https://www.kaggle.com/blastchar/telco-customer-churn
http://www.dbmarketing.com/telecom/churnreduction.html
https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

1.1 Business Problem Statement

For our main problem that we are researching, we have decided to focus on customer churn. In this sense we can see that in order to maintain a sustainable income, we must ensure that our customer relationship along with our churn is at acceptable levels. To do this we will examine the basic underlying issues of why customers disengage.

1.2 Business Goal

Our goal is to determine what factors are making customers churn and figure out methods to retain customers, thus slowing the churn rate. Decreasing churn is vital to any part of a business in order to keep revenue up and stay afloat.

1.3 Data Profile

This data came from Kaggle and is a csv file containing information about a telco company in California and shows which customers have left the company, stayed, signed up for the service(s), which service(s) the customers have, as well as some demographic and financial information. There are 11 missing values, all of which are in the total charges column, which were converted to NA's so the data can be easily usable, and there are no outliers. There were attributes. There were 21 attributes with Churn being the target.

2. Project Report

2.1 Importance

The importance of our data mining project is to effectively develop customer retention plans for a Telco company in hopes to improve the overall company churn rate. Through data mining, we will be able to analyze customer information and aim to predict future behavior that will be used to better understand why customers may exit. If a company has a high churn rate or has seen a consistent increase, they are in a position that could be damaging to their profitability and long-term success.

For Telco companies in general, the average annual churn rate is between 10 and 67 percent. This means that about 75 percent of the 17 to 20 million subscribers signing up with a new wireless carrier every year are coming from another wireless provider and hence are already churners (Hughes 2021). Not only is this damaging in the aspect that they lose customer revenue, but they will also lose the money spent to

initially attract the customer.

2.2 Background

When any company is attempting to increase their rate of retention, there are many approaches and techniques that can be utilized in order to best determine the most effective way forward. These techniques can vary from surface level to extreme deep dives into a user's history. No matter how the information is gathered, they all help add to deepen the understanding of what motivates the consumer.

One such technique is referred to as “affinity analysis”. You could look at this as a basket analysis of what the customer has purchased in the past to see what they are most interested in. For example, if you note that a customer tends to buy a certain protein bar each week, it may be advantageous to offer them a special on that brand's complementary products. When speaking on this issue, Neil Patel said, “Ask yourself: Who are your local customers and how you can turn these customers into advocates for your store”.

On a deeper level, data mining may help you to create a completely new product to better target a specific customer base. This can range from customizing an existing product to better fit a niche market, to innovation that results in a completely new product. You are looking for any gaps in service that have been overlooked and are readily available to capitalize on.

The two examples detailed here are only a small fraction of what data analysis can do. If you are looking to make your business more successful, into increasing the size of your business, or even changing direction entirely, employing these techniques is paramount to your success. Data mining is not just a simple way of examining how your customer base is reacting, but a detailed analysis of how to better predict the intricacies of the market in order to be best positioned for any changes that may come your way.

2.3 Data

This data came from Kaggle and is a csv file containing information about a telco company in California and shows which customers have left the company, stayed, signed up for the service(s), which service(s) the customers have, as well as some demographic and financial information. There are 11 missing values, all of which are in the total charges column, which were converted the to NA’s so the data can be easily usable, and there are no outliers

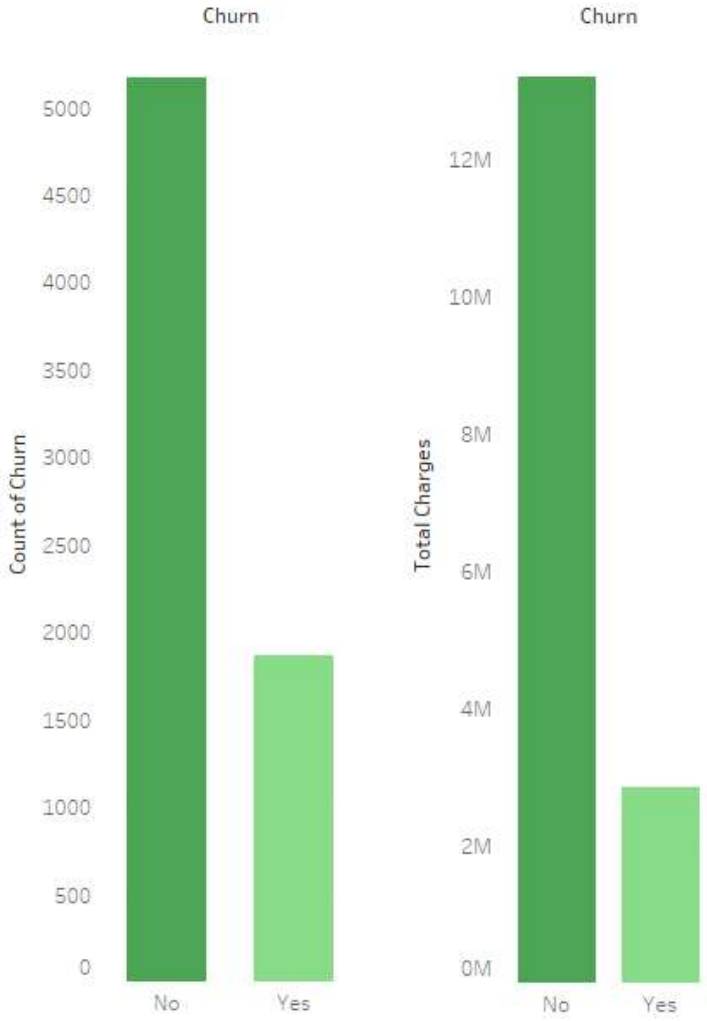
# of records: 7043
21 Attributes:
customerID - Customer ID

