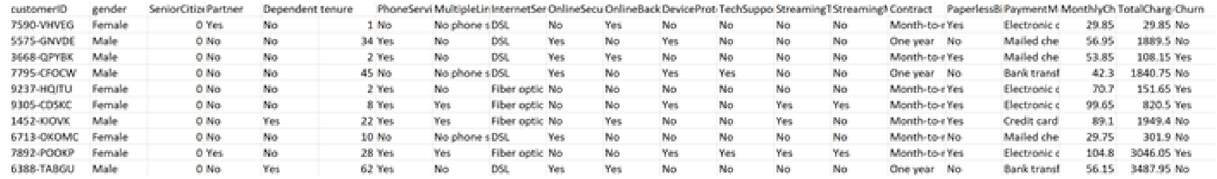
Gold Team 9
11/14/18
MSIS 510
Customer Retention: Data-Driven Insights

Introduction

Telco Direct is small telecommunications company based out of Orlando, Florida. The company provides with a variety of phone and internet packages. Additional services provided by Telco are Online Security, Online Backup, Device Protection, Tech Support, and Streaming TV and Movies. The company has three kinds of customers: phone-only, internet-only, and bundle customers. We have been provided with a dataset containing demographic, subscription, and payment information for 7,043 of Telco's customers. Our goal is to use data mining in order to develop a model that will predict customer churn; that is, will a customer cancel their service with Telco Direct at the end of the month.

We have a significant amount of data to interpret in this dataset. The target variable for our analysis will be the Churn, which is customer retention. On the next page, there is a bar chart displaying the proportion of customers that terminate their contract at the end of each month [Figure 1]. Additionally, we have three numerical columns. We have customer information on the length they have had a contract with Telco Direct measured in months. Additionally, we are provided with monthly and total charges to individual accounts. The categorical variables in the dataset contains the following customer demographic information: gender, senior citizen status, and number of dependents. There is also categorical information to what services each customer has as a part of their contract with Telco Direct. The following page contains a sample of the header of the dataset [Figure 2]. Using three different predictive models (Logistic Regression, Decision Tree, and Random Forest), our team will recommend the best model for predicting customer churn. Using the insights generated from data visualization and modeling, we will conclude with potential business solutions to combat the issue of high customer churn.

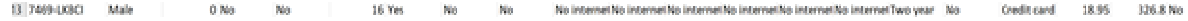
Sample Data



customerID	gender	SeniorCitiz	Partner	Dependent	tenure	PhoneServi	MultipleLin	InternetSer	OnlineSecu	OnlineBack	DeviceProt	TechSuppo	StreamingT	Streaming	Contract	PaperlessBi	PaymentM	MonthlyCh	TotalCharg	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone s	DSL	No	Yes	No	No	No	No	Month-to-r	Yes	Electronic c	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed che	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-r	Yes	Mailed che	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone s	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transf	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-r	Yes	Electronic c	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-r	Yes	Electronic c	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-r	Yes	Credit card	89.1	1949.4	No
6713-OKOMC	Female	0	No	No	10	No	No phone s	DSL	Yes	No	No	No	No	No	Month-to-r	No	Mailed che	29.75	301.9	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-r	Yes	Electronic c	104.8	3046.05	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transf	56.15	3487.95	No

The dataset we were provided has been maintained relatively well, so the steps that we took to clean it were not excessive. The total amount of customers in the dataset was 7,043. We first queried the data to check for empty or null values across any columns. There were 11 rows of customers with null values in the monthly and total charge field. These values were null because it was the first month for these customers; thus, we did not have their billing information. We decided to remove these 11 rows from the dataset because they did not represent a statistically significant proportion of the population. Ultimately, we removed rows account for .0015% of the total customer base.

One additional step that we took related to releveling the data in various service provided columns. If a customer only has phone service without internet, then each bundle column reads "No Internet Service." Using R, we releveled each instance of "No Internet Service" to read simply as "No."

