6210 Telco Project

Kelley Yao, Lisa Witherspoon, Gabriela Perez, Danica Chin

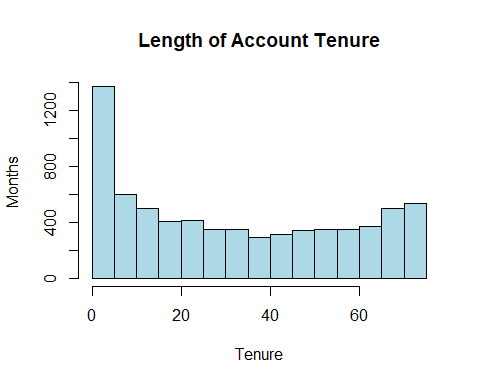
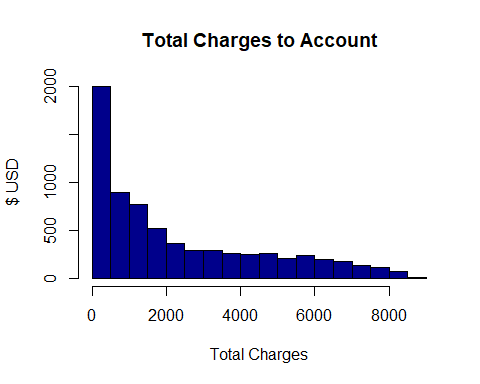
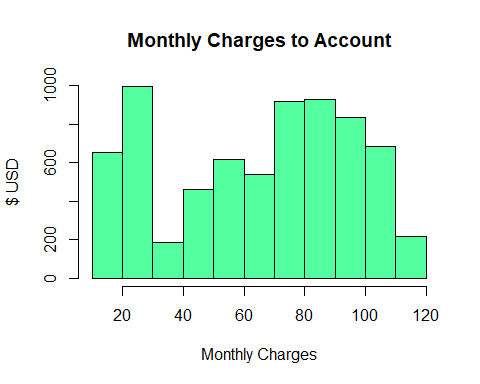
05/15/2022

## Executive Summary

Telecommunications, like many other industries, is a saturated and competitive market in which customer retention can be just as vital to revenue growth as new customer procurement. The telecommunication industry in the United States earned almost $133 billion in revenue in 2020, as the United States has one of the world’s largest smartphone-using populations, behind only countries such as China and India. Telecommunication companies are therefore incentivized to improve upon this revenue stream and build on their existing customer base, without losing the customers they already have. The concept of “churning” results in customers jumping ship to other companies and “is used as an indicator of the health and loyalty of a company’s subscriber base”, as consistent customers make the bulk of each company’s revenue stream. Predictive analytics into customer data is a major asset for companies looking to actively prevent and therefore reduce customer turnover rather than taking the alternative reactive approach to customer loss.