gender - Whether the customer is a male or a female
SeniorCitizen - Whether the customer is a senior citizen or not (1, 0)
Partner - Whether the customer has a partner or not (Yes, No)
Dependents - Whether the customer has dependents or not (Yes, No)
Tenure - Number of months the customer has stayed with the company
PhoneService - Whether the customer has a phone service or not (Yes, No)
MultipleLines - Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService - Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity - Whether the customer has online security or not (Yes, No, No internet service)
DeviceProtection - Whether the customer has tech support or not (Yes, No, No internet service)
TechSupport - Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV - Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies - Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract - The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling - Whether the customer has paperless billing or not (Yes, No)
PaymentMethod - The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges - The amount charged to the customer monthly
TotalCharges - The total amount charged to the customer
Churn - Whether the customer churned or not (Yes or No)

Table 1. Summary Statistics

	Observations	Mean	Std. Dev.	Min	Max
SeniorCitizen	7043	.01621468	.3686116	0	1
Tenure	7043	32.37115	24.55948	0	72
MonthlyCharges	7043	64.76169	30.09005	18.25	118.75
TotalCharges	7043	2283.3	2266.771	18.80	8684.80

Also use at least two visualization techniques to illustrate some of the important attributes in your data set.]



2.4 Method

We used “read.csv” to load the Telco data into R. We then dropped missing rows using the “complete cases” function and changed the Senior Citizen variable to a factor using “as.factor”. Next, we visualized the started data by comparing each character variable to churn utilizing the ggplot library to display graphs. Once the factor variables were analyzed, we then looked at the continuous variables and how they were paired with churn.

Before running any tests, we first had to prepare our data using multiple data functions. First, we organized all responses in the data with the word “No” into one category and then made each continuous variable standardized. Following this, we decided to mutate tenure from months into years and created a new tenure variable named “tenure_years”. Once the data was cleaned, dummy variables then had to be created using seed(100). The last steps in our data preparation included combining both the dummy variables and the standardized continuous variables in order to build a final set of data that we could split into a train and test group.

When building the logistic regression model, we used the stepAIC function to find the best fitting model from the prepared data. We then looked at the variable importance of this model and found that DeviceProtection and StreamingTV both had high p-values, leading us to remove them from the data. From this set of data we then tested a 50% cutoff and found that sensitivity was too low and needed to run a cutoff function to optimize our output, resulting in a 31%-32% cutoff. The decision tree model was the second model we built using the “rpart” libraries. Once finished, we then ran a confusion matrix function to compare the outputs to those of the logistic regression. The last model we created was the random forest, which was built using the “random forest” library. For our random forest we selected our ntree value to be 500 and our mtry to be 4. From here we ran a confusion matrix to again compare the output to our other models to see which resulted the best. We also charted a variable importance plot to show which variables played an important role in the churn rate. Once each model was finished , the final step was then to chart each of these models' AUC curve to select the best option.

2.5 Results

Looking at the final results of our analysis, the random forest model gave the most accurate output when finding the importance of each variable. The output returned an accuracy of 78.86% with sensitivity at 82.46% and specificity at 63.99%. The logistic regression model was the least accurate of the three at 75.59% and a near balanced sensitivity and specificity rate. The decision tree model was just slightly less accurate than the random forest model at 78.1% and had a sensitivity rate of 82.45% and a specificity rate of 61.38%.

The confusion outputs for the random forest model gave 1390 true positives and 291 true negative. The decision tree's confusion output returned slightly higher predictions with true positives 1433 and true negatives at 226. The 95% CI for the decision tree was (0.7681, 0.8036). For the logistic regression, our final output gave us 1197 true positives (No’s) and 430 true negatives (Yes’s).

Variable Importance across the three outputs was also able to show what the most important variables were when predicting our customer churn. Tenure, Charges (Monthly and Total), Type of Internet (InternetService.xFiber.optic / InternetService.xNo) , were our most important variables for this analysis.

