IEMS 304 Lab: TelePrime Data Analysis

2022-04-13

We know that TelePrime wishes to maximize its customer retention rate and therefore also would like us to come up with a predictive model for customers who may have a tendency to switch from TelePrime. This analysis focuses on the behavior of telecom customers who are more likely to leave the platform.

About the data:

- Churn: customers who left in the last month

- Services that each customer has signed up for: phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges

- Demographic info about customers – gender, age range, and if they have partners and dependents

Why is this important in real life?

In business applications, customer churn is an important metric to track because lost customers equal lost revenue. If a company loses enough customers, it can have a serious impact on its bottom line. With all the features that we have, we intend to find out the most striking behavior of customers through exploring data analysis and later on use predictive analytics techniques to determine the customers who are most likely to churn.

1. Import the churn.csv to a variable named telco. Using Rstudio, view the data to get a overview of the schema, formatting etc.

```
telco <- read.csv('/Users/kevinli/Downloads/churn.csv')
```

We read in the CSV file.

2. Like with any real-world datasets, often we may not have all the data available. Find where those missing values are and remove the rows containing a missing value. Note complete.cases is a function that returns a logical vector indicating which cases are complete. How many incomplete cases are there?

```
dim(telco[!complete.cases(telco),])
```

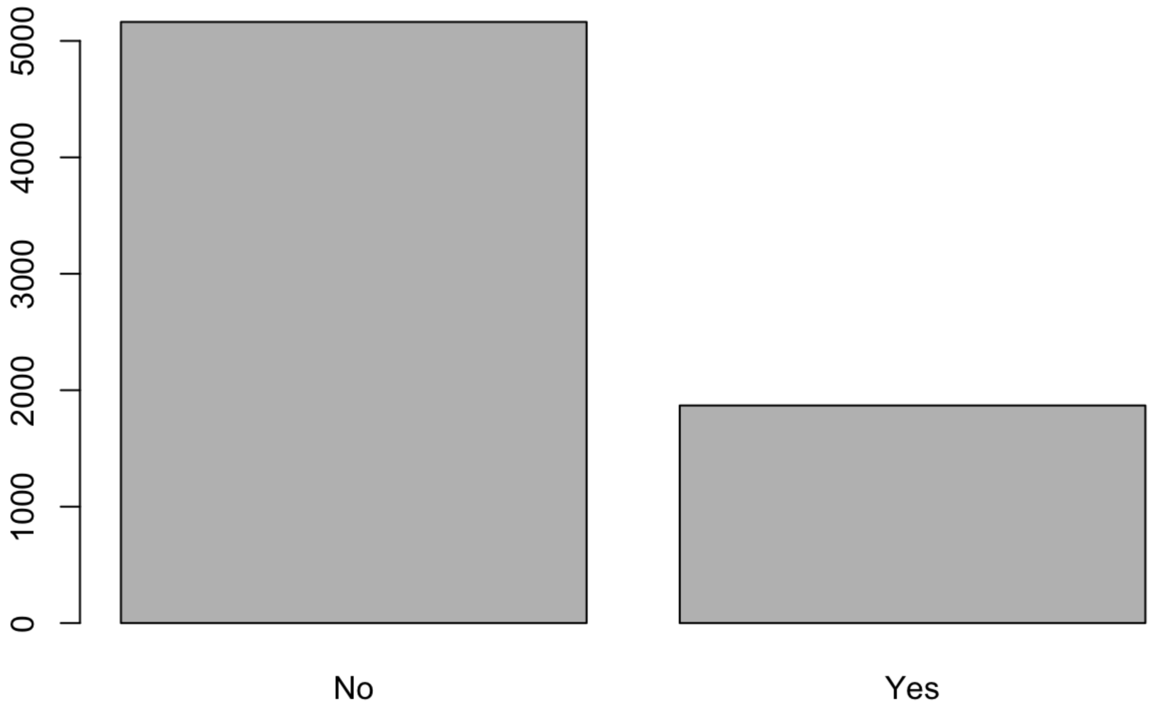
```
## [1] 11 23
```

```
missing_rows <- telco[!complete.cases(telco),]
telco <- na.omit(telco)
```

[11 23] means that there are 11 rows of incomplete data.

3. The variable of interest (i.e. the response variable or Churn) is obviously important. Count the number of Yes and Nos in Churn and display it in a bar plot. Hint: use barplot and table

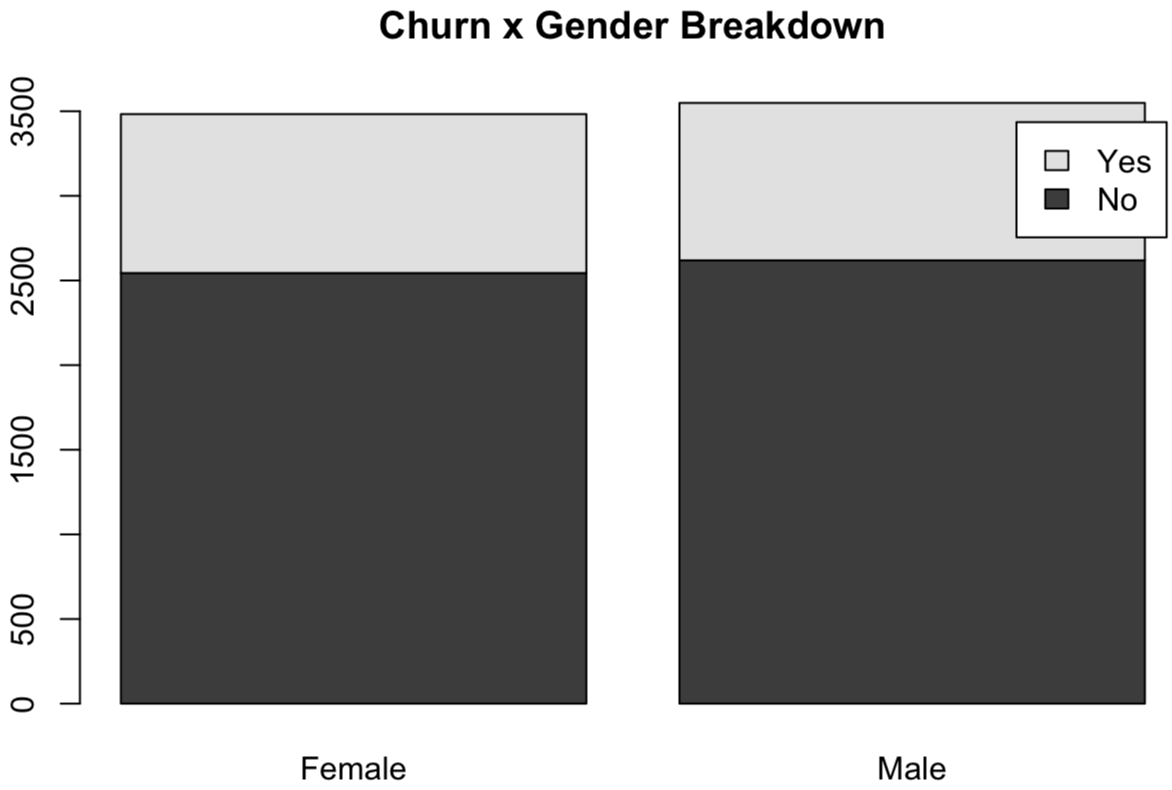
```
barplot(table(telco$Churn))
```



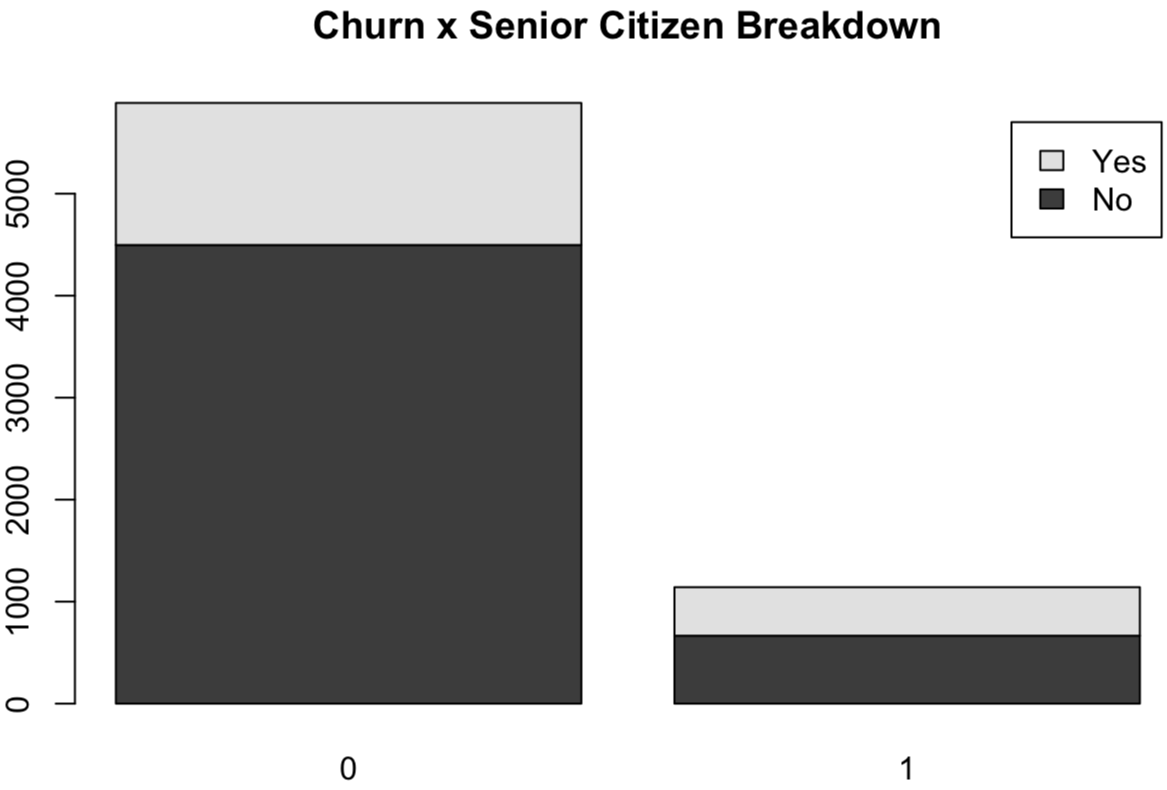
We see that there are a lot more ’No’s than ’Yes’s. Overall, around 26% of customers leave the platform within the last month.

4. But how do the other variables relate to churn? Use stacked barplots to explore the relationship between gender, Senior Citizen, Partner, Dependent,and other categorical data that you may wish to explore. The first plot is provided for you. Is there any trend that you have noticed?

