goal of the company is to reduce churn, and this model does not perform as well at predicting customers who will churn as it does predicting customers who will not churn, it is recommended to investigate further commonalities among those customer who will not churn such as whether they joined on any type of special offers. By identifying those customers who will not churn, it would be beneficial for the company to further understand this portion of their customer base.

Further, one way this model could prove very useful is by targeting marketing efforts. Since this model has a high specificity value, it is recommended to use this model to identify those customers who are not likely to churn so as to exclude or deprioritize them from future customer retention marketing efforts. This will help the organization save precious time, effort, and resources and provide a good view to the company of the effectiveness of any new initiatives since these customers are already identified as not likely to churn.