Logistic Regression:

- Accuracy 75.59%,
- Sensitivity 75.75%
- Specificity 75.53%

DecisionTrees:

- Accuracy 78.1%,
- Sensitivity 82.45%
- Specificity 61.38%

RandomForest:

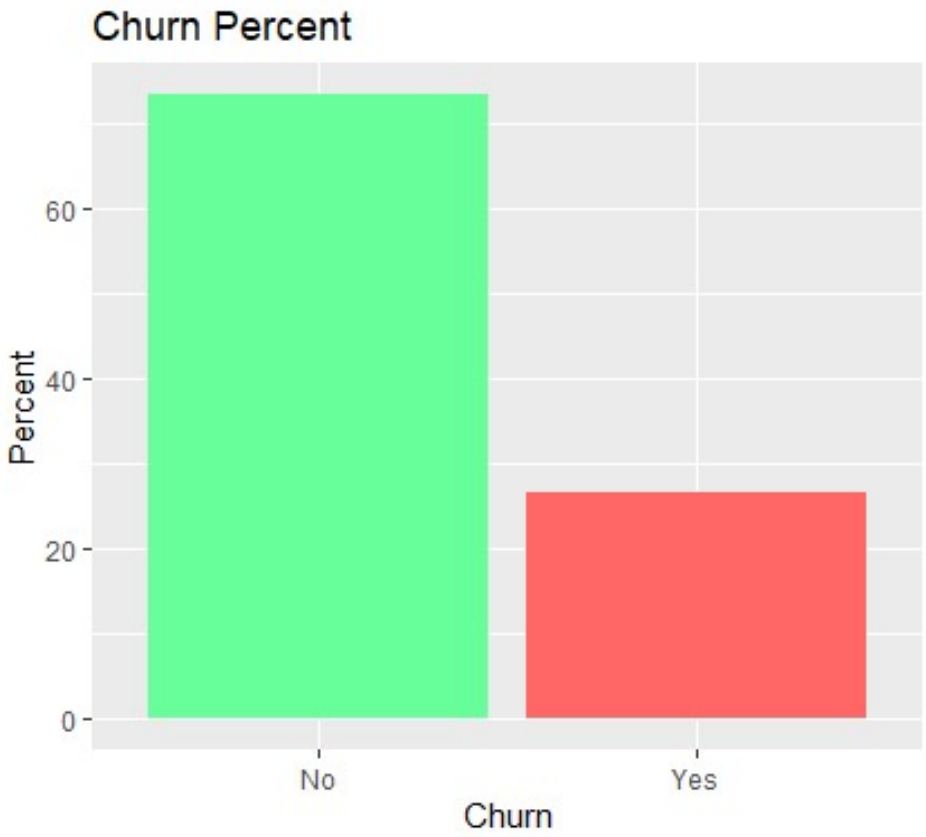
- Accuracy 78.86%,
- Sensitivity 82.46%
- Specificity 63.99%

2.6 Conclusion

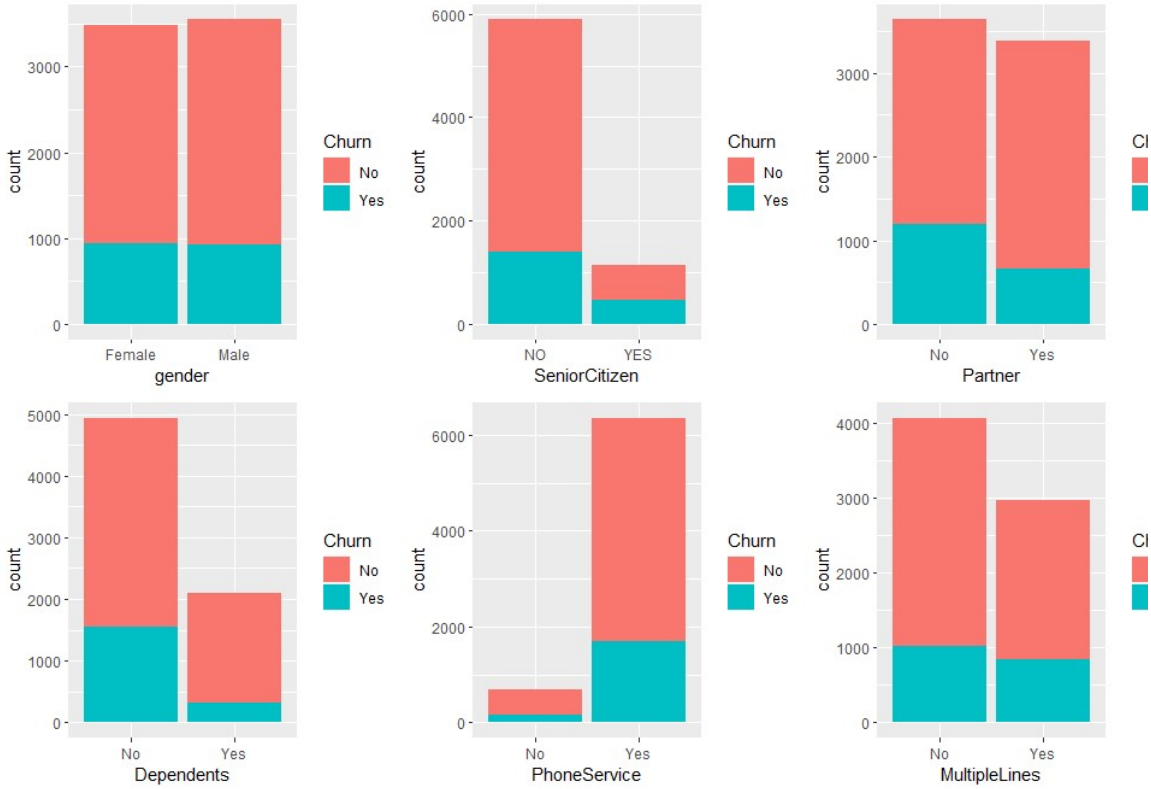
Data analysis is vital for any business that aims to succeed. Rate of retention and Churn are the two most important factors that influence whether a business will remain competitive or simply fade out. We worked with data pertaining to a Telco company with data on 21 different attributes. Our goal was to find out how much each variable affects churn, utilizing graphs and various different models to do so. In the end, it was the random forest model that provided the highest accuracy for determining the importance of each variable. The two most important important variables went hand in hand, those being total charges and tenure, which suggests that brand/company loyalty is a strong predictor of churn, thus it should be the goal for every company. Additionally, for services offered, it appeared that customers that had fiber optic internet were the least likely to switch, which means that the company should be trying to sell customers said package.