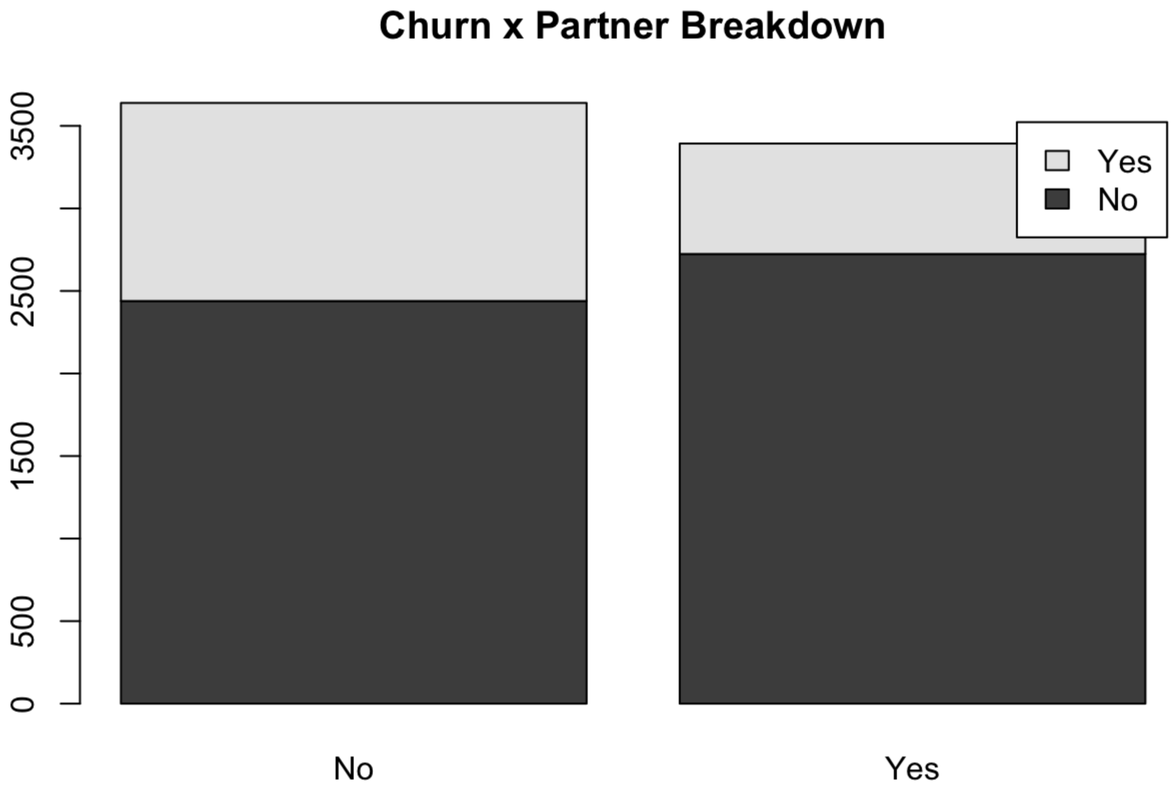
```
counts = table(telco$Churn, telco$gender)
barplot(counts, legend = rownames(counts),main="Churn x Gender Breakdown")
```



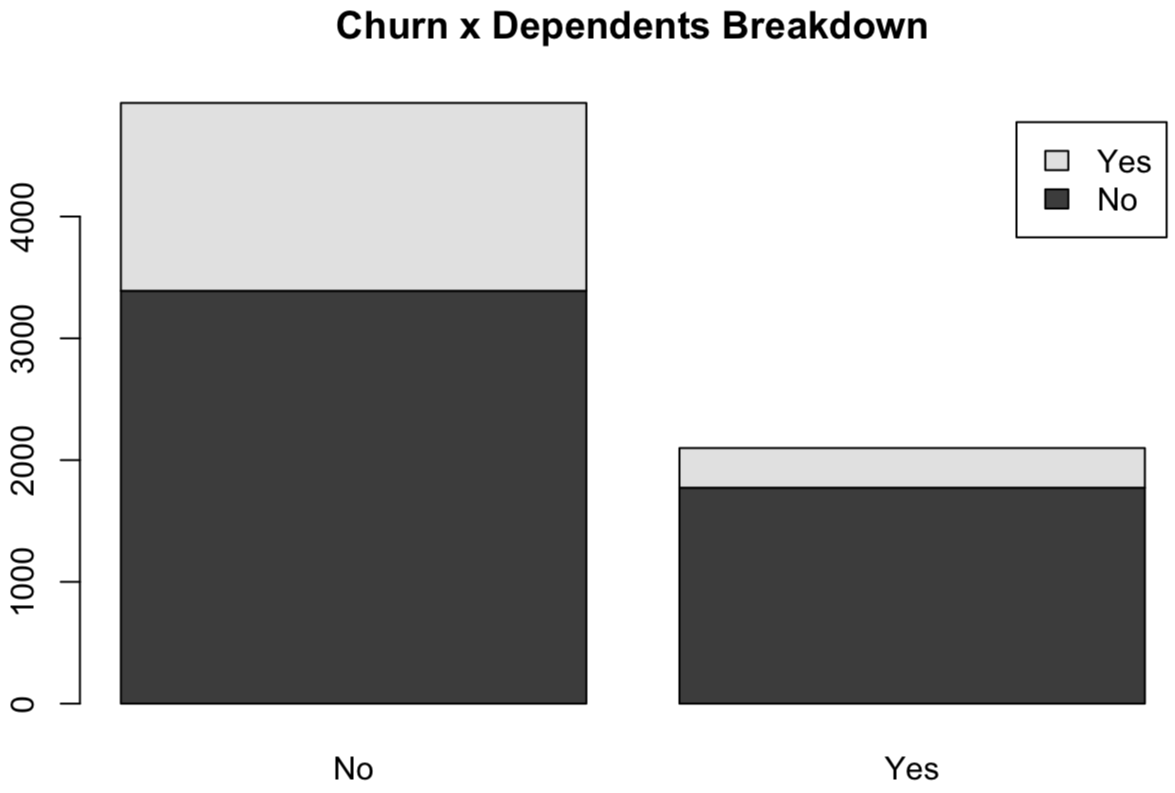
```
counts_2 = table(telco$Churn, telco$SeniorCitizen)
barplot(counts_2, legend = rownames(counts_2),main="Churn x Senior Citizen Breakdown")
```



```
counts_3 = table(telco$Churn, telco$Partner)
barplot(counts_3, legend = rownames(counts_3),main="Churn x Partner Breakdown")
```



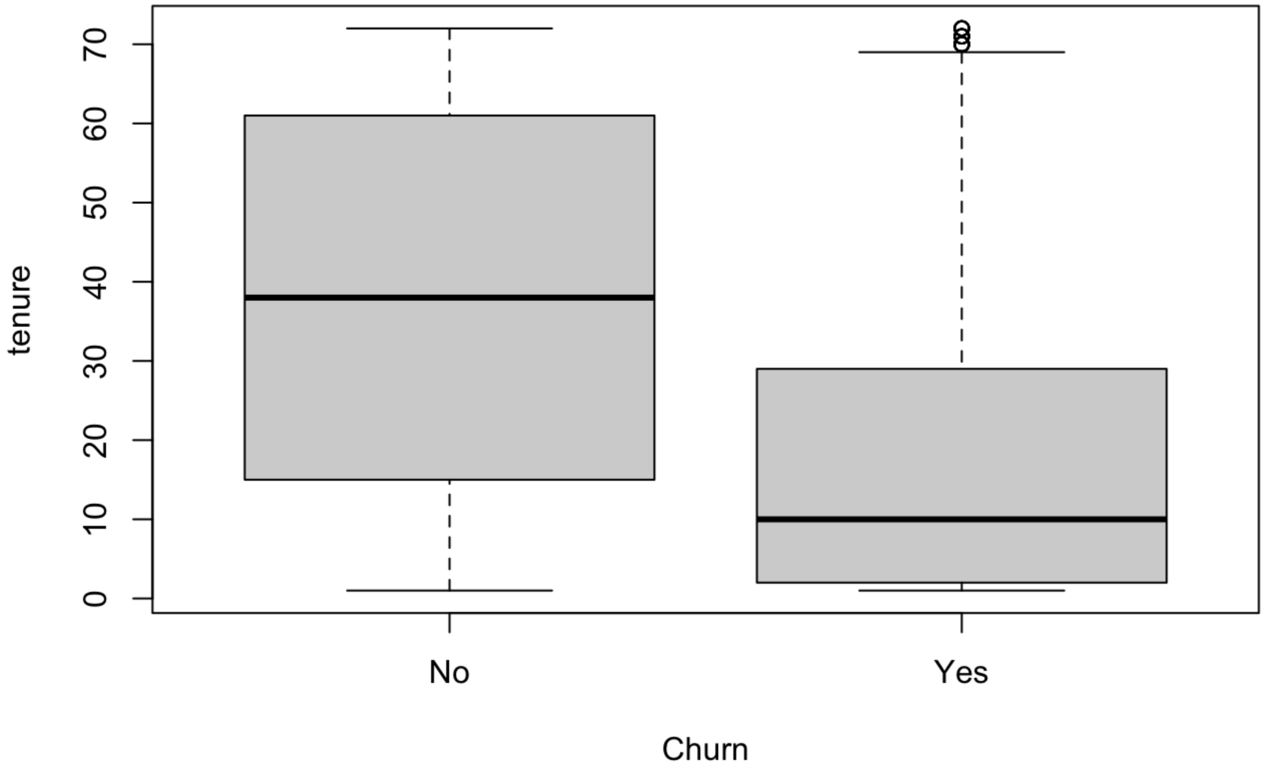
```
counts_4 = table(telco$Churn, telco$Dependents)
barplot(counts_4, legend = rownames(counts_4),main="Churn x Dependents Breakdown")
```



