TELCO CUSTOMER CHURN

Group Alpha:

Aytaj Khankishiyeva

Fatima Nurmakhamadova

David Belyaev

Vaibhav Arora

Xiaolu Shen

Class: ALY 6015 - 21495

Introduction

Objective

The main goal of this report is to build a predictive model to predict the churn rate of the telecom company.

Question

“What are significant predictor variables that affect the Telco customers churn rate ?”

Models Used

We have used 2 supervised machine learning models:

- Logistic Regression
- LASSO Regularization Regression

Data Set

- The data set used is <Telco Customer Churn> that was retrieved from Kaggle, and has 7043 observations of 21 variables.

- The dataset contains information about Telecom customers, their gender, tenure, charges, payment method, usage of internet, streaming service, and churn rate along with other variables.

Data Cleaning

```
> any(is.na(OGdata))
[1] TRUE
> #Checking the type of the variables and lengths
> summary(OGdata) # 11 Na values exist in the column 'TotalCharges'
      customerID      gender      SeniorCitizen    Partner     Dependents      tenure      PhoneService         MultipleLines     InternetService            OnlineSecurity              OnlineBackup            DeviceProtection
 0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933   Min.   : 0.00   No : 682     No              :3390   DSL        :2421   No              :3498   No              :3088   No              :3095
 0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110   1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096   No internet service:1526   No internet service:1526   No internet service:1526
 0004-TLHLJ:   1                 Median :0.0000                         Median :29.00                Yes             :2971   No         :1526   Yes             :2019   Yes             :2429   Yes             :2422
 0011-IGKFF:   1                 Mean   :0.1621                         Mean   :32.37
 0013-EXCHZ:   1                 3rd Qu.:0.0000                         3rd Qu.:55.00
 0013-MHZWF:   1                 Max.   :1.0000                         Max.   :72.00
 (Other)   :7037
              TechSupport                StreamingTV            StreamingMovies          Contract    PaperlessBilling                 PaymentMethod  MonthlyCharges    TotalCharges    Churn
 No                  :3473   No                  :2810   No                  :2785   Month-to-month:3875   No :2872         Bank transfer (automatic):1544   Min.   : 18.25   Min.   :  18.8   No :5174
 No internet service:1526   No internet service:1526   No internet service:1526   One year      :1473   Yes:4171         Credit card (automatic)  :1522   1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
 Yes                 :2044   Yes                 :2707   Yes                 :2732   Two year      :1695                  Electronic check         :2365   Median : 70.35   Median :1397.5
                                                                                                                       Mailed check             :1612   Mean   : 64.76   Mean   :2283.3
                                                                                                                                                        3rd Qu.: 89.85   3rd Qu.:3794.7
                                                                                                                                                        Max.   :118.75   Max.   :8684.8
                                                                                                                                                                         NA's   :11
>
```

```
> # Checking for missing values to make sure no more missing values are present
> any(is.na(OGdata))
[1] FALSE
>
```

- The original data set has 11 missing values in the column <TotalCharges>. Missing values are substituted with the column mean.
- Ambiguous levels from categorical variables were removed and assigned the corresponding value.
- We also remove the first column <CustomerID> that has no significance in the data analysis.

Exploratory Data Analysis

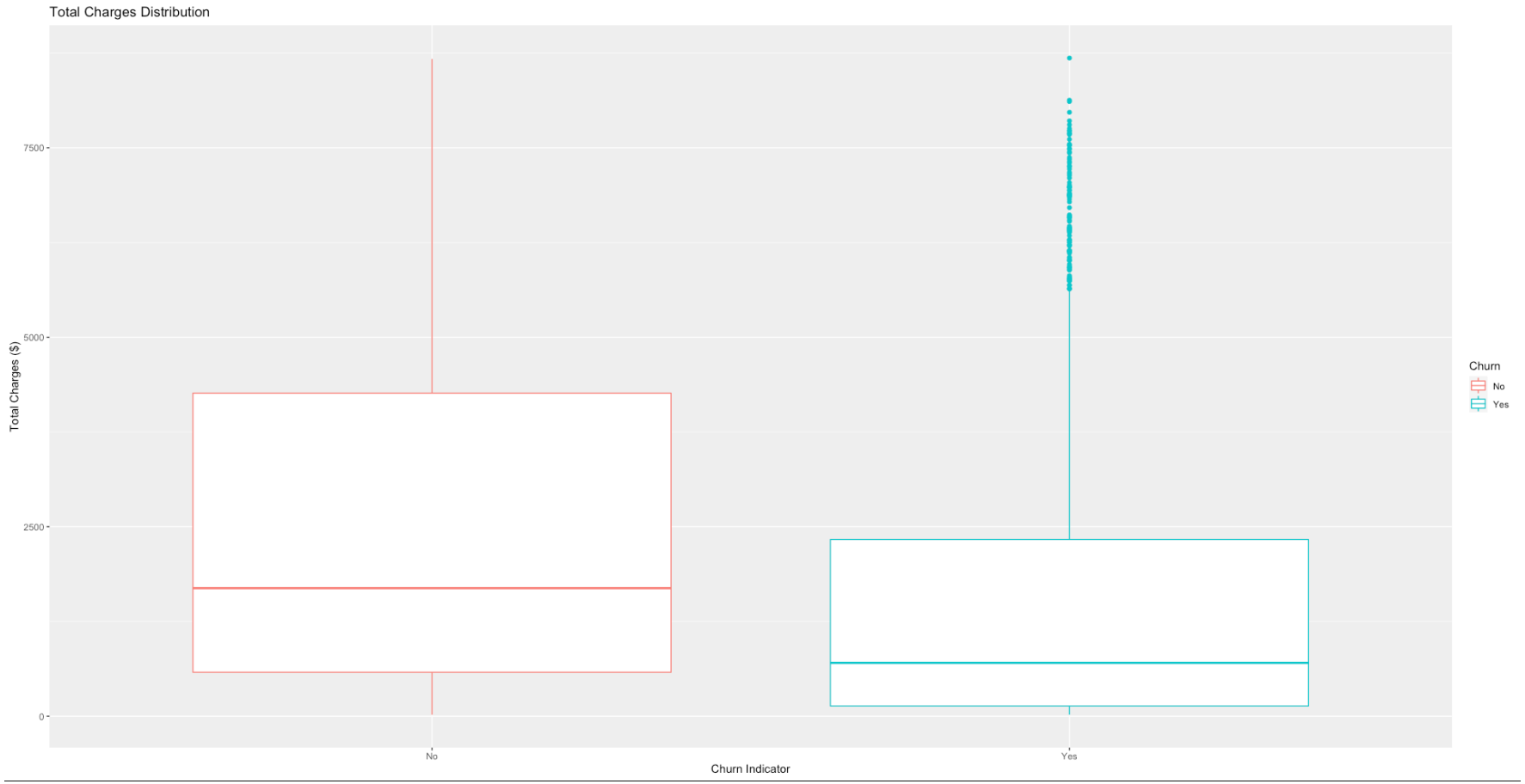
```
> summary(OGdata)
   gender     SeniorCitizen Partner    Dependents     tenure      PhoneService          MultipleLines     InternetService            OnlineSecurity             OnlineBackup
Female:3488   0:5901        No :3641   No :4933   Min.   : 0.00   No : 682     No              :4072   DSL        :2421   No                 :5024   No                 :4614
Male  :3555   1:1142        Yes:3402   Yes:2110   1st Qu.: 9.00   Yes:6361     No phone service:   0   Fiber optic:3096   No internet service:   0   No internet service:   0
                                                  Median :29.00                Yes             :2971   No         :1526   Yes                :2019   Yes                :2429
                                                  Mean   :32.37
                                                  3rd Qu.:55.00
                                                  Max.   :72.00
          DeviceProtection              TechSupport                StreamingTV              StreamingMovies          Contract    PaperlessBilling                  PaymentMethod
No                :4621     No                :4999     No                :4336     No                :4311   Month-to-month:3875   No :2872         Bank transfer (automatic):1544
No internet service:   0     No internet service:   0     No internet service:   0     No internet service:   0   One year      :1473   Yes:4171         Credit card (automatic)  :1522
Yes               :2422     Yes               :2044     Yes               :2707     Yes               :2732   Two year      :1695                    Electronic check         :2365
                                                                                                                                                      Mailed check             :1612


 MonthlyCharges    TotalCharges    Churn
 Min.   : 18.25   Min.   :  18.8   No :5174
 1st Qu.: 35.50   1st Qu.: 402.2   Yes:1869
 Median : 70.35   Median :1400.5
 Mean   : 64.76   Mean   :2283.3
 3rd Qu.: 89.85   3rd Qu.:3786.6
 Max.   :118.75   Max.   :8684.8
```