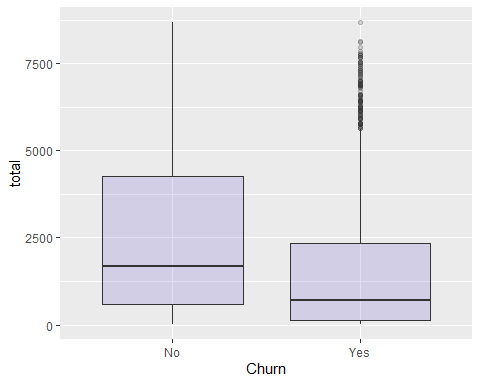
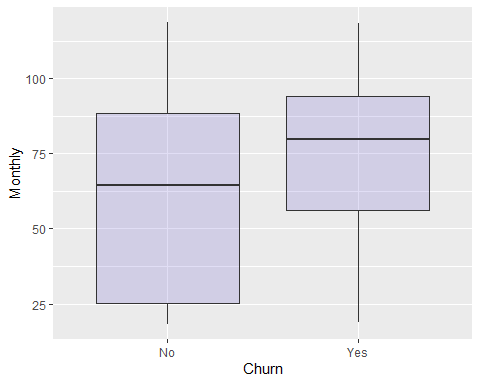
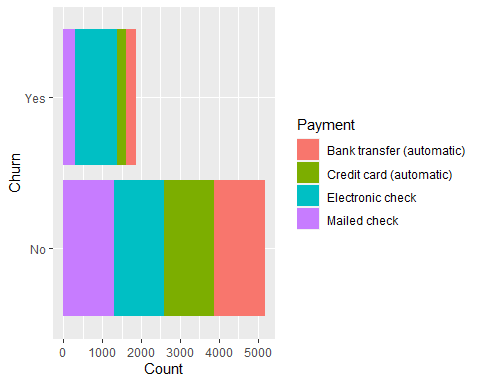
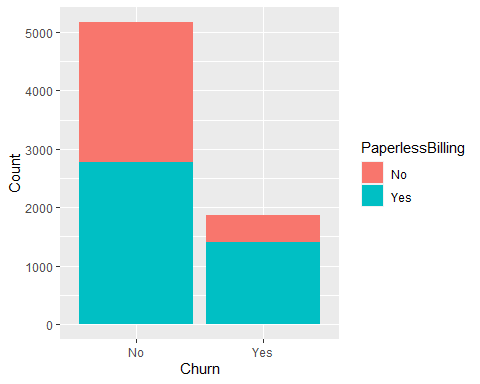
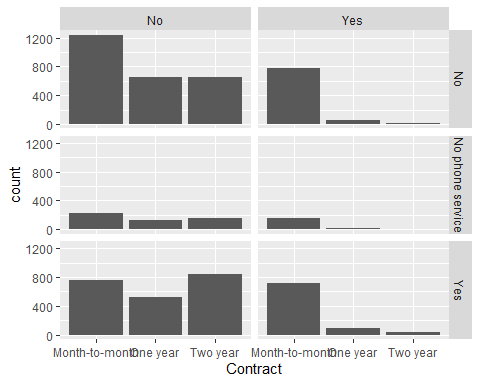
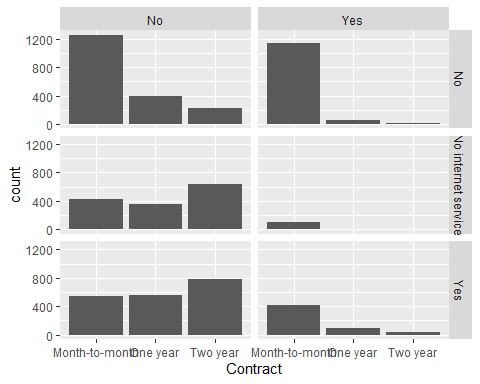
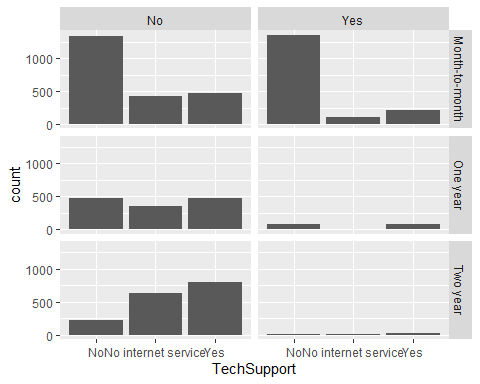
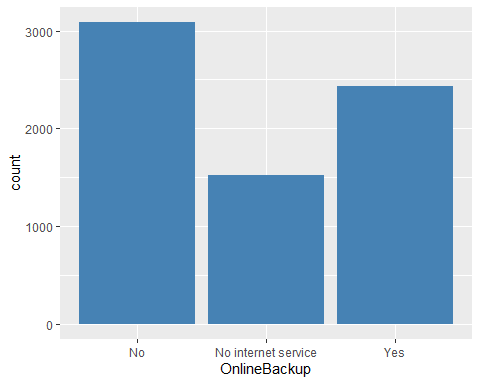
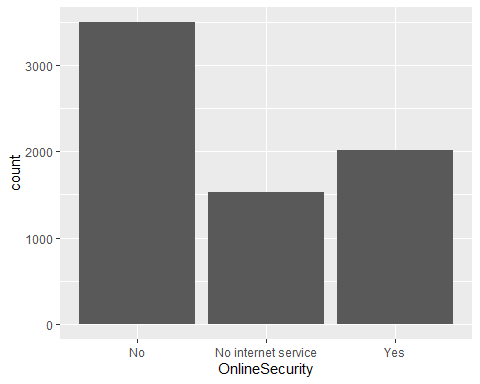
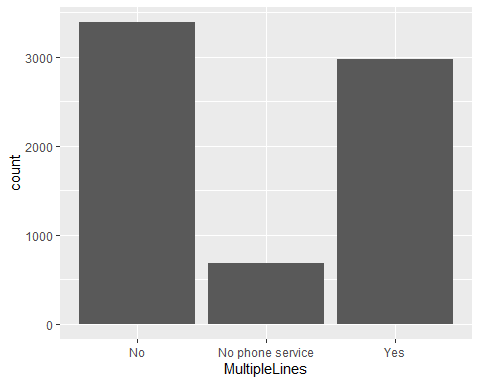
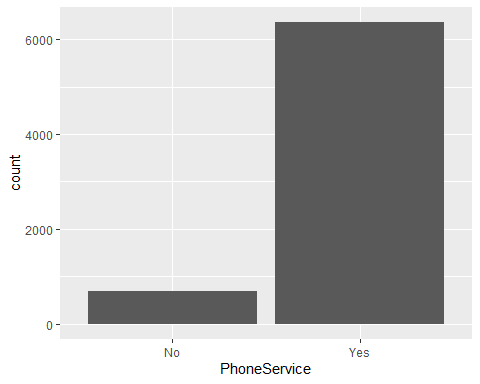
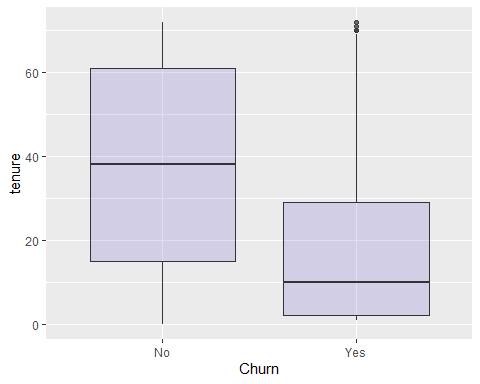