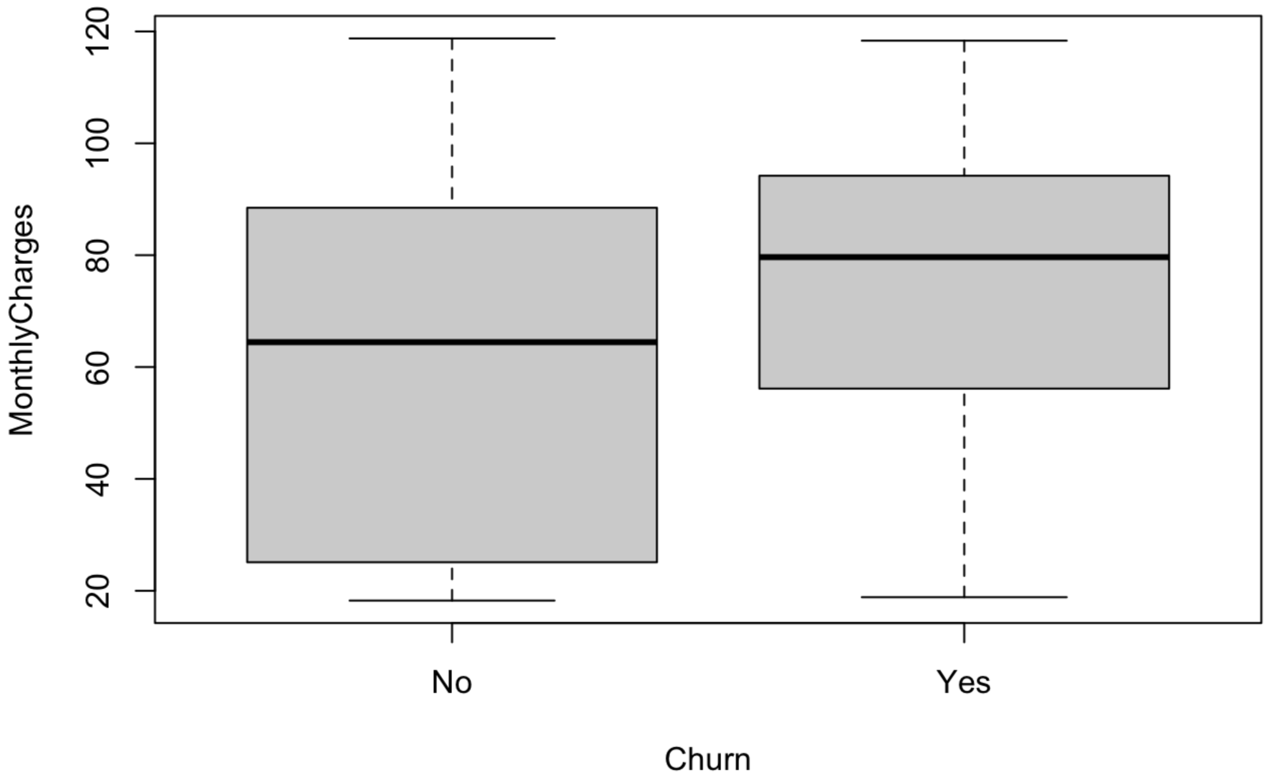
There are a few observations we can make. Churn rate is higher in senior citizens. Churn rate is higher in customers with no partners or no dependents (single). In contrast, churn rate is lower in case of customers with partners or dependents (family).

5. Now let’s explore some of the numerical data using boxplots. Create a box plot that shows the distribution off tenure for customers who have churn and thos who have not. Feel free to explore some more for more data points.

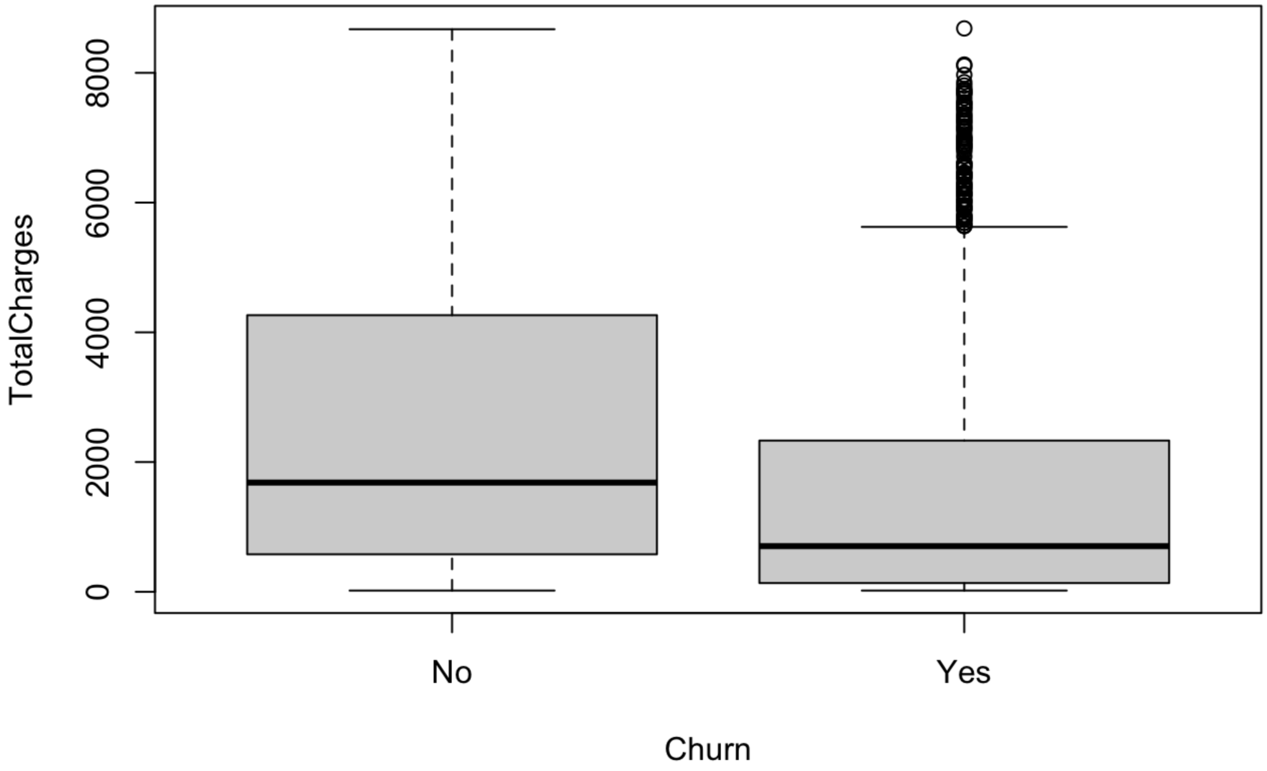
```
boxplot(tenure ~ Churn, data=telco)
```



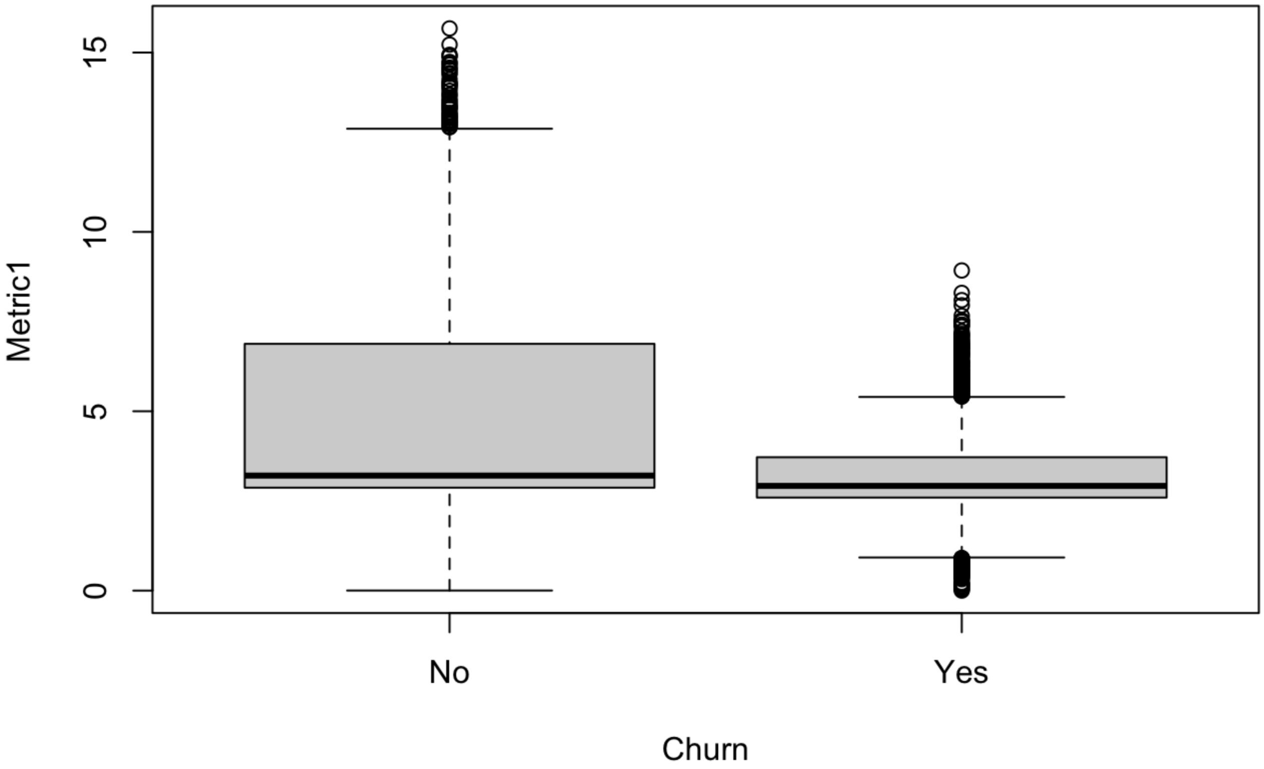
```
boxplot(MonthlyCharges ~ Churn, data=telco)
```



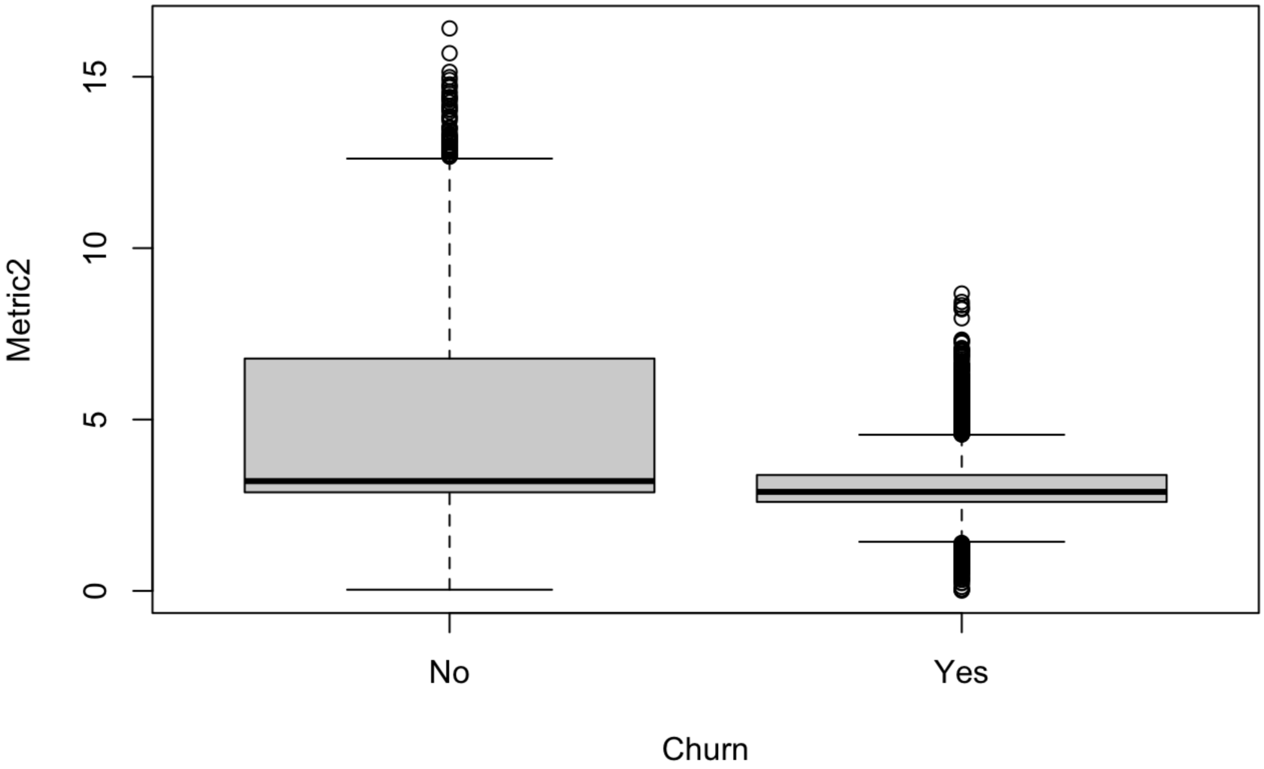
```
boxplot(TotalCharges ~ Churn, data=telco)
```



```
boxplot(Metric1 ~ Churn, data=telco)
```



```
boxplot(Metric2 ~ Churn, data=telco)
```



- Tenure: the median tenure for customers who churn is around 10 months. The lower tenure, the more churn percent.
- Monthly charges: the median monthly charges for customers who churn is above $75. The higher monthly charges, the more churn percent.
- Total charges: the median total charges for customers who churn is around $700. The lower total charges, the more churn percent.