- 20 variables with 7,043 observations
- 17 categorical
- 3 continuous and discrete

```
> describe(OGdata$tenure)
   vars    n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 7043 32.37 24.56     29   31.43 32.62   0  72    72 0.24    -1.39 0.29
> describe(OGdata$MonthlyCharges)
   vars    n  mean    sd median trimmed   mad   min    max range  skew kurtosis   se
X1    1 7043 64.76 30.09  70.35   64.97 35.66 18.25 118.75 100.5 -0.22    -1.26 0.36
> describe(OGdata$TotalCharges)
   vars    n   mean   sd  median trimmed    mad  min    max range skew kurtosis    se
X1    1 7043 2283.3 2265 1400.55 1970.38 1812.4 18.8 8684.8  8666 0.96    -0.23 26.99
```

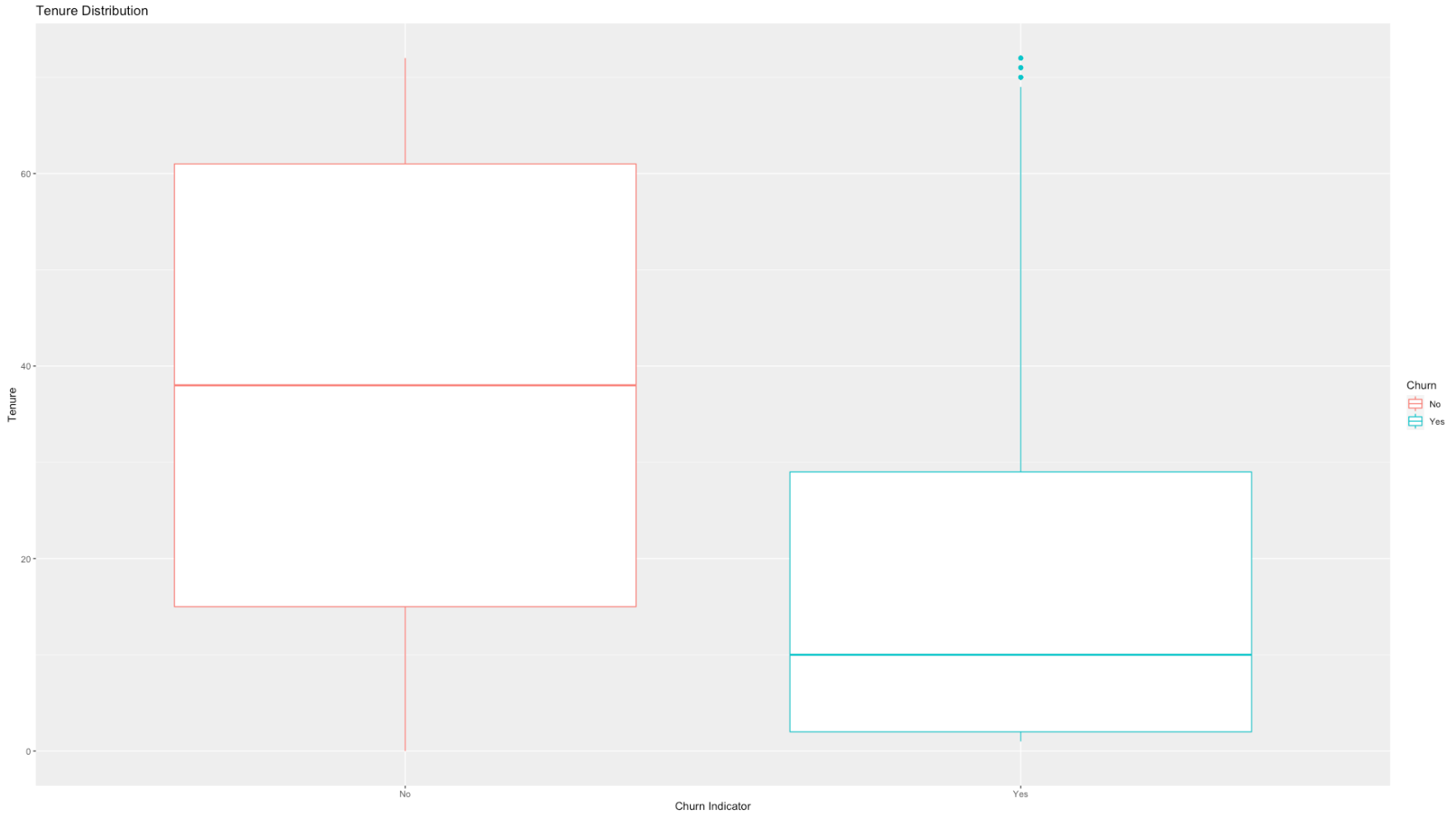
- The data for these variables is disperse
- Normally distributed variables
- Not normally distributed variables