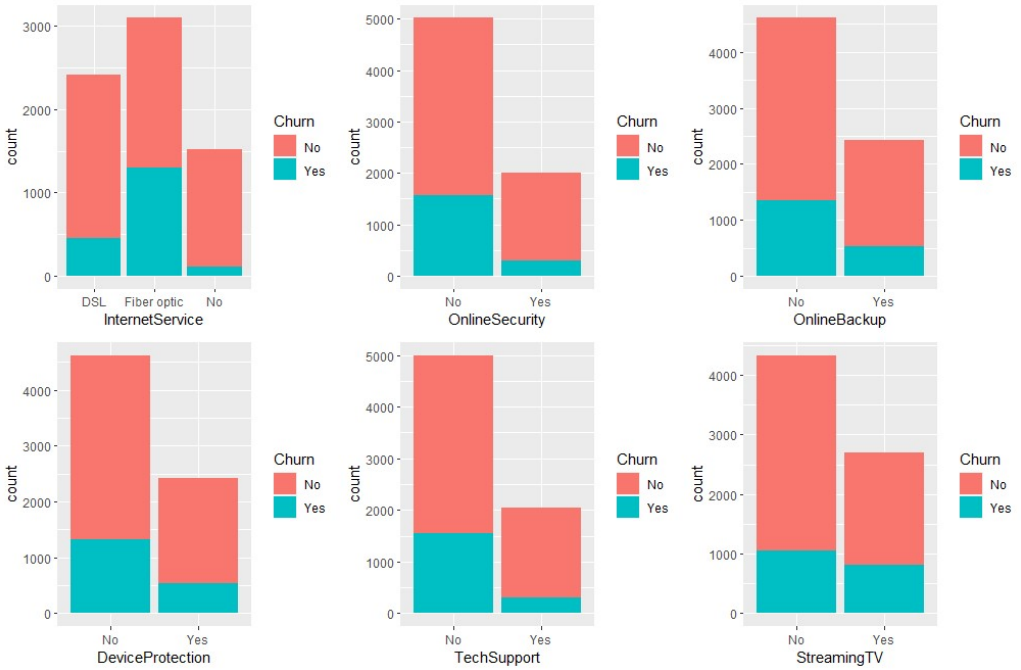
3 Appendix

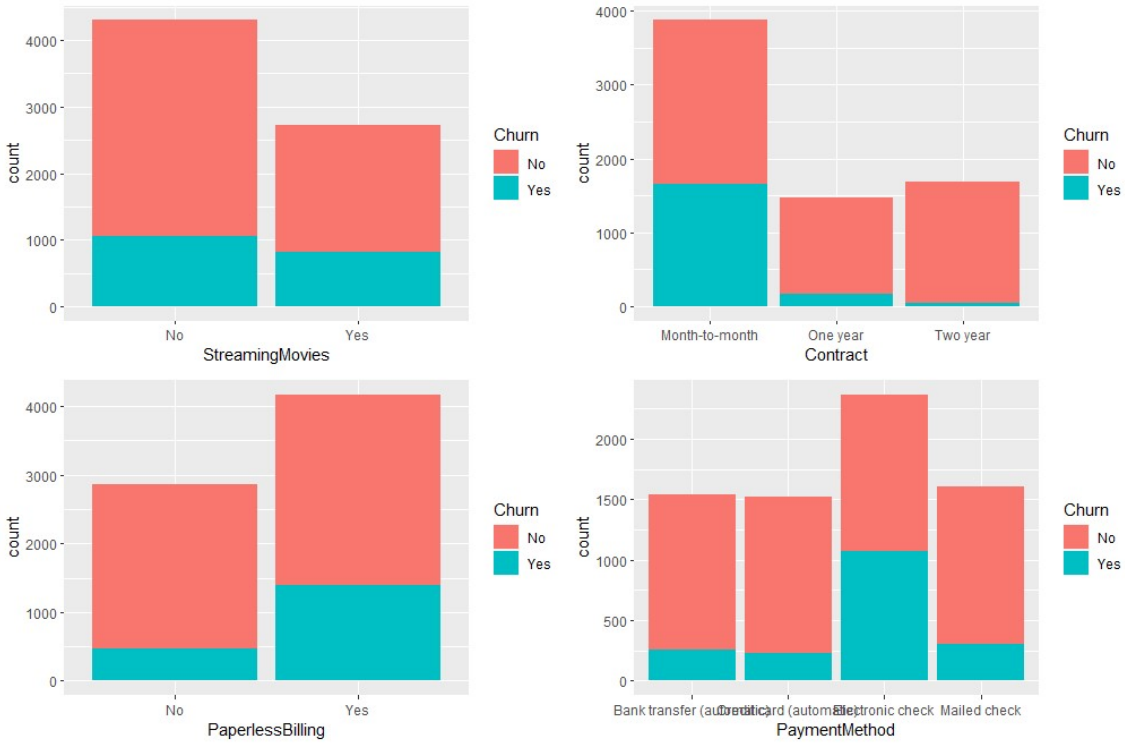
Churn Percentage Histogram



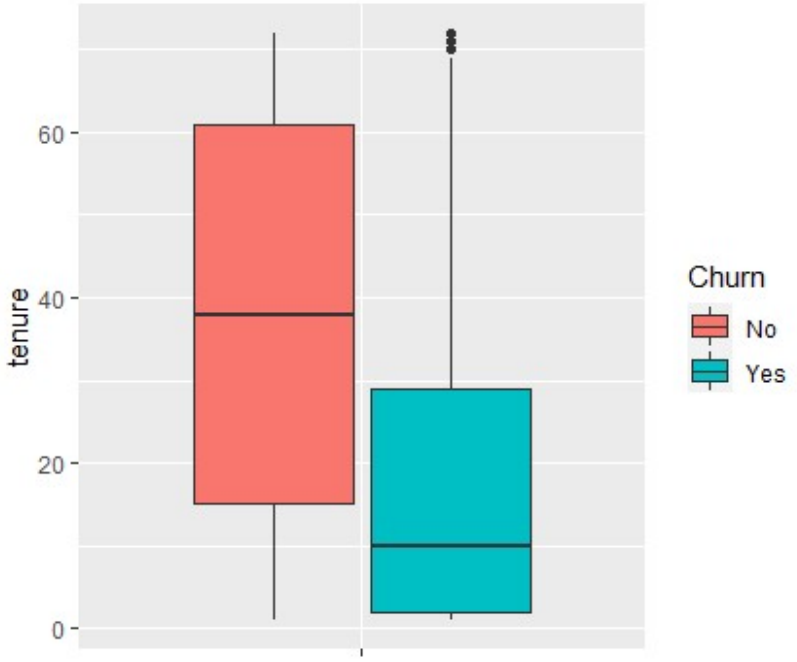
Factor Variables