6. Let’s now compute the correlation between the continuous variables

```
cor(telco[,c("tenure", "MonthlyCharges","TotalCharges","Metric1","Metric2")])
```

```
##                    tenure MonthlyCharges TotalCharges     Metric1     Metric2
## tenure         1.00000000     0.24686177   0.82588046  0.09403654  0.09295148
## MonthlyCharges 0.24686177     1.00000000   0.65106480 -0.04478265 -0.05898434
## TotalCharges   0.82588046     0.65106480   1.00000000  0.05024178  0.05811042
## Metric1        0.09403654    -0.04478265   0.05024178  1.00000000 -0.28637676
## Metric2        0.09295148    -0.05898434   0.05811042 -0.28637676  1.00000000
```

One observation here is that total charges is positive correlated with monthly charges and tenure.

8. First Replace “No internet service” and “No phone service” with just no. This cleans our categorical features so that it’s more consistent.

```
changer <- function(x) {
  if (x == "No internet service" || x == "No phone service") {
    x = "No"
  }
  else {
    x = x
  }
}
telco$OnlineSecurity <- sapply(telco$OnlineSecurity, changer)
telco$OnlineBackup<- sapply(telco$OnlineBackup, changer)
telco$DeviceProtection <- sapply(telco$DeviceProtection, changer)
telco$TechSupport <- sapply(telco$TechSupport, changer)
telco$StreamingTV <- sapply(telco$StreamingTV, changer)
telco$StreamingMovies <- sapply(telco$StreamingMovies, changer)
```

This is probably not the most efficient method. I wrote a function called 'Changer' and applied it to all the categorical columns.

9. Next, we wish to standardise the continuous features. To do this, first ensure that all the continuous features are in fact numeric, then one could use the scale function in R to normalise an entire column. This ensures that all the scales are consistent while not distorting the data. Let's temporarily put the result of the normalised continous data points to a variable named telco.

```
#identify all the continuous features
num_columns <- c("tenure", "MonthlyCharges","TotalCharges","Metric1","Metric2")
# apply as.numeric to all the numerical columns
telco[num_columns] <- sapply(telco[num_columns], as.numeric)
telco_int <- telco[,c("tenure", "MonthlyCharges","TotalCharges","Metric1","Metric2")]
telco_int <- data.frame(scale(telco_int)) # normalize the telco_int with scale function
```

Sometimes certain features maybe combined to form a better feature. Create a new feature "Metric3" that's the multiple of "Metric1" and "Metric2".Sometimes in real data sets, this step would come after the experimentation with models.

```
telco_int$Metric3 = telco_int$Metric1 * telco_int$Metric2
```

10. We can also try to create some derived features based on the continuous variables. Rather than just normalising it, we could also put them into bins. Add 1 a feature variable "tenure_bin" that takes on values of "tenure01","tenure12","tenure23","tenure34","tenure45"or "tenure56". For instance, "tenure01" feature should be 1 for those who had the account between 0 and 1 year, for all other people, this feature should be 0. Sometimes discretising of continuous variables can be helpful in statistics; however, in other cases they may prematurely limit the predictive ability of your model.

```
telco$tenure.bin = telco$tenure
telco$tenure.bin[telco$tenure.bin >= 0 & telco$tenure.bin <= 12] = "0-1 year"
telco$tenure.bin[telco$tenure.bin > 12 & telco$tenure.bin <= 24] = "1-2 years"
telco$tenure.bin[telco$tenure.bin > 24 & telco$tenure.bin <= 36] = "2-3 years"
telco$tenure.bin[telco$tenure.bin > 36 & telco$tenure.bin <= 48] = "3-4 years"
telco$tenure.bin[telco$tenure.bin > 48 & telco$tenure.bin <= 60] = "4-5 years"
telco$tenure.bin[telco$tenure.bin > 60 & telco$tenure.bin <= 72] = "5-6 years"

# convert it to factor so we can ensure that R knows it's categorical
telco$tenure.bin <- as.factor(telco$tenure.bin)
# delete tenure col
telco$tenure = NULL
```

11. Finally, we need to convert the categorical data to dummy indicator variables. This is given since in the last lab it’s assessed already.

```
#The numbers inside the bracket are simply the column numbers we are referring to
telco_cat = telco[,-c(1,3,18,19,20,21)]
dummy<- data.frame(sapply(telco_cat,function(x) data.frame(model.matrix(~x-1,data =telco
_cat))[,-1]))
head(dummy)
```

```
##   gender Partner Dependents PhoneService MultipleLines.xNo.phone.service
## 1      0       1          0            0                               1
## 2      1       0          0            1                               0
## 3      1       0          0            1                               0
## 4      1       0          0            0                               1
## 5      0       0          0            1                               0
## 6      0       0          0            1                               0
##   MultipleLines.xYes InternetService.xFiber.optic InternetService.xNo
## 1                  0                            0                   0
## 2                  0                            0                   0
## 3                  0                            0                   0
## 4                  0                            0                   0
## 5                  0                            1                   0
## 6                  1                            1                   0
##   OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV
## 1              0            1                0           0           0
## 2              1            0                1           0           0
## 3              1            1                0           0           0
## 4              1            0                1           1           0
## 5              0            0                0           0           0
## 6              0            0                1           0           1
##   StreamingMovies Contract.xOne.year Contract.xTwo.year PaperlessBilling
## 1               0                  0                  0                1
## 2               0                  1                  0                0
## 3               0                  0                  0                1
## 4               0                  1                  0                0
## 5               0                  0                  0                1
## 6               1                  0                  0                1
##   PaymentMethod.xCredit.card..automatic. PaymentMethod.xElectronic.check
## 1                                      0                               1
## 2                                      0                               0
## 3                                      0                               0
## 4                                      0                               0
## 5                                      0                               1
## 6                                      0                               1
##   PaymentMethod.xMailed.check Churn tenure.bin.x1.2.years tenure.bin.x2.3.years
## 1                           0     0                     0                     0
## 2                           1     0                     0                     1
## 3                           1     1                     0                     0
## 4                           0     0                     0                     0
## 5                           0     1                     0                     0
## 6                           0     1                     0                     0
##   tenure.bin.x3.4.years tenure.bin.x4.5.years tenure.bin.x5.6.years
## 1                     0                     0                     0
## 2                     0                     0                     0
## 3                     0                     0                     0
## 4                     1                     0                     0
## 5                     0                     0                     0
## 6                     0                     0                     0
```

12. Create the final data-set by combining the numeric and dummy data frames

```
# the cbind function combine telco_int and dummy horizontally
telco_final <- cbind(telco_int, dummy)
head(telco_final)
```

```
##        tenure MonthlyCharges TotalCharges    Metric1    Metric2    Metric3
## 1 -1.28015700     -1.1616113   -0.9941234 -0.2560194 -0.6817265  0.1745352
## 2  0.06429811     -0.2608594   -0.1737275  1.7945460 -0.6237738 -1.1193907
## 3 -1.23941594     -0.3638974   -0.9595809 -0.6542681 -0.6903462  0.4516715
## 4  0.51244982     -0.7477972   -0.1952338 -0.6502118  1.2978171 -0.8438559
## 5 -1.23941594      0.1961642   -0.9403906 -0.5557458 -0.5163032  0.2869333
## 6 -0.99496955      1.1584066   -0.6453233  1.1166705 -0.5449280 -0.6085050
##   gender Partner Dependents PhoneService MultipleLines.xNo.phone.service
## 1      0       1          0            0                               1
## 2      1       0          0            1                               0
## 3      1       0          0            1                               0
## 4      1       0          0            0                               1
## 5      0       0          0            1                               0
## 6      0       0          0            1                               0
##   MultipleLines.xYes InternetService.xFiber.optic InternetService.xNo
## 1                  0                            0                   0
## 2                  0                            0                   0
## 3                  0                            0                   0
## 4                  0                            0                   0
## 5                  0                            1                   0
## 6                  1                            1                   0
##   OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV
## 1              0            1                0           0           0
## 2              1            0                1           0           0
## 3              1            1                0           0           0
## 4              1            0                1           1           0
## 5              0            0                0           0           0
## 6              0            0                1           0           1
##   StreamingMovies Contract.xOne.year Contract.xTwo.year PaperlessBilling
## 1               0                  0                  0                1
## 2               0                  1                  0                0
## 3               0                  0                  0                1
## 4               0                  1                  0                0
## 5               0                  0                  0                1
## 6               1                  0                  0                1
##   PaymentMethod.xCredit.card..automatic. PaymentMethod.xElectronic.check
## 1                                      0                               1
## 2                                      0                               0
## 3                                      0                               0
## 4                                      0                               0
## 5                                      0                               1
## 6                                      0                               1
##   PaymentMethod.xMailed.check Churn tenure.bin.x1.2.years tenure.bin.x2.3.years
## 1                           0     0                     0                     0
## 2                           1     0                     0                     1
## 3                           1     1                     0                     0
## 4                           0     0                     0                     0
## 5                           0     1                     0                     0
## 6                           0     1                     0                     0
##   tenure.bin.x3.4.years tenure.bin.x4.5.years tenure.bin.x5.6.years
## 1                     0                     0                     0
## 2                     0                     0                     0
## 3                    0                    0                    0
## 4                    1                    0                    0
## 5                    0                    0                    0
## 6                    0                    0                    0
```

13. Using sample.split(from caTools) to split the dataset into a training dataset and a testing dataset. Set the SplitRatio to 0.7. Preserve the relative ratio of the Churn label.

```
#remember to install the package 'caTools' before running this code
library("caTools")
#Getting the indexes for the training data
indices = sample.split(telco_final$Churn, SplitRatio=0.7)
train = telco_final[indices,]
#validation here simply refers to the testing dataset
validation = telco_final[!(indices),]
```

14. Logistic Regression! Start by giving the logistic regressor the entire training dataset. Use the glm method and specify family = “binomial” Put the model into a variable named model_1 and print the summary of this model. Notice that there are a lot of variables, but many of which have very high p values.