Exploratory Data Analysis



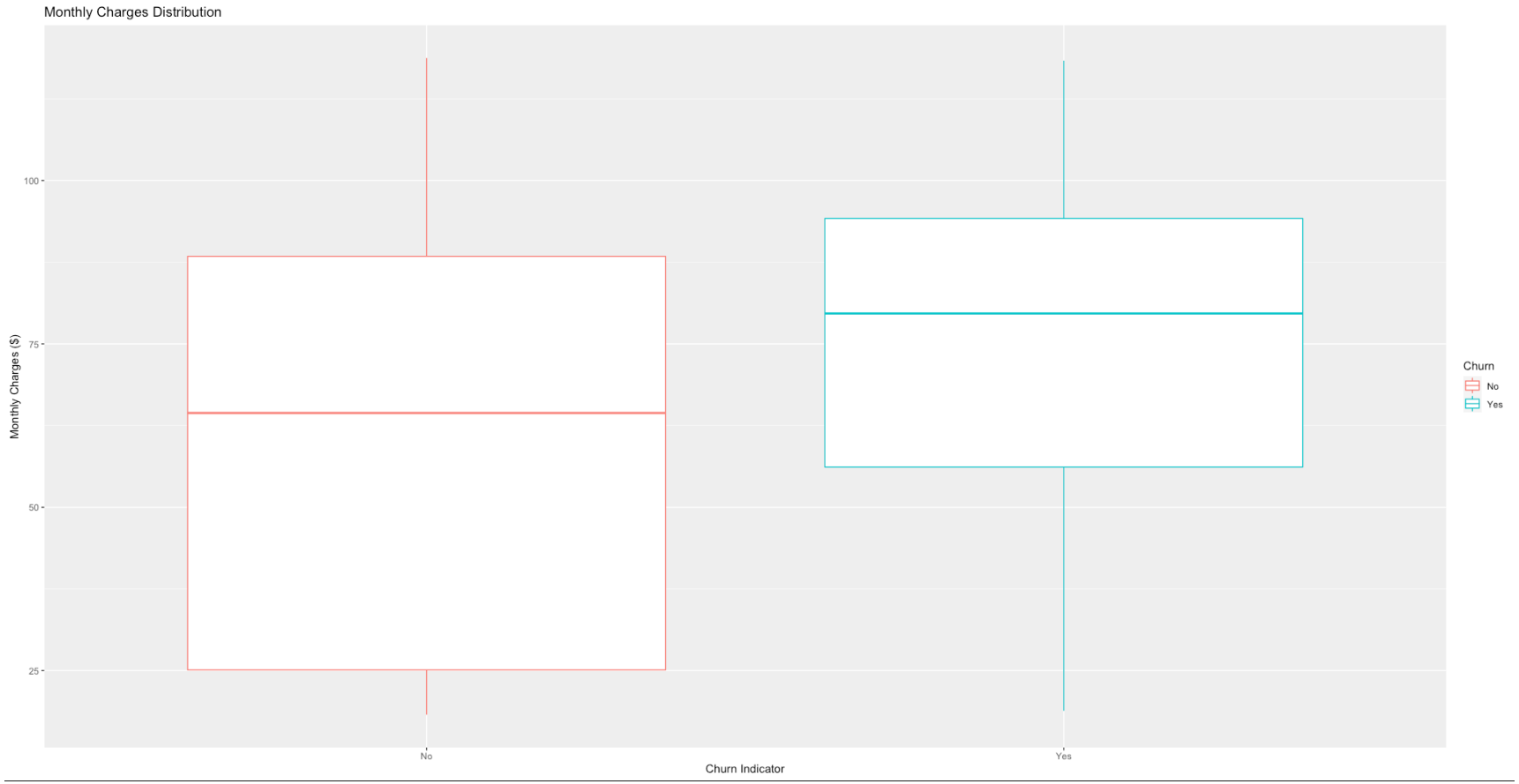
- The distribution of the Total Charges against the Churn is presented
- Outliers are detected
- The difference in the means is present in two groups