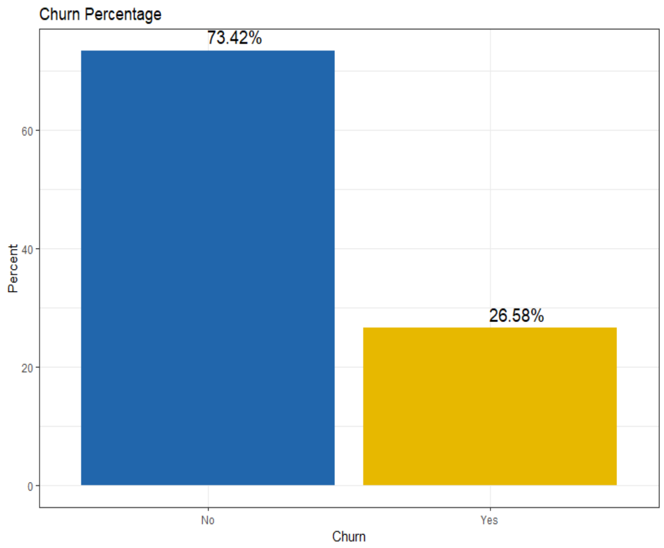


13	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet	No internet	No internet	No internet	No internet	No internet	Two year	No	Credit card	18.95	326.8	No

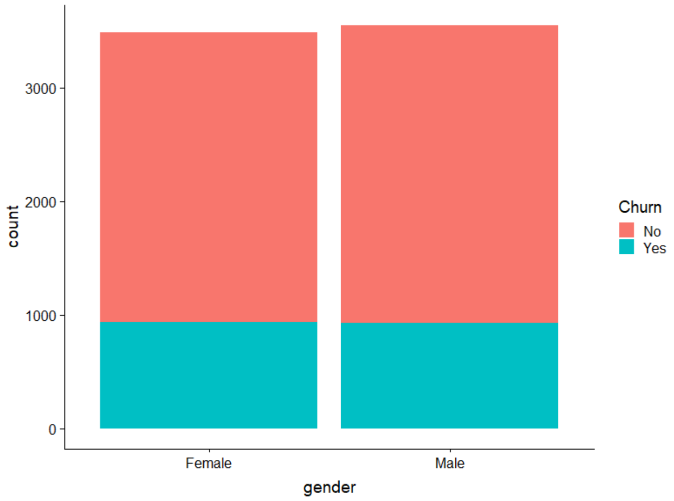
We concluded to relevel the data for multiple reasons. First, it creates cleaner and more readable models. Additionally, from a reporting and analysis perspective, these two values are identical. By choosing not to have internet service, customers implicitly turn down additional perks that come with that subscription, such as security or streaming.

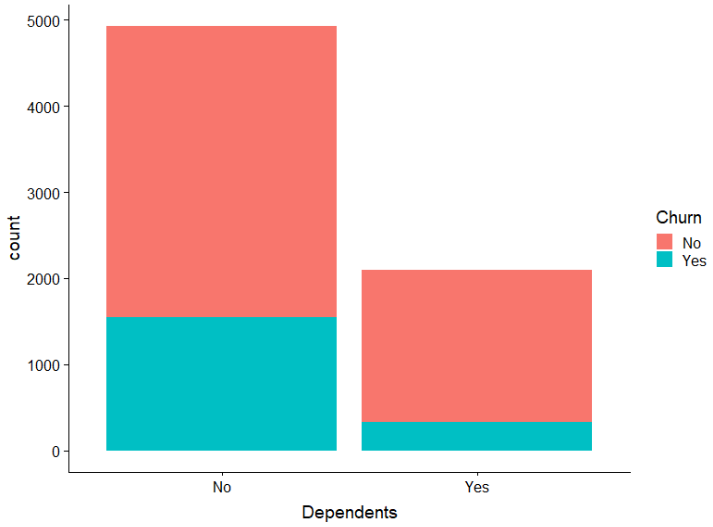
Data Visualization

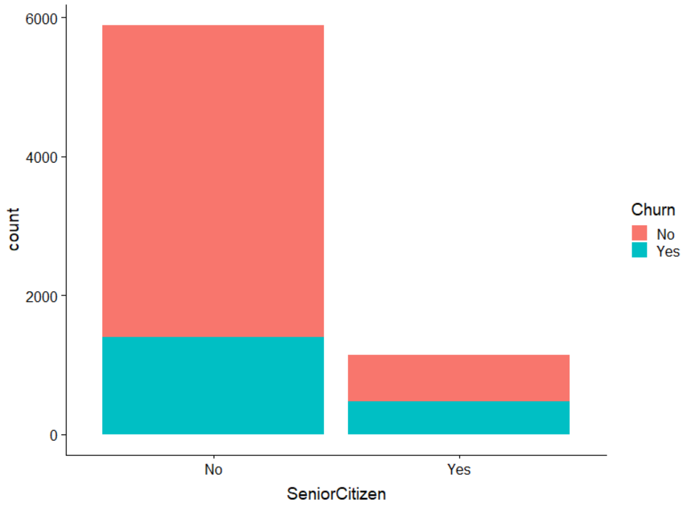
In R, we initially generate a bar chart showing the percentage of Yes and No in our target variable -- Churn. It gives us 73.42% of customers stayed with us in the last month and 26.58% customers chose to churn. Our goal is to predict the 26.58% customers and to decrease the percentage ultimately.