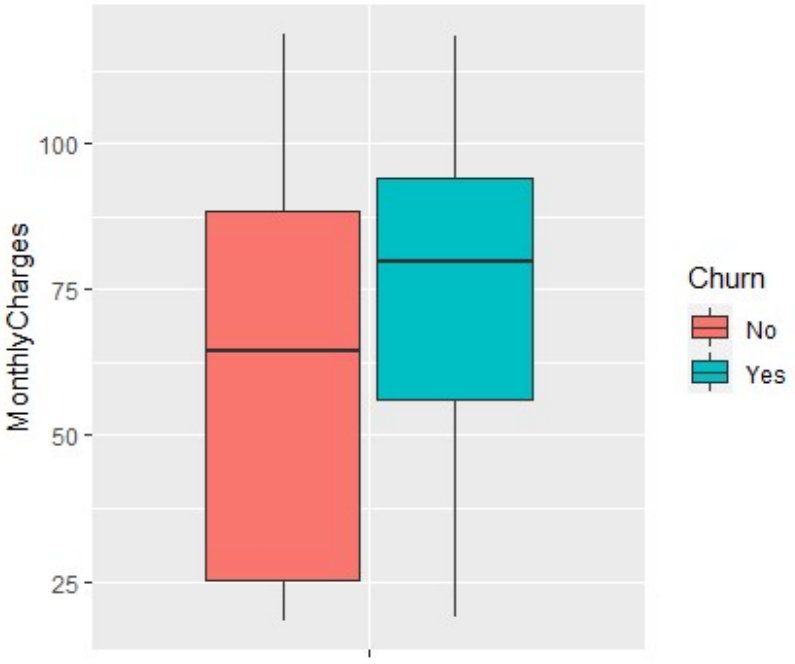

count
3000
2000
1000
0
Female
Male
gender
Churn
No
Yes
6000
4000
2000
NO
YES
SeniorCitizen
Partner
5000
Dependents
PhoneService
MultipleLines
Cl