Exploratory Data Analysis



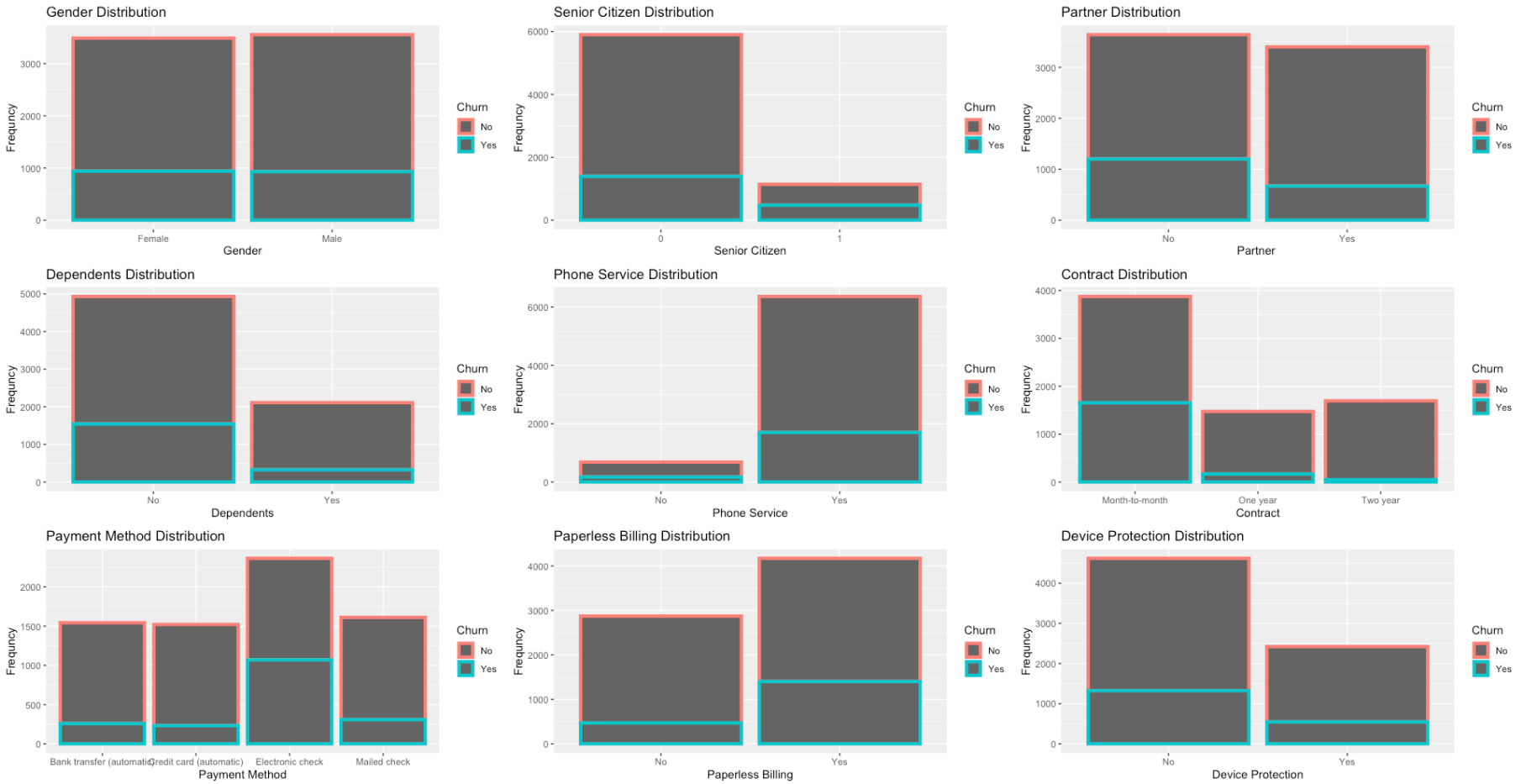
- The distribution of the Tenure against the Churn is presented
- Outliers are detected
- The difference in the means is present in two groups

Exploratory Data Analysis



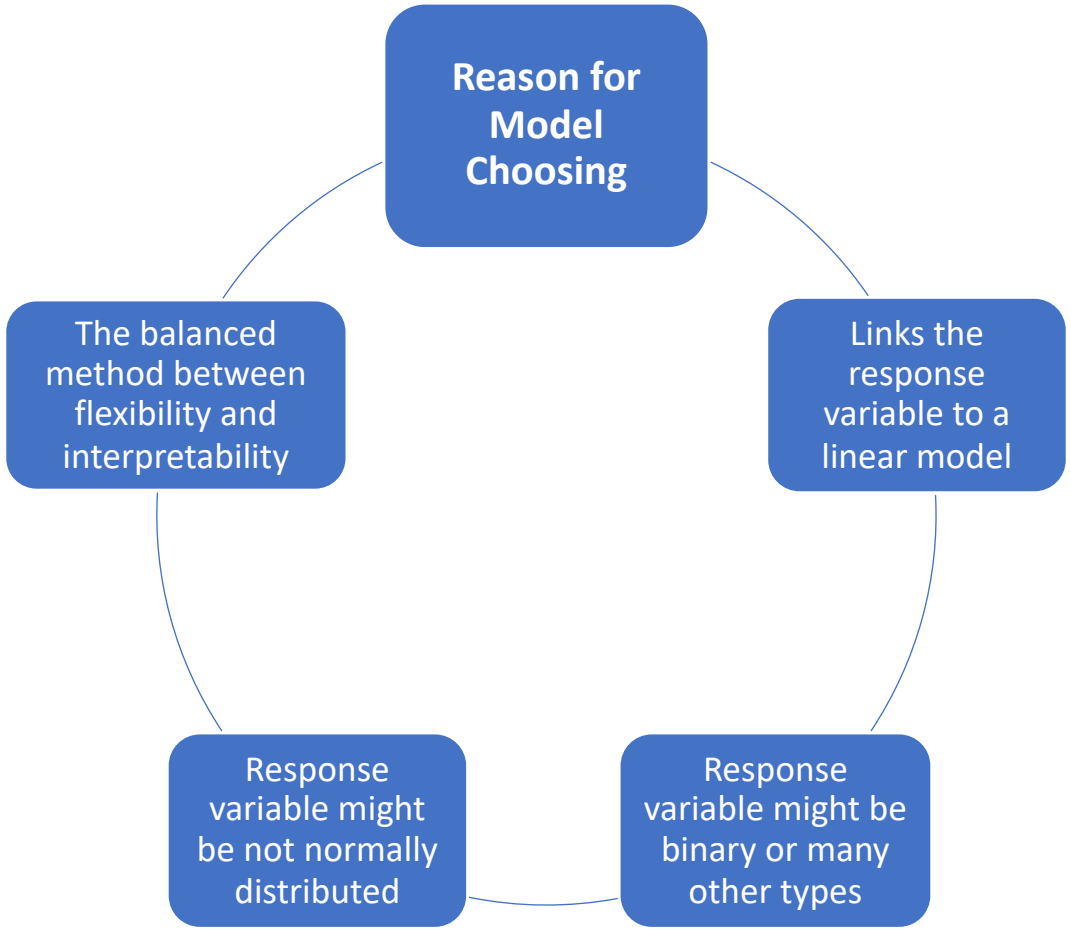
- The distribution of the Monthly Charges against the Churn is presented
- The difference in the means is present in two groups
- The distribution is different from the Total Charges