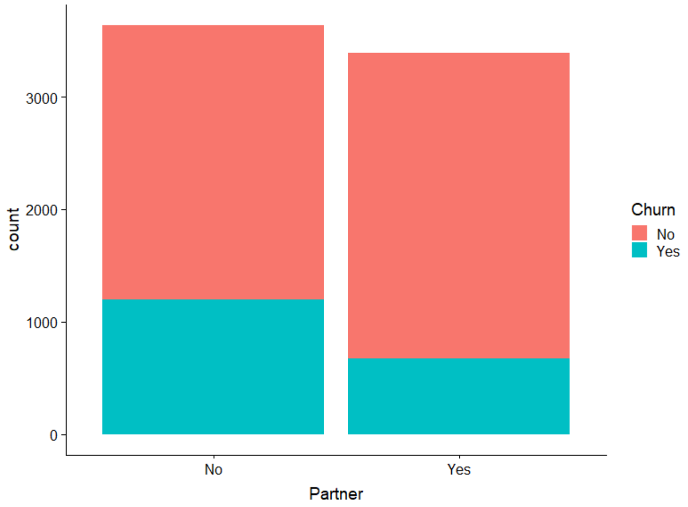


We analyze the relationship between each individual variable and the target variable. First, we analyze the customers' demographic information such as gender, dependent condition, senior citizen and partner condition. We conclude gender has almost no impact for churn. Dependent condition, senior citizenship and partner condition will be considered for the further analysis in our model. Customers with no dependents churned more. Total numbers of senior citizens who churned are less than those of non senior citizens. But the percentage of senior citizens who churned 40% to 50% which is a large portion. And customers with no partner churned more.