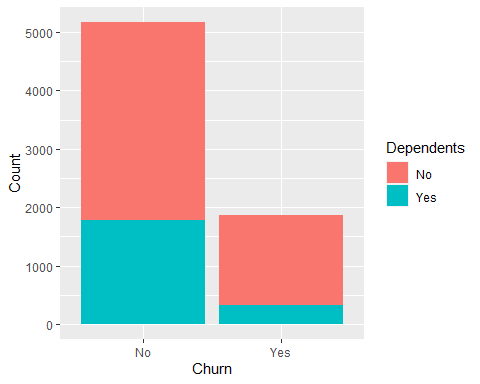
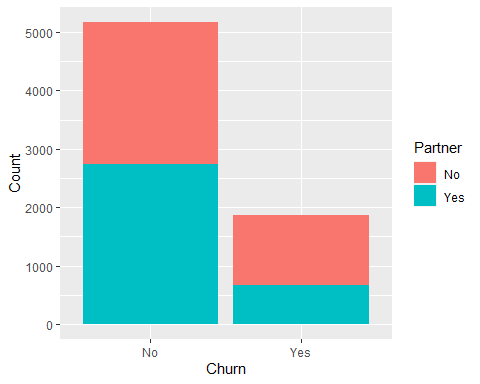
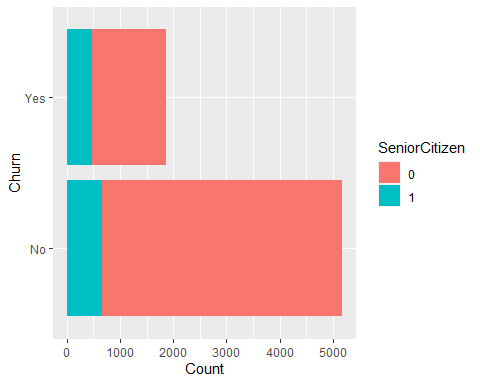
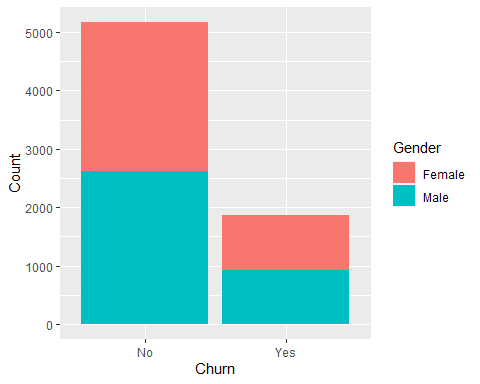
For this case study, we look at a telecommunications company, Telco, for customer data with aims to identify possible factors correlated to customer churn. These findings as well as constructing and selecting a predictive classification model forecasting customers likely to churn can contribute to higher customer retention rate through preventative action.