Exploratory Data Analysis



- A grid of bar plots for categorical variables is presented
- There are certain trends that can be extracted from the figure

GLM: Logistic Regression (LR)



Model Building

```
#- Run logistic regression using binomial family and logit link function
# 1- Fit a logistic regression model with all variables
fit.full <- glm(Churn ~ ., data = train, family=binomial(link="logit"))

#Show fitted model
summary(fit.full)


# 2- Fit a logistic regression model with only significant variables
fit.reduced <- glm(Churn ~ SeniorCitizen + MultipleLines+Contract+
                     PaperlessBilling+PaymentMethod+ten_fact,
                   data = train, family=binomial(link="logit"))

#Show fitted model
summary(fit.reduced)


# 3- Perform stepwise selection with full model
OGdata_Step <- stepAIC(fit.full, direction = 'both')

#fit a model with best predictors defined in stepwise selection
Stepwise_model <- OGdata_Step

#Show fitted model
Stepwise_model
summary(Stepwise_model)
```

LR: Summary of the Models

Full model

```
> summary(fit.full)

Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0414  -0.6719  -0.2902   0.6618   3.0594

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             3.808e-01  9.084e-01   0.419 0.675056
genderMale                             -1.268e-02  7.268e-02  -0.174 0.861512
SeniorCitizen1                          2.632e-01  9.403e-02   2.800 0.005115 **
PartnerYes                             -3.921e-02  8.666e-02  -0.452 0.650920
DependentsYes                          -1.276e-01  1.005e-01  -1.269 0.204332
PhoneServiceYes                         1.422e-01  7.280e-01   0.195 0.845150
MultipleLinesYes                        4.397e-01  1.997e-01   2.202 0.027656 *
InternetServiceFiber optic              1.655e+00  8.959e-01   1.848 0.064650 .
InternetServiceNo                      -1.599e+00  9.055e-01  -1.766 0.077338 .
OnlineSecurityYes                      -2.128e-01  2.009e-01  -1.059 0.289434
OnlineBackupYes                         7.942e-02  1.972e-01   0.403 0.687083
DeviceProtectionYes                     1.909e-01  1.985e-01   0.962 0.336253
TechSupportYes                         -1.325e-01  2.010e-01  -0.659 0.509924
StreamingTVYes                          5.576e-01  3.656e-01   1.525 0.127229
StreamingMoviesYes                      5.821e-01  3.675e-01   1.584 0.113232
ContractOne year                       -7.538e-01  1.200e-01  -6.284 3.29e-10 ***
ContractTwo year                       -1.638e+00  2.015e-01  -8.129 4.34e-16 ***
PaperlessBillingYes                     3.218e-01  8.353e-02   3.853 0.000117 ***
PaymentMethodCredit card (automatic)   -4.622e-02  1.261e-01  -0.366 0.714074
PaymentMethodElectronic check           3.544e-01  1.058e-01   3.349 0.000811 ***
PaymentMethodMailed check              -6.594e-03  1.282e-01  -0.051 0.958993
MonthlyCharges                         -2.843e-02  3.563e-02  -0.798 0.424902
TotalCharges                           -1.104e-04  6.827e-05  -1.618 0.105738
ten_fact1-2years                       -8.184e-01  1.173e-01  -6.980 2.95e-12 ***
ten_fact2-3years                       -1.191e+00  1.669e-01  -7.136 9.58e-13 ***
ten_fact3-4years                       -9.287e-01  2.221e-01  -4.181 2.90e-05 ***
ten_fact4-5years                       -1.138e+00  2.876e-01  -3.955 7.65e-05 ***
ten_fact5-6years                       -1.272e+00  3.809e-01  -3.339 0.000841 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6525.6  on 5633  degrees of freedom
Residual deviance: 4681.1  on 5606  degrees of freedom
AIC: 4737.1
```

Reduced model

```
> summary(fit.reduced)

Call:
glm(formula = Churn ~ SeniorCitizen + MultipleLines + Contract +
    PaperlessBilling + PaymentMethod + ten_fact, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8526  -0.7669  -0.3260   0.7845   3.0163

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -0.74181    0.11309  -6.560 5.39e-11 ***
SeniorCitizen1                          0.49711    0.08924   5.571 2.54e-08 ***
MultipleLinesYes                        0.54480    0.07942   6.859 6.92e-12 ***
ContractOne year                       -0.99512    0.11319  -8.791  < 2e-16 ***
ContractTwo year                       -2.14458    0.19290 -11.118  < 2e-16 ***
PaperlessBillingYes                     0.61432    0.07850   7.826 5.03e-15 ***
PaymentMethodCredit card (automatic)   -0.06081    0.12225  -0.497  0.61887
PaymentMethodElectronic check           0.60351    0.10143   5.950 2.68e-09 ***
PaymentMethodMailed check              -0.33160    0.12085  -2.744  0.00607 **
ten_fact1-2years                       -0.82643    0.10225  -8.082 6.36e-16 ***
ten_fact2-3years                       -1.24092    0.12241 -10.138  < 2e-16 ***
ten_fact3-4years                       -1.06246    0.13372  -7.945 1.94e-15 ***
ten_fact4-5years                       -1.32037    0.14715  -8.973  < 2e-16 ***
ten_fact5-6years                       -1.56490    0.17421  -8.983  < 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: 6525.6  on 5633  degrees of freedom
Residual deviance: 4957.6  on 5620  degrees of freedom
AIC: 4985.6
```