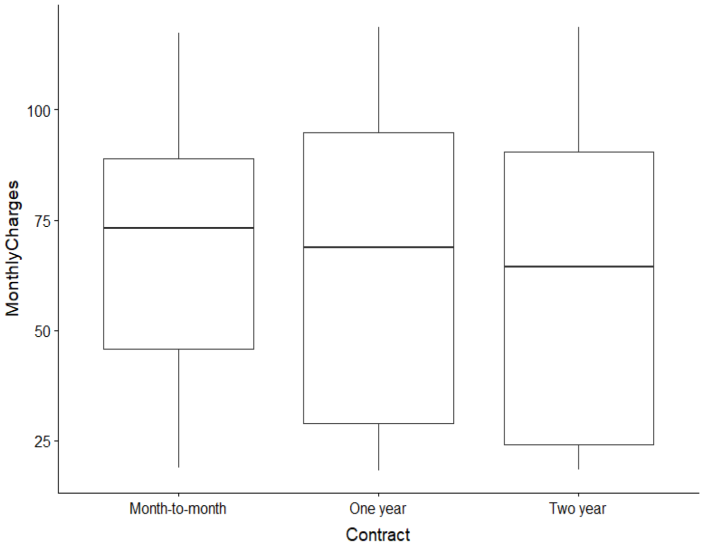

count
3000
2000
1000
0
Female
Male
gender
Churn
No
Yes

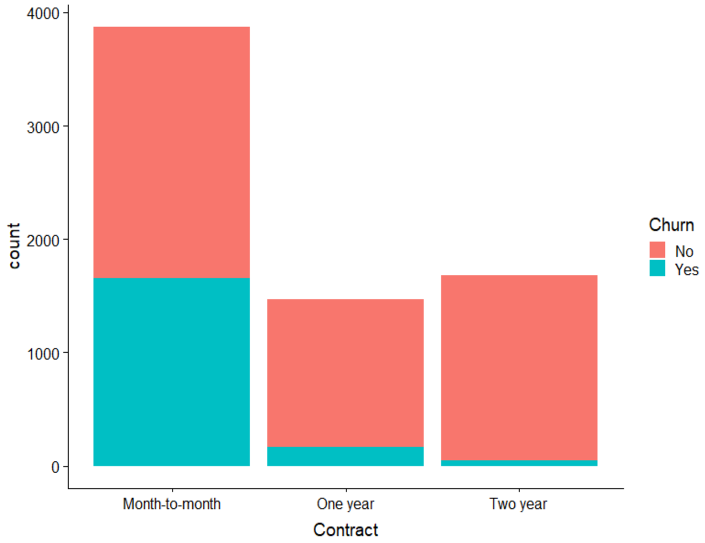

count
5000
4000
3000
2000
1000
0
No
Yes
Dependents
Churn
No
Yes



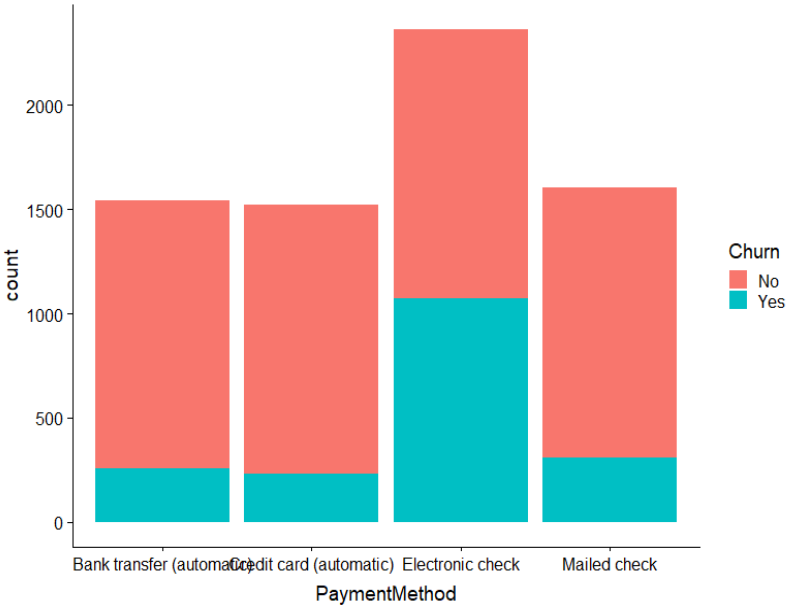


Then, we visualize the contract and churn which gives us that contract is an important variable for our model. We can tell from the below bar chart, month-to-month customers tend to churn the most comparing to one-year and two-year customers. The boxplot shows that month-to-month customers have the highest monthly fee about $75, so this indicates that these customers with higher monthly charges tend to leave the company.