We are building the first model using all variables

```
model_1 = glm(Churn ~ . , data = train, family="binomial") # specify that that we wish t
o predict Churn from all variables
summary(model_1)
```

```
## 
## Call:
## glm(formula = Churn ~ ., family = "binomial", data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.7169  -0.4219  -0.1225   0.3576   2.9871  
## 
## Coefficients: (1 not defined because of singularities)
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            -4.474924   1.878537  -2.382  0.01721
## tenure                                 -2.609899   0.368996  -7.073 1.52e-12
## MonthlyCharges                          0.284610   1.355990   0.210  0.83375
## TotalCharges                            0.372790   0.223021   1.672  0.09461
## Metric1                                -4.338005   0.398283 -10.892  < 2e-16
## Metric2                                -4.487485   0.408210 -10.993  < 2e-16
## Metric3                                -5.288911   0.666188  -7.939 2.04e-15
## gender                                  0.006942   0.092402   0.075  0.94012
## Partner                                -0.072302   0.109996  -0.657  0.51098
## Dependents                             -0.154016   0.124946  -1.233  0.21770
## PhoneService                           -0.687783   0.920688  -0.747  0.45504
## MultipleLines.xNo.phone.service               NA         NA      NA       NA
## MultipleLines.xYes                      0.413334   0.248599   1.663  0.09638
## InternetService.xFiber.optic            0.499906   1.133493   0.441  0.65919
## InternetService.xNo                    -0.753097   1.142934  -0.659  0.50995
## OnlineSecurity                         -0.383682   0.253157  -1.516  0.12962
## OnlineBackup                           -0.278481   0.249736  -1.115  0.26481
## DeviceProtection                       -0.025298   0.249186  -0.102  0.91914
## TechSupport                            -0.483720   0.257736  -1.877  0.06054
## StreamingTV                             0.214371   0.464342   0.462  0.64432
## StreamingMovies                        -0.023476   0.462636  -0.051  0.95953
## Contract.xOne.year                     -0.907185   0.153228  -5.920 3.21e-09
## Contract.xTwo.year                     -1.393861   0.232905  -5.985 2.17e-09
## PaperlessBilling                        0.484870   0.106886   4.536 5.72e-06
## PaymentMethod.xCredit.card..automatic. -0.123348   0.161129  -0.766  0.44396
## PaymentMethod.xElectronic.check         0.112845   0.133691   0.844  0.39863
## PaymentMethod.xMailed.check            -0.094894   0.164741  -0.576  0.56460
## tenure.bin.x1.2.years                   0.231915   0.221973   1.045  0.29612
## tenure.bin.x2.3.years                   0.963987   0.368083   2.619  0.00882
## tenure.bin.x3.4.years                   2.222084   0.522677   4.251 2.12e-05
## tenure.bin.x4.5.years                   3.028149   0.675944   4.480 7.47e-06
## tenure.bin.x5.6.years                   3.561342   0.842388   4.228 2.36e-05
##                                                                             
## (Intercept)                            *  
## tenure                                 ***
## MonthlyCharges                            
## TotalCharges                           .  
## Metric1                                ***
## Metric2                                ***
## Metric3                                ***
## gender                                    
## Partner                                   
## Dependents
## PhoneService
## MultipleLines.xNo.phone.service
## MultipleLines.xYes                     .
## InternetService.xFiber.optic
## InternetService.xNo
## OnlineSecurity
## OnlineBackup
## DeviceProtection
## TechSupport                            .
## StreamingTV
## StreamingMovies
## Contract.xOne.year                     ***
## Contract.xTwo.year                     ***
## PaperlessBilling                       ***
## PaymentMethod.xCredit.card..automatic.
## PaymentMethod.xElectronic.check
## PaymentMethod.xMailed.check
## tenure.bin.x1.2.years
## tenure.bin.x2.3.years                  **
## tenure.bin.x3.4.years                  ***
## tenure.bin.x4.5.years                  ***
## tenure.bin.x5.6.years                  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5699.5  on 4921  degrees of freedom
## Residual deviance: 2989.3  on 4891  degrees of freedom
## AIC: 3051.3
##
## Number of Fisher Scoring iterations: 7
```

glm stands for Generalized Linear Model. It allows us to specify the ‘binomial’ family, which means that we are using logistics regression. When we look at the summary, many of the features have super high p-values, which means it is not useful.