Stepwise model

```
> summary(Stepwise_model)

Call:
glm(formula = Churn ~ SeniorCitizen + Dependents + MultipleLines +
    InternetService + OnlineSecurity + DeviceProtection + TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod + MonthlyCharges + TotalCharges + ten_fact,
    family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0411  -0.6697  -0.2920   0.6600   3.0670

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             1.622e-01  3.011e-01   0.539 0.590032
SeniorCitizen1                          2.595e-01  9.352e-02   2.774 0.005532 **
DependentsYes                          -1.456e-01  9.185e-02  -1.585 0.112929
MultipleLinesYes                        3.961e-01  9.994e-02   3.964 7.38e-05 ***
InternetServiceFiber optic              1.445e+00  2.213e-01   6.529 6.64e-11 ***
InternetServiceNo                      -1.414e+00  1.984e-01  -7.125 1.04e-12 ***
OnlineSecurityYes                      -2.537e-01  1.018e-01  -2.492 0.012706 *
DeviceProtectionYes                     1.503e-01  9.289e-02   1.618 0.105595
TechSupportYes                         -1.726e-01  1.010e-01  -1.708 0.087671 .
StreamingTVYes                          4.766e-01  1.087e-01   4.383 1.17e-05 ***
StreamingMoviesYes                      5.004e-01  1.071e-01   4.672 2.98e-06 ***
ContractOne year                       -7.531e-01  1.199e-01  -6.279 3.40e-10 ***
ContractTwo year                       -1.638e+00  2.016e-01  -8.125 4.46e-16 ***
PaperlessBillingYes                     3.242e-01  8.344e-02   3.885 0.000102 ***
PaymentMethodCredit card (automatic)   -4.401e-02  1.260e-01  -0.349 0.726933
PaymentMethodElectronic check           3.550e-01  1.058e-01   3.355 0.000793 ***
PaymentMethodMailed check              -5.166e-03  1.280e-01  -0.040 0.967815
MonthlyCharges                         -2.038e-02  6.633e-03  -3.072 0.002128 **
TotalCharges                           -1.081e-04  6.776e-05  -1.596 0.110541
ten_fact1-2years                       -8.210e-01  1.169e-01  -7.021 2.21e-12 ***
ten_fact2-3years                       -1.194e+00  1.665e-01  -7.172 7.42e-13 ***
ten_fact3-4years                       -9.331e-01  2.217e-01  -4.208 2.58e-05 ***
ten_fact4-5years                       -1.143e+00  2.871e-01  -3.981 6.85e-05 ***
ten_fact5-6years                       -1.278e+00  3.803e-01  -3.361 0.000776 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6525.6  on 5633  degrees of freedom
Residual deviance: 4681.7  on 5610  degrees of freedom
AIC: 4729.7
```

- The full model with all 19 predictors has a lower AIC 4737.1 than the reduced model with 6 significant predictors - AIC 4985.6
- While the stepwise model with 15 predictors has the lowest AIC 4729.7 than the previous two models

LR: Fitting on the Training Set

	Actually Negative (0)	Actually Positive (1)
Predicted Negative (0)	TN	FN
Predicted Positive (0)	FP	TP

Full model

```
> #Show the model accuracy with advanced
> #confusion matrix of the train set
> confusionMatrix(predicted.classes.min_full,
+                 train$Churn, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3745  726
       Yes  391  772

               Accuracy : 0.8017
                 95% CI : (0.7911, 0.8121)
    No Information Rate : 0.7341
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4531

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5154
            Specificity : 0.9055
         Pos Pred Value : 0.6638
         Neg Pred Value : 0.8376
             Prevalence : 0.2659
         Detection Rate : 0.1370
   Detection Prevalence : 0.2064
      Balanced Accuracy : 0.7104

       'Positive' Class : Yes
```

Reduced model

```
> #Show the model accuracy with advanced
> #confusion matrix of the train set
> confusionMatrix(predicted.classes.min_reduced,
+                 train$Churn, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3762  851
       Yes  374  647

               Accuracy : 0.7826
                 95% CI : (0.7716, 0.7933)
    No Information Rate : 0.7341
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3801

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4319
            Specificity : 0.9096
         Pos Pred Value : 0.6337
         Neg Pred Value : 0.8155
             Prevalence : 0.2659
         Detection Rate : 0.1148
   Detection Prevalence : 0.1812
      Balanced Accuracy : 0.6707

       'Positive' Class : Yes
```

Stepwise model

```
> #Show the model accuracy with advanced
> #confusion matrix of the train set
> confusionMatrix(predicted.classes.min_step,
+                 train$Churn, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3743  726
       Yes  393  772

               Accuracy : 0.8014
                 95% CI : (0.7907, 0.8117)
    No Information Rate : 0.7341
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4524

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5154
            Specificity : 0.9050
         Pos Pred Value : 0.6627
         Neg Pred Value : 0.8375
             Prevalence : 0.2659
         Detection Rate : 0.1370
   Detection Prevalence : 0.2068
      Balanced Accuracy : 0.7102

       'Positive' Class : Yes
```

- The reduced model mispredicted 851 customers attrition, whereas full and stepwise models mispredicted 726 customers attrition
- The full and stepwise models have higher accuracy and lower misclassification error rate for almost 2% than the reduced model
- All models are at 26.6% prevalence of the positive class which is 'Yes' in the dataset. Thus, we have less churned customers and a higher of those who stayed

- Type II error
- Type I error

LR: Fitting on the Test Set

	Actually Negative (0)	Actually Positive (1)
Predicted Negative (0)	TN	FN
Predicted Positive (0)	FP	TP

Full model

```
> #Show the model accuracy with advanced
> #confusion matrix of the test set
> confusionMatrix(predicted.classes.min_test_full,
+                 test$Churn, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  947 188
       Yes  91 183

               Accuracy : 0.802
                 95% CI : (0.7802, 0.8225)
    No Information Rate : 0.7367
    P-Value [Acc > NIR] : 6.053e-09

                  Kappa : 0.4428

 Mcnemar's Test P-Value : 9.064e-09

            Sensitivity : 0.4933
            Specificity : 0.9123
         Pos Pred Value : 0.6679
         Neg Pred Value : 0.8344
             Prevalence : 0.2633
         Detection Rate : 0.1299
   Detection Prevalence : 0.1945
      Balanced Accuracy : 0.7028

       'Positive' Class : Yes
```