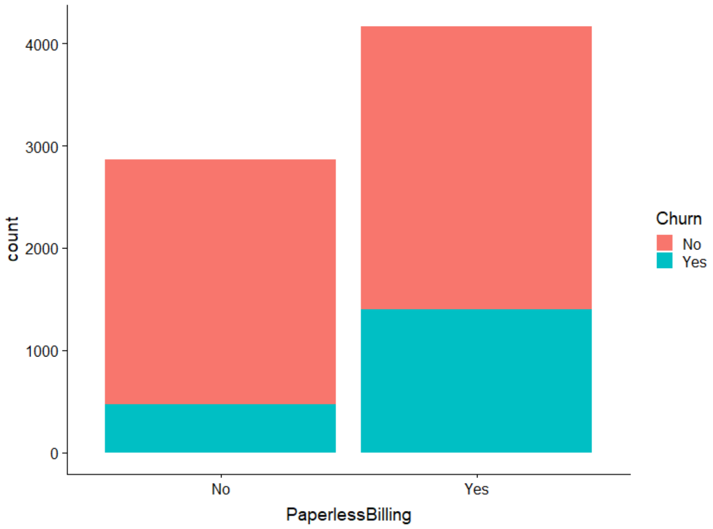




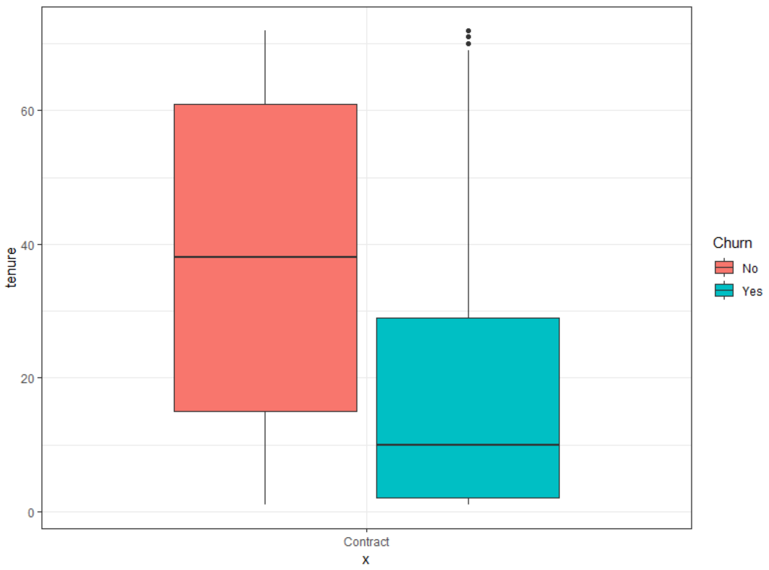
We then analyze whether the payment method will have a big impact on churn using a bar chart. It shows that customers using electronic check as a payment method are the easiest to leave the company which makes sense. Later we will show the relationship between payment method and people who use these payment method.



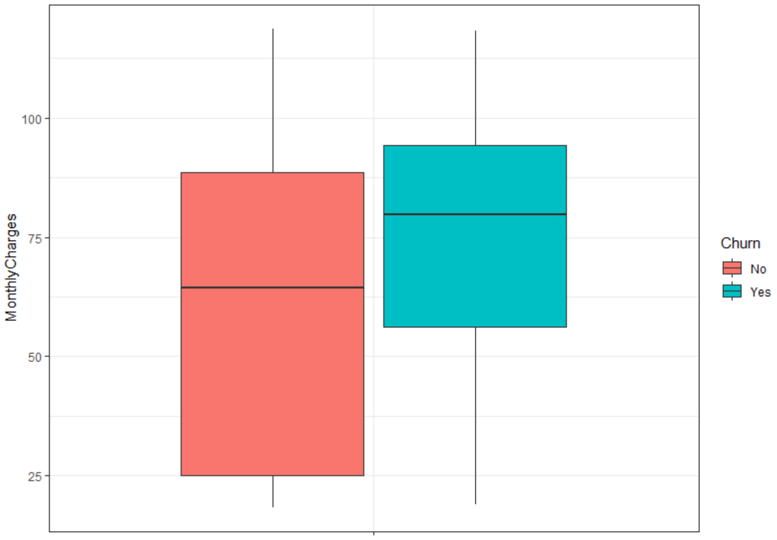
We continue to analyze the impact of paperless billing on churn. We can see that the PaperlessBilling variable has a large impact for our target variable which means we should consider about including this factor into our model.



Then we make a boxplot to figure out the average tenure for those customers who chose to leave us in the last month. We can get the result from the chart that they have a 10-month tenure period with us. This is to help giving the business insight that 10-month is a tricky point. The company could convince customers who already has less than or close to 10 month contract to change to a one-year contract by offering them great deals.



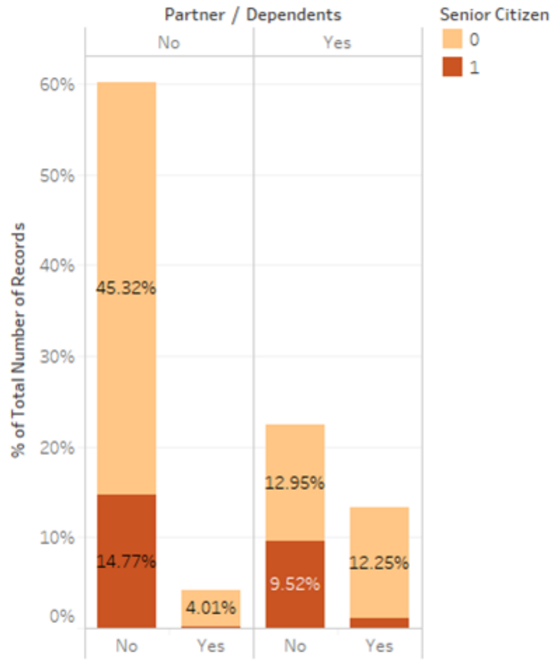
This boxplot tells us that churned customers have been charged more for monthly fees than customers who stay with the company. This also could be a reason for them to churn so we will take MonthlyCharges as an important variable and put it into the logistic regression.