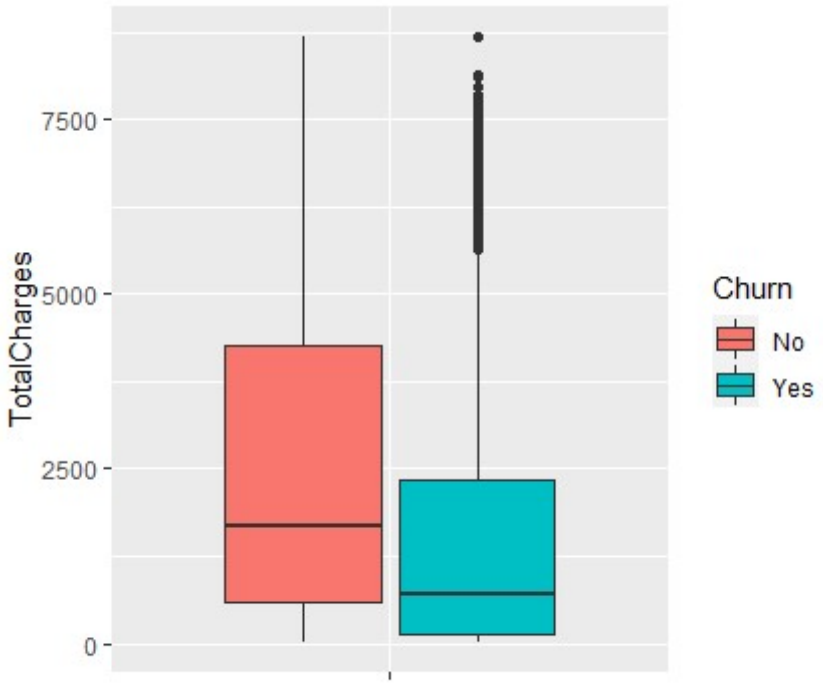

count
Churn
No
Yes
DSL
Fiber optic
No
InternetService
OnlineSecurity
OnlineBackup
DeviceProtection
TechSupport
StreamingTV


count
Churn
No
Yes
StreamingMovies
Month-to-month
One year
Two year
Contract
PaperlessBilling
Bank transfer (automatic)
Credit card (automatic)
Electronic check
Mailed check
PaymentMethod