mean min first\_qt median third\_qt max  
tenure 32.37115 0.00 9.00 29.000 55.000 72.00  
MonthlyCharges 64.76169 18.25 35.50 70.350 89.850 118.75  
TotalCharges 2283.30044 18.80 401.45 1397.475 3794.738 8684.80



## Data Visualization

Plot variables in relation to dependent variable Churn

 ## **Predictive Modeling**

This case requires a predictive classification model. Our group chose to pre-emptively construct two different models to test- a classification tree and a logistic regression model- and chose one model to implement on the data set based on performance evaluation metrics. Data pre-processing included removing empty records as well as combining variable outcomes to avoid multicollinearity during regression analysis (see Data pre-processing). Using 10-folds cross validation to partition the data set, a logistic model was constructed and tuned using step-wise backwards regression and a classification tree was created using the algorithm from rpart library in R.

## Logistic Regression Model

The first model we attempted was a logistic regression model, as it seemed fitting. Our variable of interest was the churn variable, with a distinct “Yes” or “No” binary outcome, which we transform into a factor variable with levels “1, 0”.

We first create the entire logistic regression model with all variables intact, then use backwards stepwise regression to eliminate variables that are not significant in predicting “churn”. Stepwise regression is a technical guide that assists in removing variables one by one, where the smaller the AIC (Aikake information criterion), the better the model is.

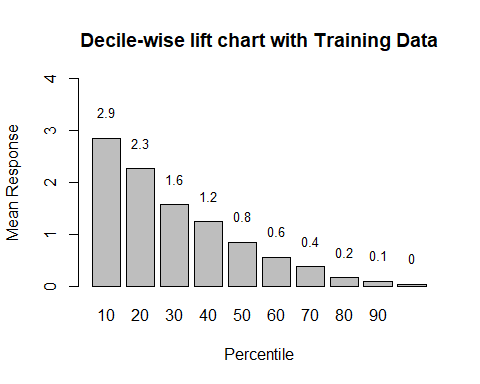
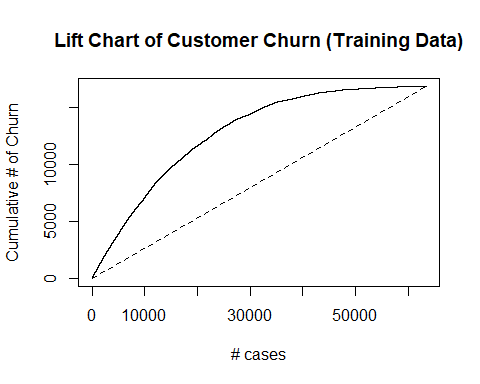
Call:  
glm(formula = Churn ~ ., family = "binomial", data = telco.logit,   
 na.action = na.omit)  
   
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-1.9180 -0.6791 -0.2855 0.7282 3.4300   
  