We used Tableau for some more visualizations of data to determine the other relationships between a combination of various attributes and the target Variable.

1. Impact of Demographic data - SeniorCitizen, Partner and Dependent on Churn.

Churn : Y

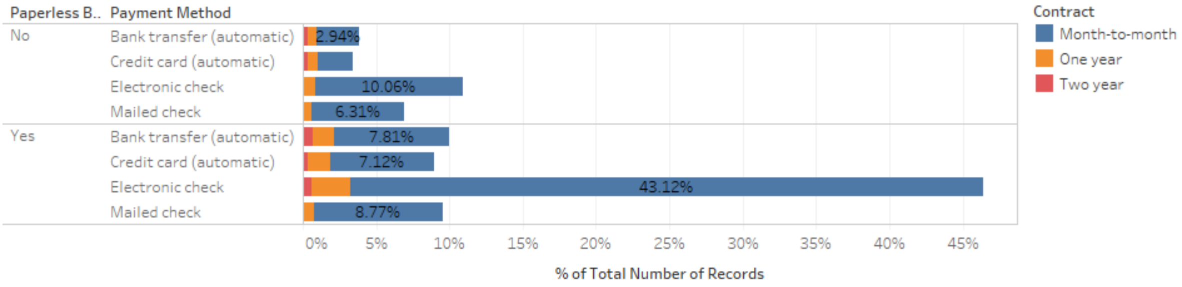


% of Total Number of Records for each Dependents broken down by Partner. Color shows details about Senior Citizen. The data is filtered on Churn and Multiple Lines. The Churn filter keeps Yes. The Multiple Lines filter keeps No, No phone service and Yes. Percents are based on the whole table.

This graph shows that for all the users, the maximum churn is with people with no partners and no dependents. Also, it looked like senior citizen was a huge factor but drilling down we can see that it is a small percentage, while non senior citizens are the ones who churn the most.

2. We analyze the various payment and term related measures to see the impact it has on Churn.

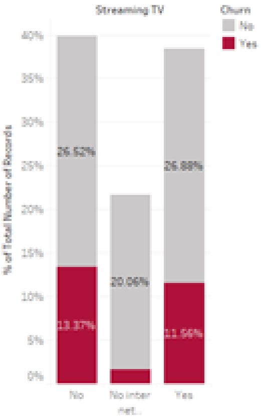
Churn: Y , By PaperLess billing and Payment Method



We can see that while paperless billing has a higher percentage of people Churning, for paperless as well as paper billing, Churn is more for Electronic Check payment method.

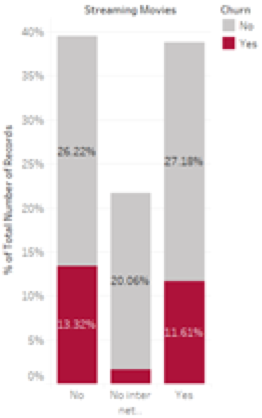
3. Now we will examine the relationship between the various services offered :

Churn % by if they
Stream TV



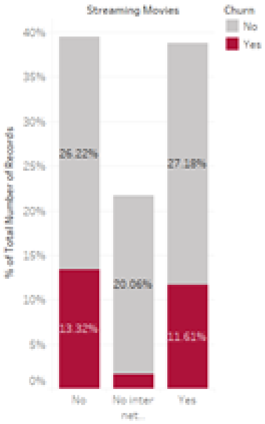
% of Total Number of Records for each Streaming TV. Color shows details about Churn. Percents are based on the whole table.

Churn % by if they
Stream Movies



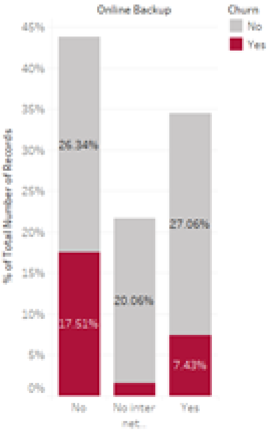
% of Total Number of Records for each Streaming Movies. Color shows details about Churn. Percents are based on the whole table.

Churn % by if they
Stream Movies



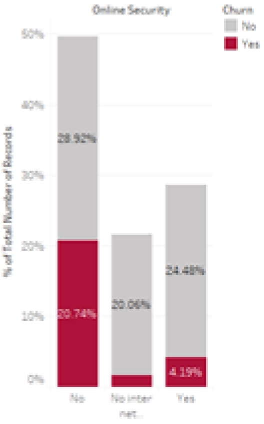
% of Total Number of Records for each Streaming Movies. Color shows details about Churn. Percents are based on the whole table.