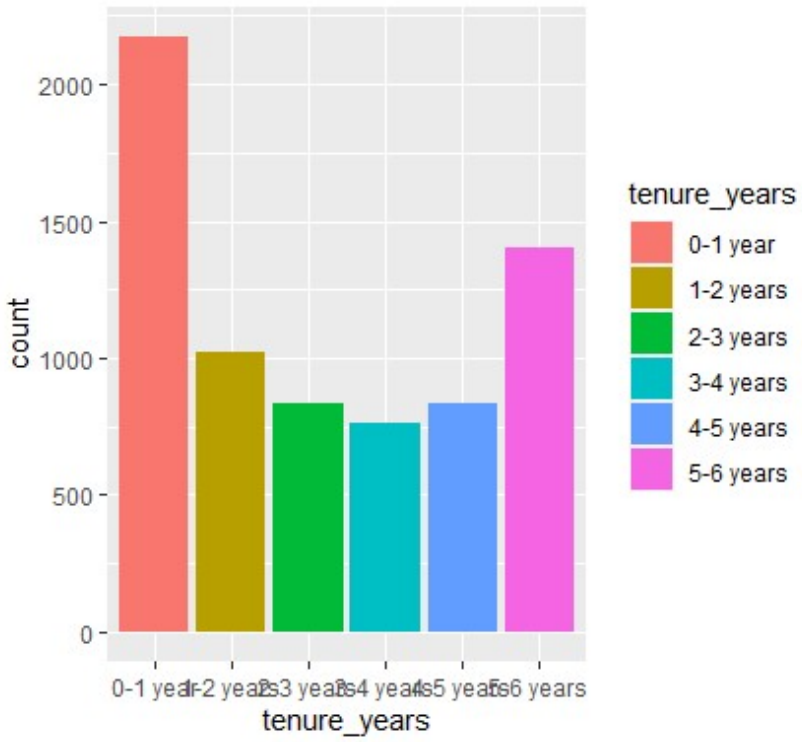
Continuous Variables



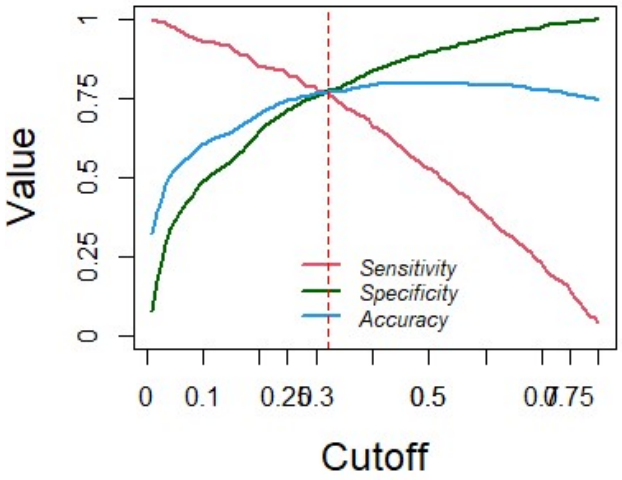

MonthlyCharges
100
75
50
25
Churn
No
Yes



Tenure Years Graph



Logistic Regression Cutoff Prediction



Logistic Regression Output

```
Call:
glm(formula = Churn ~ tenure + MonthlyCharges + SeniorCitizen +
    Partner + InternetService.xFiber.optic + InternetService.xNo +
    OnlineSecurity + OnlineBackup + TechSupport + StreamingTV +
    Contract.xOne.year + Contract.xTwo.year + PaperlessBilling +
    PaymentMethod.xElectronic.check + tenure_years.x1.2.years +
    tenure_years.x5.6.years, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9011  -0.6656  -0.2884   0.6922   3.1765

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.33139    0.17722  -7.513 5.79e-14 ***
tenure                          -0.81446    0.07291 -11.171  < 2e-16 ***
MonthlyCharges                   0.22584    0.15644   1.444 0.148841
SeniorCitizen                    0.15105    0.09893   1.527 0.126801
Partner                         -0.03739    0.08429  -0.444 0.657355
InternetService.xFiber.optic     0.64208    0.20221   3.175 0.001497 **
InternetService.xNo             -0.77204    0.18791  -4.109 3.98e-05 ***
OnlineSecurity                  -0.40827    0.10608  -3.849 0.000119 ***
OnlineBackup                    -0.19730    0.09453  -2.087 0.036880 *
TechSupport                     -0.40875    0.10784  -3.790 0.000150 ***
StreamingTV                      0.27704    0.11655   2.377 0.017455 *
Contract.xOne.year              -0.84029    0.12948  -6.490 8.59e-11 ***
Contract.xTwo.year              -1.71205    0.21954  -7.798 6.27e-15 ***
PaperlessBilling                 0.29894    0.08892   3.362 0.000774 ***
PaymentMethod.xElectronic.check  0.42222    0.08271   5.105 3.31e-07 ***
tenure_years.x1.2.years         -0.36192    0.10426  -3.471 0.000518 ***
tenure_years.x5.6.years          0.50492    0.19034   2.653 0.007984 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Logistic Regression Matrix:

actual_churn	pred_churn No	pred_churn Yes
No	1197	352
Yes	131	430

Logistic Regression Variable Importance

tenure	MonthlyCharges	SeniorCitizen
2.566687	13.185353	1.093448
Partner	InternetService.xFiber.optic	InternetService.xNo
1.131568	6.586757	2.235547
OnlineSecurity	OnlineBackup	TechSupport
1.222028	1.302089	1.284460
StreamingTV	Contract.xOne.year	Contract.xTwo.year
2.178736	1.314885	1.399638
PaperlessBilling	PaymentMethod.xElectronic.check	tenure_years.x1.2.years
1.133354	1.131859	1.032772
tenure_years.x5.6.years		
1.788311		

Decision Tree Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1433  116
         1  335  226

               Accuracy : 0.7863
                 95% CI : (0.7681, 0.8036)
    No Information Rate : 0.8379
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3746

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8105
            Specificity : 0.6608
         Pos Pred Value : 0.9251
         Neg Pred Value : 0.4029
             Prevalence : 0.8379
         Detection Rate : 0.6791
   Detection Prevalence : 0.7341
      Balanced Accuracy : 0.7357

       'Positive' Class : 0
```