 Coefficients: (7 not defined because of singularities)  
 Estimate Std. Error z value  
 (Intercept) 1.16528747 0.81513548 1.430  
 genderMale -0.02183273 0.06480439 -0.337  
 SeniorCitizen 0.21677504 0.08453049 2.564  
 PartnerYes -0.00038400 0.07782976 -0.005  
 DependentsYes -0.14848782 0.08973115 -1.655  
 tenure -0.06058758 0.00623568 -9.716  
 PhoneServiceYes 0.17146779 0.64869215 0.264  
 MultipleLinesNo phone service NA NA NA  
 MultipleLinesYes 0.44839544 0.17725585 2.530  
 InternetServiceFiber optic 1.74747491 0.79807951 2.190  
 InternetServiceNo -1.78629472 0.80726809 -2.213  
 OnlineSecurityNo internet service NA NA NA  
 OnlineSecurityYes -0.20542004 0.17868789 -1.150  
 OnlineBackupNo internet service NA NA NA  
 OnlineBackupYes 0.02604185 0.17540119 0.148  
 DeviceProtectionNo internet service NA NA NA  
 DeviceProtectionYes 0.14737500 0.17637397 0.836  
 TechSupportNo internet service NA NA NA  
 TechSupportYes -0.18049681 0.18060228 -0.999  
 StreamingTVNo internet service NA NA NA  
 StreamingTVYes 0.59050741 0.32630948 1.810  
 StreamingMoviesNo internet service NA NA NA  
 StreamingMoviesYes 0.59929571 0.32668370 1.834  
 ContractOne year -0.66079529 0.10758532 -6.142  
 ContractTwo year -1.35710624 0.17644517 -7.691  
 PaperlessBillingYes 0.34235364 0.07449538 4.596  
 PaymentMethodCredit card (automatic) -0.08779182 0.11407927 -0.770  
 PaymentMethodElectronic check 0.30446727 0.09449652 3.222  
 PaymentMethodMailed check -0.05758719 0.11491136 -0.501  
 MonthlyCharges -0.04034353 0.03175504 -1.270  
 TotalCharges 0.00032894 0.00007063 4.657  
 Pr(>|z|)   
 (Intercept) 0.15284   
 genderMale 0.73619   
 SeniorCitizen 0.01033 \*   
 PartnerYes 0.99606   
 DependentsYes 0.09796 .   
 tenure < 0.0000000000000002 \*\*\*  
 PhoneServiceYes 0.79153   
 MultipleLinesNo phone service NA   
 MultipleLinesYes 0.01142 \*   
 InternetServiceFiber optic 0.02855 \*   
 InternetServiceNo 0.02691 \*   
 OnlineSecurityNo internet service NA   
 OnlineSecurityYes 0.25031   
 OnlineBackupNo internet service NA   
 OnlineBackupYes 0.88197   
 DeviceProtectionNo internet service NA   
 DeviceProtectionYes 0.40339   
 TechSupportNo internet service NA   
 TechSupportYes 0.31759   
 StreamingTVNo internet service NA   
 StreamingTVYes 0.07035 .   
 StreamingMoviesNo internet service NA   
 StreamingMoviesYes 0.06658 .   
 ContractOne year 0.0000000008145911 \*\*\*  
 ContractTwo year 0.0000000000000146 \*\*\*  
 PaperlessBillingYes 0.0000043143131817 \*\*\*  
 PaymentMethodCredit card (automatic) 0.44156   
 PaymentMethodElectronic check 0.00127 \*\*   
 PaymentMethodMailed check 0.61627   
 MonthlyCharges 0.20392   
 TotalCharges 0.0000032026251598 \*\*\*  
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 (Dispersion parameter for binomial family taken to be 1)  
   
 Null deviance: 8143.4 on 7031 degrees of freedom  
 Residual deviance: 5826.3 on 7008 degrees of freedom  
 (11 observations deleted due to missingness)  
 AIC: 5874.3  
   