Using a function stepAIC from the MASS library, we could iteratively add and remove variables to improve model performance. The best part of it is that it’s mostly automatic.

```
library(MASS)
model_2<- stepAIC(model_1, direction="both")
```

```
## Start:  AIC=3051.29
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 + 
##     Metric3 + gender + Partner + Dependents + PhoneService + 
##     MultipleLines.xNo.phone.service + MultipleLines.xYes + InternetService.xFiber.opt
ic + 
##     InternetService.xNo + OnlineSecurity + OnlineBackup + DeviceProtection + 
##     TechSupport + StreamingTV + StreamingMovies + Contract.xOne.year + 
##     Contract.xTwo.year + PaperlessBilling + PaymentMethod.xCredit.card..automatic. + 
##     PaymentMethod.xElectronic.check + PaymentMethod.xMailed.check + 
##     tenure.bin.x1.2.years + tenure.bin.x2.3.years + tenure.bin.x3.4.years + 
##     tenure.bin.x4.5.years + tenure.bin.x5.6.years
## 
## 
## Step:  AIC=3051.29
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 + 
##     Metric3 + gender + Partner + Dependents + PhoneService + 
##     MultipleLines.xYes + InternetService.xFiber.optic + InternetService.xNo + 
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + 
##     StreamingTV + StreamingMovies + Contract.xOne.year + Contract.xTwo.year + 
##     PaperlessBilling + PaymentMethod.xCredit.card..automatic. + 
##     PaymentMethod.xElectronic.check + PaymentMethod.xMailed.check + 
##     tenure.bin.x1.2.years + tenure.bin.x2.3.years + tenure.bin.x3.4.years + 
##     tenure.bin.x4.5.years + tenure.bin.x5.6.years
## 
##                                           Df Deviance    AIC
## - StreamingMovies                          1   2989.3 3049.3
## - gender                                   1   2989.3 3049.3
## - DeviceProtection                         1   2989.3 3049.3
## - MonthlyCharges                           1   2989.3 3049.3
## - InternetService.xFiber.optic             1   2989.5 3049.5
## - StreamingTV                              1   2989.5 3049.5
## - PaymentMethod.xMailed.check              1   2989.6 3049.6
## - Partner                                  1   2989.7 3049.7
## - InternetService.xNo                      1   2989.7 3049.7
## - PhoneService                             1   2989.8 3049.8
## - PaymentMethod.xCredit.card..automatic.  1   2989.9 3049.9
## - PaymentMethod.xElectronic.check          1   2990.0 3050.0
## - tenure.bin.x1.2.years                    1   2990.4 3050.4
## - OnlineBackup                             1   2990.5 3050.5
## - Dependents                               1   2990.8 3050.8
## <none>                                        2989.3 3051.3
## - OnlineSecurity                           1   2991.6 3051.6
## - MultipleLines.xYes                       1   2992.1 3052.1
## - TotalCharges                             1   2992.2 3052.2
## - TechSupport                              1   2992.8 3052.8
## - tenure.bin.x2.3.years                    1   2996.2 3056.2
## - tenure.bin.x5.6.years                    1   3007.2 3067.2
## - tenure.bin.x3.4.years                    1   3007.4 3067.5
## - tenure.bin.x4.5.years                    1   3009.5 3069.5
## - PaperlessBilling                         1   3010.0 3070.0
## - Contract.xOne.year                       1   3026.1 3086.1
## - Contract.xTwo.year                       1   3029.4 3089.5
## - tenure                                  1   3041.1 3101.1
## - Metric3                                 1   3061.9 3121.9
## - Metric1                                 1   3148.3 3208.3
## - Metric2                                 1   3149.6 3209.6
## 
## Step:  AIC=3049.29
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 + 
##     Metric3 + gender + Partner + Dependents + PhoneService + 
##     MultipleLines.xYes + InternetService.xFiber.optic + InternetService.xNo + 
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + 
##     StreamingTV + Contract.xOne.year + Contract.xTwo.year + PaperlessBilling + 
##     PaymentMethod.xCredit.card..automatic. + PaymentMethod.xElectronic.check + 
##     PaymentMethod.xMailed.check + tenure.bin.x1.2.years + tenure.bin.x2.3.years + 
##     tenure.bin.x3.4.years + tenure.bin.x4.5.years + tenure.bin.x5.6.years
## 
##                                          Df Deviance    AIC
## - gender                                  1   2989.3 3047.3
## - DeviceProtection                        1   2989.3 3047.3
## - PaymentMethod.xMailed.check             1   2989.6 3047.6
## - MonthlyCharges                          1   2989.7 3047.7
## - Partner                                 1   2989.7 3047.7
## - PaymentMethod.xCredit.card..automatic.  1   2989.9 3047.9
## - PaymentMethod.xElectronic.check         1   2990.0 3048.0
## - tenure.bin.x1.2.years                   1   2990.4 3048.4
## - Dependents                              1   2990.8 3048.8
## - StreamingTV                             1   2991.0 3049.0
## <none>                                        2989.3 3049.3
## - TotalCharges                            1   2992.2 3050.2
## - InternetService.xFiber.optic            1   2992.4 3050.4
## + StreamingMovies                         1   2989.3 3051.3
## - OnlineBackup                            1   2994.2 3052.2
## - PhoneService                            1   2994.5 3052.5
## - InternetService.xNo                     1   2994.5 3052.5
## - tenure.bin.x2.3.years                   1   2996.2 3054.2
## - OnlineSecurity                          1   2997.5 3055.5
## - MultipleLines.xYes                      1   3000.7 3058.7
## - TechSupport                             1   3001.6 3059.6
## - tenure.bin.x5.6.years                   1   3007.3 3065.3
## - tenure.bin.x3.4.years                   1   3007.5 3065.5
## - tenure.bin.x4.5.years                   1   3009.6 3067.5
## - PaperlessBilling                        1   3010.0 3068.0
## - Contract.xOne.year                      1   3026.1 3084.1
## - Contract.xTwo.year                      1   3029.5 3087.5
## - tenure                                  1   3041.3 3099.3
## - Metric3                                 1   3061.9 3119.9
## - Metric1                                 1   3148.3 3206.3
## - Metric2                                 1   3149.6 3207.6
## 
## Step:  AIC=3047.3
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 + 
##     Metric3 + Partner + Dependents + PhoneService + MultipleLines.xYes + 
##     InternetService.xFiber.optic + InternetService.xNo + OnlineSecurity + 
##     OnlineBackup + DeviceProtection + TechSupport + StreamingTV + 
##     Contract.xOne.year + Contract.xTwo.year + PaperlessBilling + 
##     PaymentMethod.xCredit.card..automatic. + PaymentMethod.xElectronic.check + 
##     PaymentMethod.xMailed.check + tenure.bin.x1.2.years + tenure.bin.x2.3.years + 
##     tenure.bin.x3.4.years + tenure.bin.x4.5.years + tenure.bin.x5.6.years
## 
##                                          Df Deviance    AIC
## - DeviceProtection                        1   2989.3 3045.3
## - PaymentMethod.xMailed.check             1   2989.6 3045.6
## - MonthlyCharges                          1   2989.7 3045.7
## - Partner                                 1   2989.7 3045.7
## - PaymentMethod.xCredit.card..automatic.  1   2989.9 3045.9
## - PaymentMethod.xElectronic.check         1   2990.0 3046.0
## - tenure.bin.x1.2.years                   1   2990.4 3046.4
## - Dependents                              1   2990.8 3046.8
## - StreamingTV                             1   2991.0 3047.0
## <none>                                        2989.3 3047.3
## - TotalCharges                            1   2992.2 3048.2
## - InternetService.xFiber.optic            1   2992.4 3048.4
## + gender                                  1   2989.3 3049.3
## + StreamingMovies                         1   2989.3 3049.3
## - OnlineBackup                            1   2994.3 3050.3
## - PhoneService                            1   2994.5 3050.5
## - InternetService.xNo                     1   2994.5 3050.5
## - tenure.bin.x2.3.years                   1   2996.2 3052.2
## - OnlineSecurity                          1   2997.5 3053.5
## - MultipleLines.xYes                      1   3000.7 3056.7
## - TechSupport                             1   3001.6 3057.6
## - tenure.bin.x5.6.years                   1   3007.3 3063.3
## - tenure.bin.x3.4.years                   1   3007.5 3063.5
## - tenure.bin.x4.5.years                   1   3009.6 3065.6
## - PaperlessBilling                        1   3010.0 3066.0
## - Contract.xOne.year                      1   3026.1 3082.1
## - Contract.xTwo.year                      1   3029.5 3085.5
## - tenure                                  1   3041.3 3097.3
## - Metric3                                 1   3061.9 3117.9
## - Metric1                                 1   3148.4 3204.4
## - Metric2                                 1   3149.7 3205.7
## 
## Step:  AIC=3045.31
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 + 
##     Metric3 + Partner + Dependents + PhoneService + MultipleLines.xYes + 
##     InternetService.xFiber.optic + InternetService.xNo + OnlineSecurity + 
##     OnlineBackup + TechSupport + StreamingTV + Contract.xOne.year + 
##     Contract.xTwo.year + PaperlessBilling + PaymentMethod.xCredit.card..automatic. + 
##     PaymentMethod.xElectronic.check + PaymentMethod.xMailed.check + 
##     tenure.bin.x1.2.years + tenure.bin.x2.3.years + tenure.bin.x3.4.years + 
##     tenure.bin.x4.5.years + tenure.bin.x5.6.years
## 
##                                          Df Deviance    AIC
## - PaymentMethod.xMailed.check             1   2989.6 3043.6
## - Partner                                 1   2989.8 3043.7
## - MonthlyCharges                         1   2989.8 3043.8
## - PaymentMethod.xCredit.card..automatic.  1   2989.9 3043.9
## - PaymentMethod.xElectronic.check         1   2990.0 3044.0
## - tenure.bin.x1.2.years                   1   2990.4 3044.4
## - Dependents                              1   2990.8 3044.8
## <none>                                        2989.3 3045.3
## - StreamingTV                             1   2991.4 3045.4
## - TotalCharges                            1   2992.2 3046.2
## + DeviceProtection                        1   2989.3 3047.3
## + gender                                  1   2989.3 3047.3
## + StreamingMovies                         1   2989.3 3047.3
## - InternetService.xFiber.optic            1   2993.4 3047.4
## - OnlineBackup                            1   2994.4 3048.4
## - PhoneService                            1   2995.2 3049.2
## - InternetService.xNo                     1   2995.6 3049.6
## - tenure.bin.x2.3.years                   1   2996.2 3050.2
## - OnlineSecurity                          1   2997.8 3051.8
## - MultipleLines.xYes                      1   3001.4 3055.5
## - TechSupport                             1   3001.9 3055.9
## - tenure.bin.x5.6.years                   1   3007.3 3061.3
## - tenure.bin.x3.4.years                   1   3007.6 3061.5
## - tenure.bin.x4.5.years                   1   3009.6 3063.6
## - PaperlessBilling                        1   3010.0 3064.0
## - Contract.xOne.year                      1   3026.3 3080.3
## - Contract.xTwo.year                      1   3029.7 3083.7
## - tenure                                  1   3041.3 3095.3
## - Metric3                                 1   3062.0 3116.0
## - Metric1                                 1   3148.4 3202.4
## - Metric2                                 1   3149.8 3203.8
##
## Step:  AIC=3043.64
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 +
##     Metric3 + Partner + Dependents + PhoneService + MultipleLines.xYes +
##     InternetService.xFiber.optic + InternetService.xNo + OnlineSecurity +
##     OnlineBackup + TechSupport + StreamingTV + Contract.xOne.year +
##     Contract.xTwo.year + PaperlessBilling + PaymentMethod.xCredit.card..automatic. +
##     PaymentMethod.xElectronic.check + tenure.bin.x1.2.years +
##     tenure.bin.x2.3.years + tenure.bin.x3.4.years + tenure.bin.x4.5.years +
##     tenure.bin.x5.6.years
##
##                                          Df Deviance    AIC
## - PaymentMethod.xCredit.card..automatic.  1   2990.0 3042.0
## - Partner                                 1   2990.0 3042.0
## - MonthlyCharges                          1   2990.1 3042.1
## - tenure.bin.x1.2.years                   1   2990.7 3042.7
## - Dependents                              1   2991.2 3043.2
## <none>                                        2989.6 3043.6
## - StreamingTV                             1   2991.7 3043.7
## - PaymentMethod.xElectronic.check         1   2991.7 3043.7
## - TotalCharges                            1   2992.4 3044.4
## + PaymentMethod.xMailed.check             1   2989.3 3045.3
## + DeviceProtection                        1   2989.6 3045.6
## + StreamingMovies                        1   2989.6 3045.6
## + gender                                 1   2989.6 3045.6
## - InternetService.xFiber.optic           1   2993.8 3045.8
## - OnlineBackup                           1   2994.7 3046.7
## - PhoneService                           1   2995.6 3047.6
## - InternetService.xNo                    1   2996.0 3048.0
## - tenure.bin.x2.3.years                  1   2996.4 3048.4
## - OnlineSecurity                         1   2998.2 3050.2
## - MultipleLines.xYes                     1   3001.7 3053.7
## - TechSupport                            1   3002.3 3054.3
## - tenure.bin.x5.6.years                  1   3007.5 3059.5
## - tenure.bin.x3.4.years                  1   3007.7 3059.7
## - tenure.bin.x4.5.years                  1   3009.8 3061.8
## - PaperlessBilling                       1   3010.5 3062.5
## - Contract.xOne.year                     1   3026.6 3078.6
## - Contract.xTwo.year                     1   3030.2 3082.2
## - tenure                                 1   3041.4 3093.4
## - Metric3                                1   3062.5 3114.5
## - Metric1                                1   3149.0 3201.0
## - Metric2                                1   3150.5 3202.5
## 
## Step:  AIC=3041.96
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 + 
##     Metric3 + Partner + Dependents + PhoneService + MultipleLines.xYes + 
##     InternetService.xFiber.optic + InternetService.xNo + OnlineSecurity + 
##     OnlineBackup + TechSupport + StreamingTV + Contract.xOne.year + 
##     Contract.xTwo.year + PaperlessBilling + PaymentMethod.xElectronic.check + 
##     tenure.bin.x1.2.years + tenure.bin.x2.3.years + tenure.bin.x3.4.years + 
##     tenure.bin.x4.5.years + tenure.bin.x5.6.years
## 
##                                          Df Deviance    AIC
## - Partner                                1   2990.3 3040.4
## - MonthlyCharges                         1   2990.4 3040.4
## - tenure.bin.x1.2.years                  1   2991.1 3041.0
## - Dependents                             1   2991.6 3041.6
## <none>                                       2990.0 3042.0
## - StreamingTV                            1   2992.0 3042.0
## - TotalCharges                           1   2992.8 3042.8
## - PaymentMethod.xElectronic.check        1   2993.3 3043.3
## + PaymentMethod.xCredit.card..automatic.  1   2989.6 3043.6
## + PaymentMethod.xMailed.check            1   2989.9 3043.9
## + DeviceProtection                       1   2989.9 3043.9
## + StreamingMovies                        1   2989.9 3044.0
## + gender                                 1   2989.9 3044.0
## - InternetService.xFiber.optic           1   2994.1 3044.1
## - OnlineBackup                           1   2995.0 3045.0
## - PhoneService                           1   2995.9 3045.9
## - InternetService.xNo                    1   2996.3 3046.3
## - tenure.bin.x2.3.years                  1   2996.8 3046.8
## - OnlineSecurity                         1   2998.4 3048.4
## - MultipleLines.xYes                     1   3002.1 3052.0
## - TechSupport                            1   3002.6 3052.6
## - tenure.bin.x5.6.years                    1   3007.9 3057.9
## - tenure.bin.x3.4.years                    1   3008.1 3058.1
## - tenure.bin.x4.5.years                    1   3010.2 3060.2
## - PaperlessBilling                         1   3010.6 3060.6
## - Contract.xOne.year                       1   3026.9 3076.9
## - Contract.xTwo.year                       1   3030.6 3080.6
## - tenure                                   1   3042.2 3092.2
## - Metric3                                  1   3062.7 3112.7
## - Metric1                                  1   3149.2 3199.2
## - Metric2                                  1   3150.6 3200.6
## 
## Step:  AIC=3040.35
## Churn ~ tenure + MonthlyCharges + TotalCharges + Metric1 + Metric2 + 
##     Metric3 + Dependents + PhoneService + MultipleLines.xYes + 
##     InternetService.xFiber.optic + InternetService.xNo + OnlineSecurity + 
##     OnlineBackup + TechSupport + StreamingTV + Contract.xOne.year + 
##     Contract.xTwo.year + PaperlessBilling + PaymentMethod.xElectronic.check + 
##     tenure.bin.x1.2.years + tenure.bin.x2.3.years + tenure.bin.x3.4.years + 
##     tenure.bin.x4.5.years + tenure.bin.x5.6.years
## 
##                                           Df Deviance    AIC
## - MonthlyCharges                           1   2990.8 3038.8
## - tenure.bin.x1.2.years                    1   2991.4 3039.4
## <none>                                         2990.3 3040.4
## - StreamingTV                              1   2992.4 3040.4
## - Dependents                               1   2993.1 3041.1
## - TotalCharges                             1   2993.2 3041.2
## - PaymentMethod.xElectronic.check          1   2993.7 3041.7
## + Partner                                  1   2990.0 3042.0
## + PaymentMethod.xCredit.card..automatic.   1   2990.0 3042.0
## + PaymentMethod.xMailed.check              1   2990.3 3042.3
## + DeviceProtection                         1   2990.3 3042.3
## + StreamingMovies                          1   2990.3 3042.3
## + gender                                   1   2990.3 3042.4
## - InternetService.xFiber.optic             1   2994.5 3042.5
## - OnlineBackup                             1   2995.3 3043.4
## - PhoneService                             1   2996.1 3044.1
## - InternetService.xNo                      1   2996.8 3044.8
## - tenure.bin.x2.3.years                    1   2997.2 3045.2
## - OnlineSecurity                           1   2998.9 3046.9
## - MultipleLines.xYes                       1   3002.3 3050.3
## - TechSupport                              1   3002.9 3050.9
## - tenure.bin.x5.6.years                    1   3008.4 3056.4
## - tenure.bin.x3.4.years                    1   3008.7 3056.7
## - tenure.bin.x4.5.years                    1   3010.7 3058.7
## - PaperlessBilling                         1   3011.0 3059.0
## - Contract.xOne.year                       1   3027.2 3075.2
## - Contract.xTwo.year                       1   3030.9 3078.9
## - tenure                                   1   3043.5 3091.5
## - Metric3                                  1   3063.0 3111.0
## - Metric1                                  1   3149.5 3197.5
## - Metric2                                  1   3150.8 3198.8
```

```
##
## Step:  AIC=3038.81
## Churn ~ tenure + TotalCharges + Metric1 + Metric2 + Metric3 +
##     Dependents + PhoneService + MultipleLines.xYes + InternetService.xFiber.optic +
##     InternetService.xNo + OnlineSecurity + OnlineBackup + TechSupport +
##     StreamingTV + Contract.xOne.year + Contract.xTwo.year + PaperlessBilling +
##     PaymentMethod.xElectronic.check + tenure.bin.x1.2.years +
##     tenure.bin.x2.3.years + tenure.bin.x3.4.years + tenure.bin.x4.5.years +
##     tenure.bin.x5.6.years
##
##                                           Df Deviance    AIC
## - tenure.bin.x1.2.years                    1   2991.9 3037.9
## <none>                                         2990.8 3038.8
## - Dependents                               1   2993.6 3039.6
## - PaymentMethod.xElectronic.check          1   2994.3 3040.3
## + MonthlyCharges                           1   2990.3 3040.4
## + StreamingMovies                          1   2990.4 3040.4
## + Partner                                  1   2990.4 3040.4
## + PaymentMethod.xCredit.card..automatic.   1   2990.5 3040.5
## + PaymentMethod.xMailed.check              1   2990.7 3040.7
## - TotalCharges                             1   2994.8 3040.7
## + DeviceProtection                         1   2990.8 3040.8
## + gender                                   1   2990.8 3040.8
## - OnlineBackup                             1   2995.4 3041.4
## - tenure.bin.x2.3.years                    1   2997.6 3043.6
## - PhoneService                             1   2998.0 3044.0
## - StreamingTV                              1   2999.0 3045.0
## - OnlineSecurity                           1   2999.0 3045.0
## - TechSupport                              1   3003.4 3049.5
## - MultipleLines.xYes                       1   3006.8 3052.8
## - tenure.bin.x5.6.years                    1   3008.6 3054.6
## - tenure.bin.x3.4.years                    1   3009.0 3055.0
## - tenure.bin.x4.5.years                    1   3010.9 3057.0
## - PaperlessBilling                         1   3011.6 3057.6
## - InternetService.xFiber.optic             1   3018.1 3064.1
## - InternetService.xNo                      1   3020.3 3066.3
## - Contract.xOne.year                       1   3027.2 3073.2
## - Contract.xTwo.year                       1   3030.9 3076.9
## - tenure                                   1   3044.9 3090.9
## - Metric3                                  1   3063.4 3109.4
## - Metric1                                  1   3150.0 3196.0
## - Metric2                                  1   3151.3 3197.3
##
## Step:  AIC=3037.93
## Churn ~ tenure + TotalCharges + Metric1 + Metric2 + Metric3 +
##     Dependents + PhoneService + MultipleLines.xYes + InternetService.xFiber.optic +
##     InternetService.xNo + OnlineSecurity + OnlineBackup + TechSupport +
##     StreamingTV + Contract.xOne.year + Contract.xTwo.year + PaperlessBilling +
##     PaymentMethod.xElectronic.check + tenure.bin.x2.3.years +
##     tenure.bin.x3.4.years + tenure.bin.x4.5.years + tenure.bin.x5.6.years
##
##                                           Df Deviance    AIC
## <none>                                        2991.9 3037.9
## - Dependents                              1   2994.6 3038.6
## + tenure.bin.x1.2.years                   1   2990.8 3038.8
## - PaymentMethod.xElectronic.check         1   2995.4 3039.4
## + MonthlyCharges                          1   2991.4 3039.4
## + StreamingMovies                         1   2991.5 3039.5
## + Partner                                 1   2991.6 3039.6
## - TotalCharges                            1   2995.6 3039.6
## + PaymentMethod.xCredit.card..automatic.  1   2991.6 3039.6
## + PaymentMethod.xMailed.check             1   2991.9 3039.9
## + DeviceProtection                        1   2991.9 3039.9
## + gender                                  1   2991.9 3039.9
## - OnlineBackup                            1   2996.4 3040.4
## - PhoneService                            1   2999.1 3043.1
## - OnlineSecurity                          1   3000.2 3044.2
## - tenure.bin.x2.3.years                   1   3000.3 3044.3
## - StreamingTV                             1   3000.4 3044.4
## - TechSupport                             1   3004.5 3048.5
## - MultipleLines.xYes                      1   3007.9 3051.9
## - PaperlessBilling                        1   3012.5 3056.5
## - InternetService.xFiber.optic            1   3019.5 3063.5
## - InternetService.xNo                     1   3021.6 3065.5
## - tenure.bin.x5.6.years                   1   3021.9 3065.9
## - tenure.bin.x3.4.years                   1   3024.4 3068.4
## - Contract.xOne.year                      1   3028.8 3072.7
## - tenure.bin.x4.5.years                   1   3028.8 3072.8
## - Contract.xTwo.year                      1   3033.1 3077.1
## - Metric3                                 1   3064.7 3108.7
## - tenure                                  1   3092.3 3136.3
## - Metric1                                 1   3151.2 3195.2
## - Metric2                                 1   3152.6 3196.6
```

```
summary(model_2)
```

```
##
## Call:
## glm(formula = Churn ~ tenure + TotalCharges + Metric1 + Metric2 +
##     Metric3 + Dependents + PhoneService + MultipleLines.xYes +
##     InternetService.xFiber.optic + InternetService.xNo + OnlineSecurity +
##     OnlineBackup + TechSupport + StreamingTV + Contract.xOne.year +
##     Contract.xTwo.year + PaperlessBilling + PaymentMethod.xElectronic.check +
##     tenure.bin.x2.3.years + tenure.bin.x3.4.years + tenure.bin.x4.5.years +
##     tenure.bin.x5.6.years, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7416  -0.4190  -0.1236   0.3610   2.9580
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -4.57191    0.39409 -11.601  < 2e-16 ***
## tenure                          -2.34659    0.24511  -9.574  < 2e-16 ***
## TotalCharges                     0.39622    0.21091   1.879 0.060301 .
## Metric1                         -4.33987    0.39827 -10.897  < 2e-16 ***
## Metric2                         -4.48951    0.40823 -10.997  < 2e-16 ***
## Metric3                         -5.29251    0.66624  -7.944 1.96e-15 ***
## Dependents                      -0.18739    0.11422  -1.641 0.100888
## PhoneService                    -0.49893    0.18681  -2.671 0.007568 **
## MultipleLines.xYes               0.45404    0.11423   3.975 7.05e-05 ***
## InternetService.xFiber.optic     0.73908    0.14143   5.226 1.73e-07 ***
## InternetService.xNo             -1.01581    0.18934  -5.365 8.09e-08 ***
## OnlineSecurity                  -0.34364    0.11943  -2.877 0.004010 **
## OnlineBackup                    -0.23256    0.10951  -2.124 0.033692 *
## TechSupport                     -0.43378    0.12256  -3.539 0.000401 ***
## StreamingTV                      0.32861    0.11340   2.898 0.003760 **
## Contract.xOne.year              -0.90154    0.15222  -5.923 3.17e-09 ***
## Contract.xTwo.year              -1.39831    0.23112  -6.050 1.45e-09 ***
## PaperlessBilling                 0.48187    0.10667   4.518 6.26e-06 ***
## PaymentMethod.xElectronic.check  0.18679    0.09968   1.874 0.060943 .
## tenure.bin.x2.3.years            0.64929    0.22511   2.884 0.003922 **
## tenure.bin.x3.4.years            1.77599    0.31377   5.660 1.51e-08 ***
## tenure.bin.x4.5.years            2.44557    0.40663   6.014 1.81e-09 ***
## tenure.bin.x5.6.years            2.83036    0.51926   5.451 5.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5699.5  on 4921  degrees of freedom
## Residual deviance: 2991.9  on 4899  degrees of freedom
## AIC: 3037.9
##
## Number of Fisher Scoring iterations: 7
```