Reduced model

```
> #Show the model accuracy with advanced
> #confusion matrix of the test set
> confusionMatrix(predicted.classes.min_test_reduced,
+                 test$Churn, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  940 212
       Yes  98 159

               Accuracy : 0.78
                 95% CI : (0.7574, 0.8014)
    No Information Rate : 0.7367
    P-Value [Acc > NIR] : 9.782e-05

                  Kappa : 0.3708

 Mcnemar's Test P-Value : 1.381e-10

            Sensitivity : 0.4286
            Specificity : 0.9056
         Pos Pred Value : 0.6187
         Neg Pred Value : 0.8160
             Prevalence : 0.2633
         Detection Rate : 0.1128
   Detection Prevalence : 0.1824
      Balanced Accuracy : 0.6671

       'Positive' Class : Yes
```

Stepwise model

```
> #Show the model accuracy with advanced
> #confusion matrix of the test set
> confusionMatrix(predicted.classes.min_test_step,
+                 test$Churn, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  950 189
       Yes  88 182

               Accuracy : 0.8034
                 95% CI : (0.7817, 0.8239)
    No Information Rate : 0.7367
    P-Value [Acc > NIR] : 2.817e-09

                  Kappa : 0.4447

 Mcnemar's Test P-Value : 1.873e-09

            Sensitivity : 0.4906
            Specificity : 0.9152
         Pos Pred Value : 0.6741
         Neg Pred Value : 0.8341
             Prevalence : 0.2633
         Detection Rate : 0.1292
   Detection Prevalence : 0.1916
      Balanced Accuracy : 0.7029

       'Positive' Class : Yes
```

- As in the reduced model in the train set, the full and stepwise models in the test set now also have higher FN values than TP
- All models have almost the same accuracy level as in the train set. But the stepwise model has higher accuracy than the full model for 0.14%.
- All models are at 26.3% prevalence of the positive class as in the train set

- Type II error
- Type I error

LR: Performance Evaluation

	Accuracy	Sensitivity	Specificity	Precision	AUC
Full Model	**0.802**	**0.4933**	**0.9123**	**0.6678**	**0.845**
Reduced Model	0.78	0.4286	0.9056	0.6186	0.8209
Stepwise Model	**0.8034**	**0.4906**	**0.9152**	**0.6740**	**0.8454**

- The Full Model with 19 variables performs better on the model's Sensitivity
- The Reduced Model with 6 significant predictors is the most underperformed model
- Whereas the Stepwise Model with 15 variables performs better on the model's Accuracy, Specificity, Precision, and AUC
- The Stepwise model having 3 predictors fewer than the Full Model seems to perform better than the other models

LASSO Regression Model

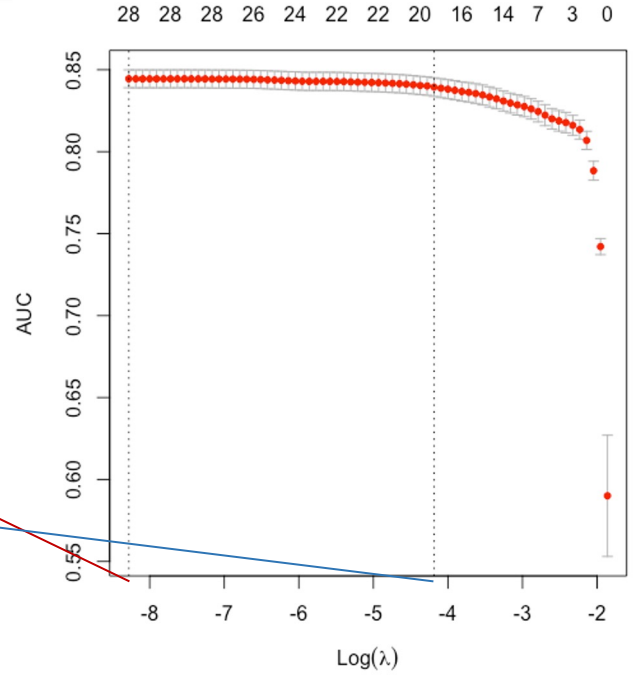
Reason for Model Choosing

- Prevent the over-fitting problem
- Penalty is the **absolute value** of coefficients -> Shrinks coefficients to **zero** -> **dimension reduction** and **feature selection**

Model Building

```
# Use cv.glmnet function to cross validate to find the best lambda values
# The cross validation method I chose is K-fold and here, K is 10.
CV.L1 <- cv.glmnet(trainX, trainY, alpha = 1, nfolds = 10, family="binomial", type.measure='auc')
```

```
> # lambda.min
> CV.L1$lambda.min
[1] 0.00143047
> # log of lambda.min
> log(CV.L1$lambda.min)
[1] -6.549752
> # one-standard-error lambda
> CV.L1$lambda.1se
[1] 0.01464131
> # log of one-standard-error lambda
> log(CV.L1$lambda.1se)
[1] -4.223908
```



LASSO Regression Model

Fitting on the **training** set:

minimum λ

```
> confusionMatrix(predicted.train.min, trainY, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3762  731
       Yes  374  767

               Accuracy : 0.8039
                 95% CI : (0.7933, 0.8142)
    No Information Rate : 0.7341
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4563

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5120
            Specificity : 0.9096
         Pos Pred Value : 0.6722
         Neg Pred Value : 0.8373
             Prevalence : 0.2659
         Detection Rate : 0.1361
   Detection Prevalence : 0.2025
      Balanced Accuracy : 0.7108

       'Positive' Class : Yes
```

Type II error

Type I error

1-standard-error λ

```
> confusionMatrix(predicted.train.1se, trainY, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3817  827
       Yes  319  671

               Accuracy : 0.7966
                 95% CI : (0.7858, 0.807)
    No Information Rate : 0.7341
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4158

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4479
            Specificity : 0.9229
         Pos Pred Value : 0.6778
         Neg Pred Value : 0.8219
             Prevalence : 0.2659
         Detection Rate : 0.1191
   Detection Prevalence : 0.1757
      Balanced Accuracy : 0.6854

       'Positive' Class : Yes
```