 Number of Fisher Scoring iterations: 6

Stepwise Model Path   
 Analysis of Deviance Table  
   
 Initial Model:  
 Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +   
 PhoneService + MultipleLines + InternetService + OnlineSecurity +   
 OnlineBackup + DeviceProtection + TechSupport + StreamingTV +   
 StreamingMovies + Contract + PaperlessBilling + PaymentMethod +   
 MonthlyCharges + TotalCharges  
   
 Final Model:  
 Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +   
 InternetService + OnlineSecurity + TechSupport + StreamingTV +   
 StreamingMovies + Contract + PaperlessBilling + PaymentMethod +   
 MonthlyCharges + TotalCharges  
   
   
 Step Df Deviance Resid. Df Resid. Dev AIC  
 1 7008 5826.272 5874.272  
 2 - PhoneService 0 0.0000000000 7008 5826.272 5874.272  
 3 - Partner 1 0.0000243419 7009 5826.272 5872.272  
 4 - OnlineBackup 1 0.0220270678 7010 5826.294 5870.294  
 5 - gender 1 0.1163653710 7011 5826.411 5868.411  
 6 - DeviceProtection 1 1.4307215836 7012 5827.842 5867.842

## Cross validation procedure

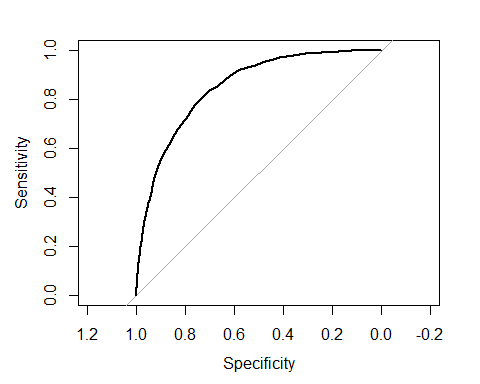
In order to train the dataset, we randomized the dataset, then split it into 10 equal sections to cross validate. The for loop in the below code takes the finalized model after backwards stepwise regression is performed and trains the model utilizing the sections we randomly split. Prediction propensity is then calculated and utilized to check the efficacy by creating a confusion matrix, gain chart and ROC/AUC.

Confusion Matrix and Statistics  
   
 Reference  
 Prediction 1 0  
 1 9343 4780  
 0 7478 41687  
   
 Accuracy : 0.8063   
 95% CI : (0.8032, 0.8094)   
 No Information Rate : 0.7342   
 P-Value [Acc > NIR] : < 0.00000000000000022  
   
 Kappa : 0.477   
   
 Mcnemar's Test P-Value : < 0.00000000000000022  
   
 Sensitivity : 0.5554   
 Specificity : 0.8971   
 Pos Pred Value : 0.6615   
 Neg Pred Value : 0.8479   
 Prevalence : 0.2658   
 Detection Rate : 0.1476   
 Detection Prevalence : 0.2232   
 Balanced Accuracy : 0.7263   
   
 'Positive' Class : 1



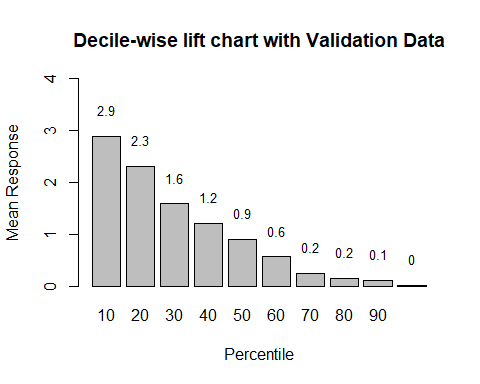
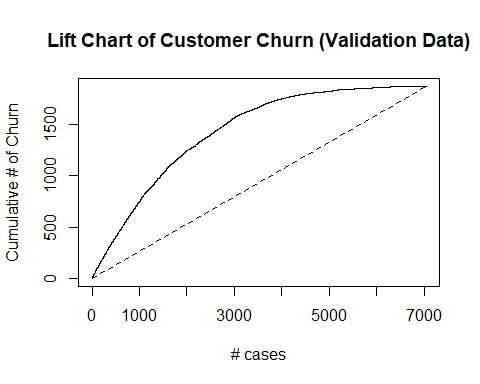
Setting levels: control = 0, case = 1