The Akaike information criterion is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. StepAIC function reduces the AIC visibly, as shown by the results. We could use p values to get rid of a few more redundant features, for instance StreamingMovies has a relatively high p value and can be removed.

15. Create a new logistics regression model based on features not pruned and print its summary. Feel free to tweak the model as you see fit.

```
model_3 <-glm(formula = Churn ~ tenure + MonthlyCharges + InternetService.xFiber.optic + InternetService.xNo + OnlineSecurity + StreamingTV + StreamingMovies + Contract.xOne.year + Contract.xTwo.year + PaperlessBilling + PaymentMethod.xElectronic.check + tenure.bin.x2.3.years + tenure.bin.x3.4.years + tenure.bin.x4.5.years + tenure.bin.x5.6.years, family = "binomial", data = train)
summary(model_3)
```

```
## 
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + InternetService.xFiber.optic + 
##     InternetService.xNo + OnlineSecurity + StreamingTV + StreamingMovies + 
##     Contract.xOne.year + Contract.xTwo.year + PaperlessBilling + 
##     PaymentMethod.xElectronic.check + tenure.bin.x2.3.years + 
##     tenure.bin.x3.4.years + tenure.bin.x4.5.years + tenure.bin.x5.6.years, 
##     family = "binomial", data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0203  -0.6782  -0.2852   0.6539   3.1842  
## 
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                     -3.17046    0.23984 -13.219  < 2e-16 ***
## tenure                          -1.84275    0.17207 -10.709  < 2e-16 ***
## MonthlyCharges                  -0.41855    0.16558  -2.528 0.011476 *  
## InternetService.xFiber.optic     1.38920    0.20568   6.754 1.44e-11 ***
## InternetService.xNo             -1.09958    0.19402  -5.667 1.45e-08 ***
## OnlineSecurity                  -0.28471    0.10669  -2.669 0.007616 ** 
## StreamingTV                      0.46074    0.11193   4.116 3.85e-05 ***
## StreamingMovies                  0.38053    0.11061   3.440 0.000581 ***
## Contract.xOne.year              -0.89634    0.13153  -6.815 9.45e-12 ***
## Contract.xTwo.year              -1.50109    0.20528  -7.313 2.62e-13 ***
## PaperlessBilling                 0.50056    0.08978   5.575 2.47e-08 ***
## PaymentMethod.xElectronic.check  0.32220    0.08347   3.860 0.000113 ***
## tenure.bin.x2.3.years            0.57140    0.19006   3.006 0.002643 ** 
## tenure.bin.x3.4.years            1.61303    0.26355   6.120 9.33e-10 ***
## tenure.bin.x4.5.years            2.22813    0.33777   6.597 4.21e-11 ***
## tenure.bin.x5.6.years            2.72117    0.42141   6.457 1.07e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 5699.5  on 4921  degrees of freedom
## Residual deviance: 4055.9  on 4906  degrees of freedom
## AIC: 4087.9
## 
## Number of Fisher Scoring iterations: 6
```

All of the features we have here have very low p-values.