LASSO Regression Model

Fitting on the **test** set:

minimum λ

```
> confusionMatrix(predicted.test.min, testY, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  951 192
       Yes  87 179

               Accuracy : 0.802
                 95% CI : (0.7802, 0.8225)
    No Information Rate : 0.7367
    P-Value [Acc > NIR] : 6.053e-09

                  Kappa : 0.4385

 Mcnemar's Test P-Value : 4.775e-10

            Sensitivity : 0.4825
            Specificity : 0.9162
         Pos Pred Value : 0.6729
         Neg Pred Value : 0.8320
             Prevalence : 0.2633
         Detection Rate : 0.1270
   Detection Prevalence : 0.1888
      Balanced Accuracy : 0.6993

       'Positive' Class : Yes
```

Type II error

Type I error

1-standard-error λ

```
> confusionMatrix(predicted.test.1se, testY, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  963 208
       Yes  75 163

               Accuracy : 0.7991
                 95% CI : (0.7773, 0.8198)
    No Information Rate : 0.7367
    P-Value [Acc > NIR] : 2.652e-08

                  Kappa : 0.4149

 Mcnemar's Test P-Value : 4.275e-15

            Sensitivity : 0.4394
            Specificity : 0.9277
         Pos Pred Value : 0.6849
         Neg Pred Value : 0.8224
             Prevalence : 0.2633
         Detection Rate : 0.1157
   Detection Prevalence : 0.1689
      Balanced Accuracy : 0.6835

       'Positive' Class : Yes
```

LASSO Regression Model

	accuracy	sensitivity	specificity	precision	AUC
min λ on train	**0.8039**	**0.5120**	0.9096	0.6722	**0.8465**
min λ on test	**0.802**	**0.4825**	0.9162	0.6729	**0.8453**
1SE λ on train	0.7966	0.4479	**0.9229**	**0.6778**	0.8400
1SE λ on test	0.7991	0.4394	**0.9277**	**0.6849**	0.8400

Performance Evaluation:

- The LASSO regression model with **minimum λ** performs better on model's **accuracy** and **sensitivity**;
- On the other hand, the LASSO regression model with 1-standard-error performs better on specificity and precision.

Conclusion

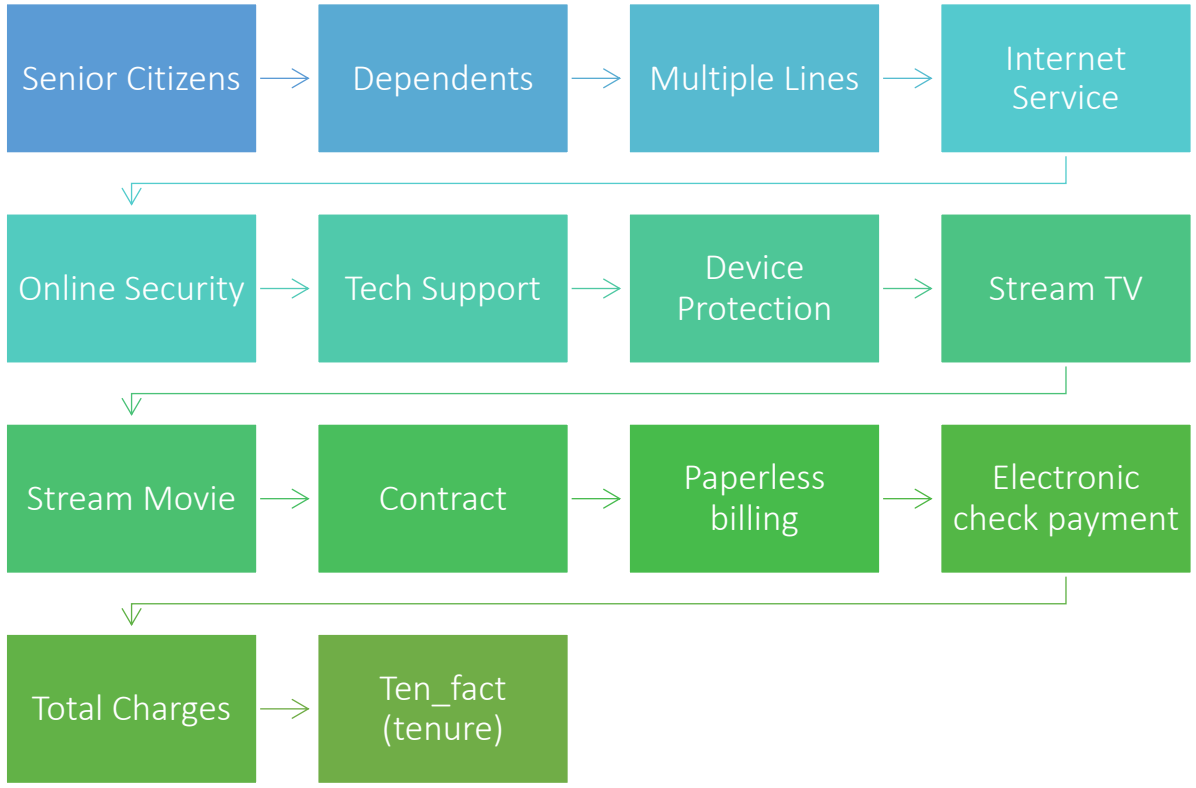
- Estimation of Logistic and LASSO Regression models and each classifiers, we can conclude that Stepwise model with 15 predictors and the lowest AIC showed better performance in Logistic Regression.

- In this case while model with minimum lambda on training data achieved the highest accuracy and sensitivity, and the highest specificity and precision with 1 standard error lambda on testing data. AUC in Stepwise model and minimum lambda(on train set) are so similar, within 84.54% and 84.65%, respectively

- Although that sensitivity is fewer higher in stepwise model rather than min lambda model on train data, rest of metrics are higher on min lambda model on train data set

Recommendations

Company should focus on these variables to control Churn variables.
Because they are mostly significant variables as the result of all models.



Reference:

- BlastChar. (2018, February 23). *Telco customer churn*. Kaggle. Retrieved January 26, 2022, from https://www.kaggle.com/blastchar/telco-customer-churn