Setting direction: controls < cases

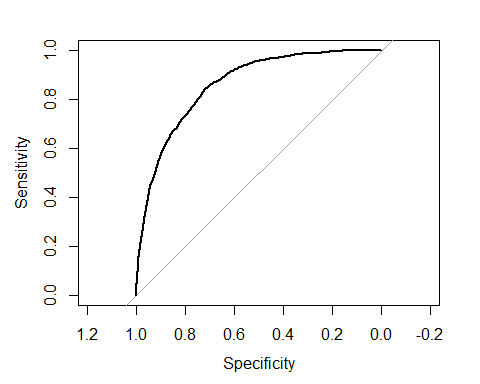


Area under the curve: 0.8482

Confusion Matrix and Statistics  
   
 Reference  
 Prediction 1 0  
 1 1078 521  
 0 791 4642  
   
 Accuracy : 0.8134   
 95% CI : (0.8041, 0.8225)   
 No Information Rate : 0.7342   
 P-Value [Acc > NIR] : < 0.00000000000000022  
   
 Kappa : 0.4989   
   
 Mcnemar's Test P-Value : 0.0000000000001115   
   
 Sensitivity : 0.5768   
 Specificity : 0.8991   
 Pos Pred Value : 0.6742   
 Neg Pred Value : 0.8544   
 Prevalence : 0.2658   
 Detection Rate : 0.1533   
 Detection Prevalence : 0.2274   
 Balanced Accuracy : 0.7379   
   
 'Positive' Class : 1



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



Area under the curve: 0.8584

## Classification Tree model

When creating a classification tree, we first use 10-fold cross validation to partition the dataset into training and validation subset. By partitioning the dataset, we were able to increase the clarity of the node with the target variable—churn. By running the decision tree, it was resulted the variables that associated with “Churn” were Contract (Month-to-month, One year, Two year), Internet Service (DSL, Fiber Optic, No), and Tenure (number of months). We then made a prediction to test the customer churn. Based on the model, it has correctly predicted 1147 to be churn but classified an additional 722 to be churn. The model by analogy has misclassified 402 of the customers to be non-churned when actually they are. Based on the training data set, the accuracy is 79.2% and sensitivity and specificity is (40.07%, 93.36%). Whereas for the validation set’s accuracy has increase to 84.02% and sensitivity increased to 61.37%, and specificity decreased to 92.21%.

Confusion Matrix and Statistics  
   
 Reference  
 Prediction Yes No  
 Yes 6921 3210  
 No 9900 43356  
   
 Accuracy : 0.7932   
 95% CI : (0.79, 0.7963)   
 No Information Rate : 0.7346   
 P-Value [Acc > NIR] : < 0.00000000000000022  
   
 Kappa : 0.3924   
   
 Mcnemar's Test P-Value : < 0.00000000000000022  
   
 Sensitivity : 0.4114   
 Specificity : 0.9311   
 Pos Pred Value : 0.6832   
 Neg Pred Value : 0.8141   
 Prevalence : 0.2654   
 Detection Rate : 0.1092   
 Detection Prevalence : 0.1598   
 Balanced Accuracy : 0.6713   
   
 'Positive' Class : Yes

Confusion Matrix and Statistics  
   
 Reference  
 Prediction Yes No  
 Yes 1114 371  
 No 755 4803  
   
 Accuracy : 0.8401   
 95% CI : (0.8314, 0.8486)   
 No Information Rate : 0.7346   
 P-Value [Acc > NIR] : < 0.00000000000000022  
   
 Kappa : 0.5612   
   
 Mcnemar's Test P-Value : < 0.00000000000000022  
   
 Sensitivity : 0.5960   
 Specificity : 0.9283   
 Pos Pred Value : 0.7502   
 Neg Pred Value : 0.8642   
 Prevalence : 0.2654   
 Detection Rate : 0.1582   
 Detection Prevalence : 0.2108   
 Balanced Accuracy : 0.7622   
   