16. Store the final model to final_model

```
final_model <- model_3
```

17. The following code gives the evaluated performance of your model based on a 50% cut off.

```
pred <- predict(final_model, type = "response", newdata = validation)
summary(pred)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00412 0.04414 0.19148 0.26136 0.43099 0.86905
```

```
# Using probability cutoff of 50%.
# if pred meets the cut off of 0.5 then "Yes" else "No"
pred_churn <- factor(ifelse(pred >= 0.5, "Yes", "No"))
# if validation$Churn is 1 then"Yes" else "No"
actual_churn <- factor(ifelse(validation$Churn == 1, "Yes", "No"))
performance = table(actual_churn,pred_churn)
performance
```

```
##             pred_churn
## actual_churn   No  Yes
##          No  1413  136
##          Yes  290  271
```

We see first from the summary that the “mean” is 26.9%, meaning that 26.9 % of customers will churn. This is quite similar to the boxplot we saw in the beginning of the lab. This means our accuracy is quite high. We also see that the true negative rate is much higher than the true positive rate. We correctly identify more negatives than

Accuracy: is the number of data points correctly identified over all data points.

Sensitivity: (True Positive rate) measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition).

Specificity: (True Negative rate) measures the proportion of negatives that are correctly identified (i.e. the proportion of those who do not have the condition (unaffected) who are correctly identified as not having the condition).

Ideally we would want our model to be have a balance between sensitivity, specificity, as opposed to simply optimising for accuracy and specificity. As using a cutoff of 0.5, we are getting a good accuracy and specificity. But, the sensitivity is not optimized. Hence, we need to find the optimal probability cutoff which will give us the maximum accuracy, sensitivity, and specificity.

19. This is our current results. You can see that the sensitivity is quite low.

```
#install.packages('caret')
#install.packages('dplyr')
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```r
pred = predict(final_model, type = "response", newdata = validation)
pred.cutoff = factor(ifelse(pred >= 0.5, "Yes", "No"))
actual = factor(ifelse(validation$Churn == 1, "Yes", "No"))
cm = confusionMatrix(pred.cutoff, actual, positive = "Yes")
cm.lr = cm
pred.lr = pred.cutoff

# generating result
Accuracy = cm$overall[1]
Sensitivity = cm$byClass[1]
Specificity = cm$byClass[2]
print(Accuracy)
```

```
##  Accuracy
## 0.7981043
```

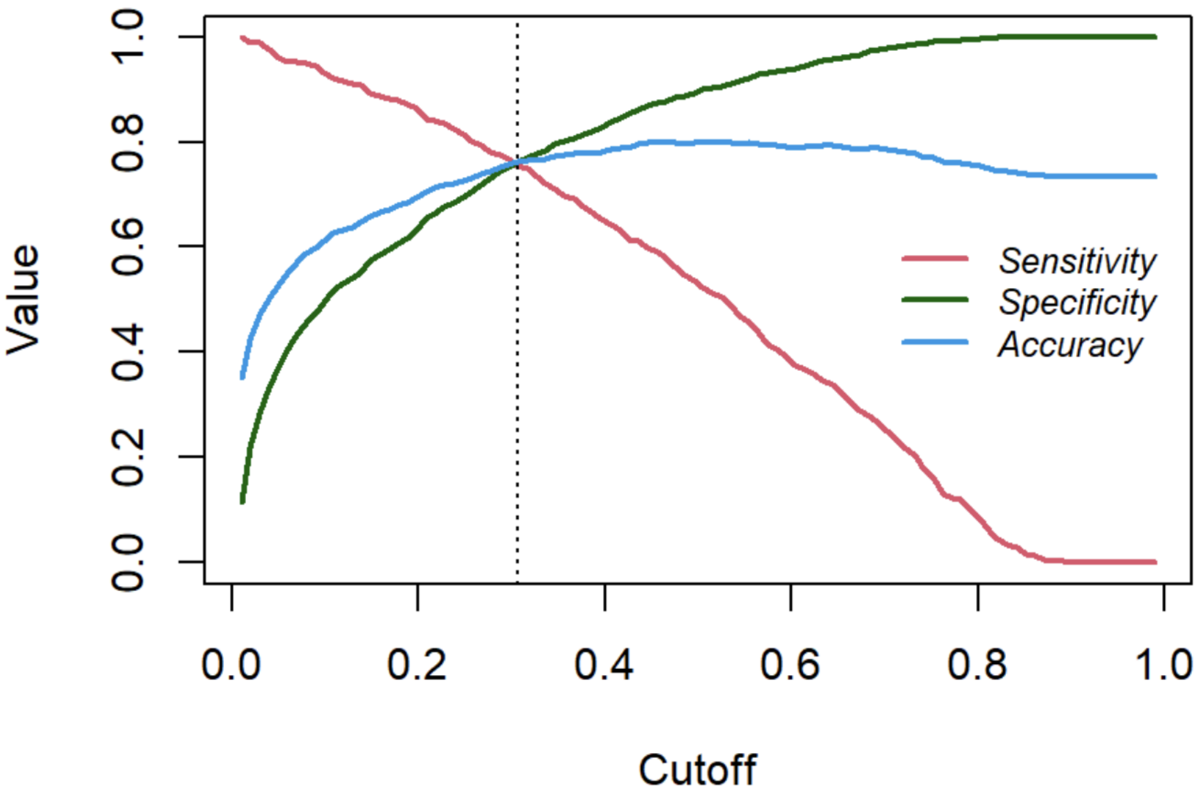
```r
print(Sensitivity)
```

```
## Sensitivity
##    0.483066
```

```r
print(Specificity)
```

```
## Specificity
##   0.9122014
```

This is a figure that I got from online. It shows sensitivity, specificity, accuracy as a function of the cutoff value. We see that the logistics regression with a cutoff probability value of 0.307 gives us better values.



Let’s evaluate our model again with the new cutoff value.

```
library(caret)
pred = predict(final_model, type = "response", newdata = validation)
pred.cutoff = factor(ifelse(pred >= 0.307, "Yes", "No"))
actual = factor(ifelse(validation$Churn == 1, "Yes", "No"))
cm = confusionMatrix(pred.cutoff, actual, positive = "Yes")
cm.lr = cm
pred.lr = pred.cutoff

# generating result
Accuracy = cm$overall[1]
Sensitivity = cm$byClass[1]
Specificity = cm$byClass[2]
print(Accuracy)
```

```
##  Accuracy
## 0.7720379
```

```
print(Sensitivity)
```

```
## Sensitivity
##   0.7522282
```

```
print(Specificity)
```

```
## Specificity
##   0.7792124
```

20. Use the tree library to produce a decision tree that solves this classification problem. Print a summary of the model. Note it is also helpful to plot and use the text function to generate a visualisation of the tree itself.

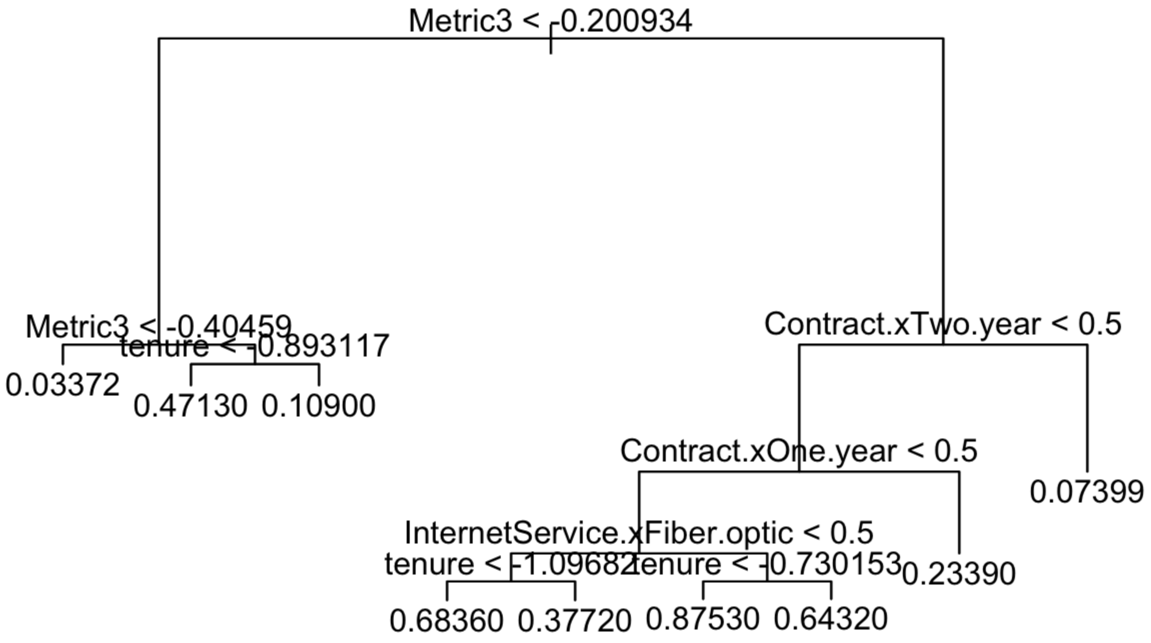
```
require(tree)
```

```
## Loading required package: tree
```

```
model_tree = tree(Churn ~ ., data = train)
summary(model_tree)
```

```
##
## Regression tree:
## tree(formula = Churn ~ ., data = train)
## Variables actually used in tree construction:
## [1] "Metric3"                       "tenure"
## [3] "Contract.xTwo.year"            "Contract.xOne.year"
## [5] "InternetService.xFiber.optic"
## Number of terminal nodes:  9
## Residual mean deviance:  0.1065 = 523.1 / 4913
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.87530 -0.07399 -0.03372  0.00000  0.12470  0.96630
```

```
plot(model_tree)
text(model_tree, pretty = 0)
```


Metric3 < -0.200934
Metric3 < -0.40459
tenure < -0.893117
0.03372
0.47130
0.10900
Contract.xTwo.year < 0.5
Contract.xOne.year < 0.5
0.07399
InternetService.xFiber.optic < 0.5
tenure < -1.09682
tenure < -0.730153
0.23390
0.68360
0.37720
0.87530
0.64320