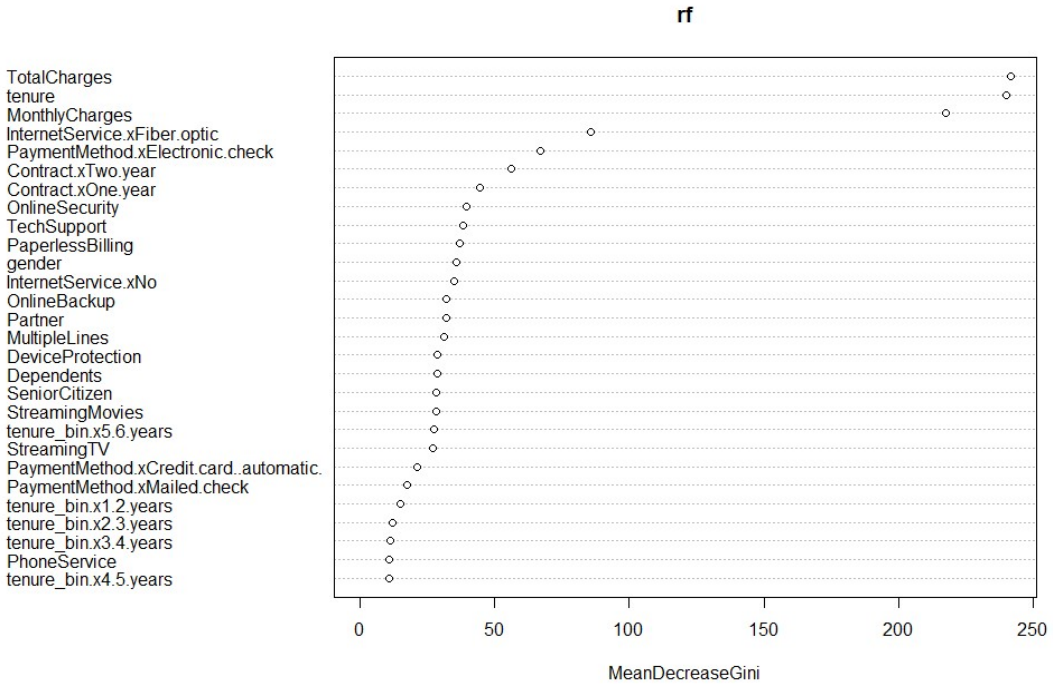
Decision Tree Variable Importance:

tenure	TotalCharges	InternetService.xFiber.optic
26	23	18
MonthlyCharges	PaymentMethod.xElectronic.check	MultipleLines
17	6	5
StreamingTV	PaymentMethod.xMailed.check	Partner
4	1	1

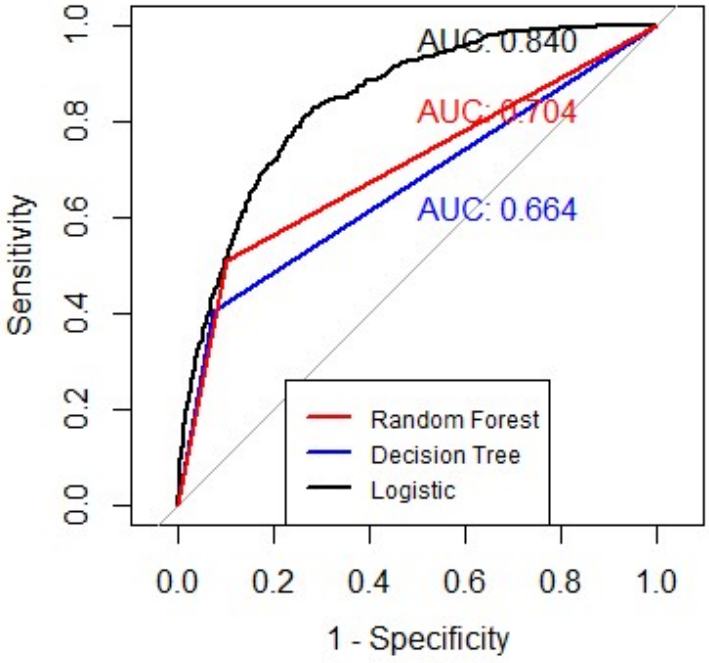
Random Forest Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1390  159
         1  270  291

               Accuracy : 0.7967
                 95% CI : (0.7789, 0.8137)
    No Information Rate : 0.7867
    P-Value [Acc > NIR] : 0.1377

                  Kappa : 0.4441

 Mcnemar's Test P-Value : 1.091e-07

            Sensitivity : 0.8373
            Specificity : 0.6467
         Pos Pred Value : 0.8974
         Neg Pred Value : 0.5187
             Prevalence : 0.7867
         Detection Rate : 0.6588
   Detection Prevalence : 0.7341
      Balanced Accuracy : 0.7420

       'Positive' Class : 0
```

Random Forest Variable Importance



AUC Comparison



A brief Summary of all the models:

Logistic Regression:

- Accuracy 75.59%,
- Sensitivity 75.75%
- Specificity 75.53%

DecisionTrees:

- Accuracy 78.1%,
- Sensitivity 82.45%
- Specificity 61.38%

RandomForest:

- Accuracy 78.86%,
- Sensitivity 82.46%

- Specificity 63.99%

Coding Citation

"Rahman, F (2018, September). Telco Customer Churn-LogisticRegression, Version 15. Retrieved
April 3 , 2021 from https://www.kaggle.com/farazrahman/telco-customer-churn-logisticregression."