 'Positive' Class : Yes

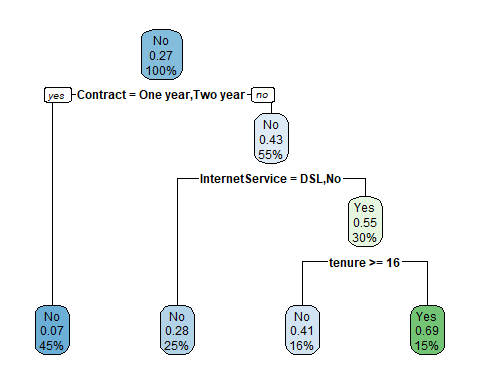
## Model Deployment with Classification Tree

Results:

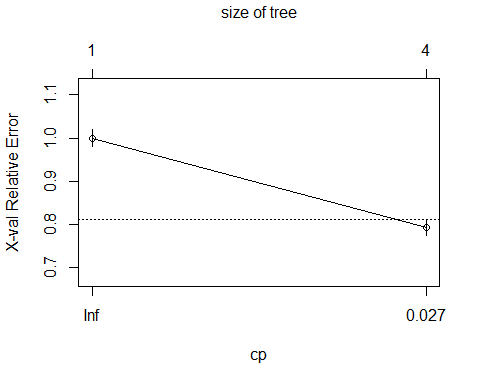
The best-pruned classification tree results showed the variables most associated to churning were Contract (Month-to-month, One year, Two year), Internet Service (DSL, Fiber Optic, No), and Tenure (number of months). The logistic regression showed SeniorCitizen, Dependents, Tenure, MultipleLines, InternetService, OnlineSecurity, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges and TotalCharges as significant predictors for churning.

Analysis:

Based on performance evaluation metrics from both models, we opted to use the classification tree model to better predict churn rate on the Telco data; although accuracy alone cannot determine a better model, the combined evaluation metrics (accuracy, sensitivity, etc.) of the classification tree model showed an improved performance on the validation data and was ultimately chosen to be used for the full data set.



Classification tree:  
 rpart(formula = Churn ~ ., data = telco.df, method = "class",   
 control = rpart.control(xval = 10))  
   
 Variables actually used in tree construction:  
 [1] Contract InternetService tenure   
   
 Root node error: 1869/7043 = 0.26537  
   
 n= 7043   
   
 CP nsplit rel error xerror xstd  
 1 0.070983 0 1.00000 1.00000 0.019826  
 2 0.010000 3 0.78705 0.79347 0.018307



## References

DiRienzo, N. (2020, August 24). ISTA 321 - Data Mining. 8 Classification. Retrieved May 9, 2022, fromhttps://bookdown.org/ndirienzo/ista\_321\_data\_mining/classification.html

Lunn, P. D., & Lyons, S. (2018). Consumer switching intentions for telecoms services: evidence from Ireland. Heliyon, 4(5), e00618. <https://doi.org/10.1016/j.heliyon.2018.e00618>

Malhotra, Arvind. (2013). Exploring switching behavior of US mobile phone customers. Journal of Services Marketing. 27. 13-24. 10.1108/08876041311296347.

S. O’Dea. (2020, February 27). AT&T, Verizon, Sprint, T-Mobile Churn Rate 2018. Statista. Retrieved May 8, 2022, from <https://www.statista.com/statistics/283511/average-monthly-churn-rate-top-wireless-carriers-us/>

Salhieh, S. M. (2019). Modeling the rationality of customers’ switching mobile services behavior. International Journal of Engineering Business Management. <https://doi.org/10.1177/1847979019865411>

YouGov. (2021, March 23). Cost is the main reason Americans switch phone carriers. why else do they jump ship? YouGov. Retrieved May 8, 2022, from <https://today.yougov.com/topics/technology/articles-reports/2021/03/23/americans-switch-phone-carriers>