% of users who use
Online Backup



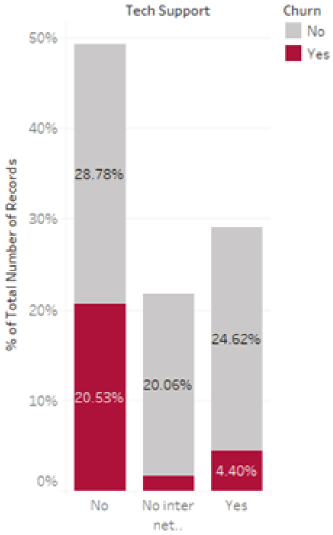
% of Total Number of Records for each Online Backup. Color shows details about Churn.

Churn % by online
security

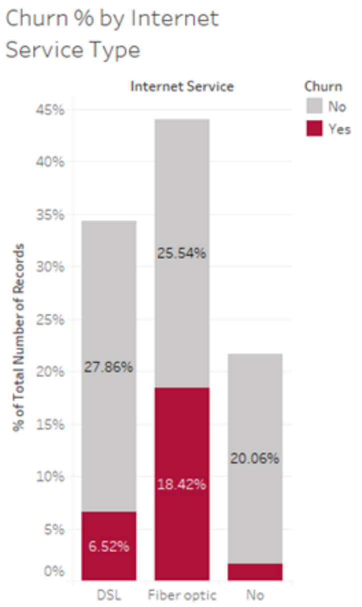


% of Total Number of Records for each Online Security. Color shows details about Churn. Percents are based on the whole table.

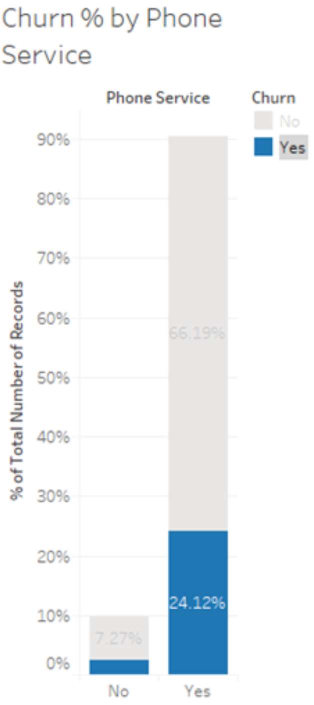
Churn % by if they avail
Tech Support



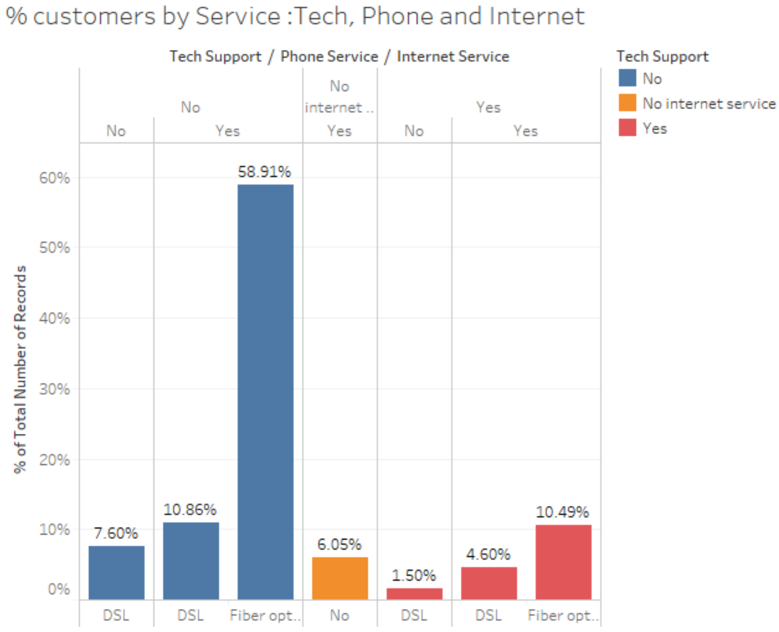
% of Total Number of Records for each Tech Support. Color shows details about Churn. Percents are based on the whole table.



% of Total Number of Records for each Internet Service. Color shows details about Churn. Percents are based on the whole table.

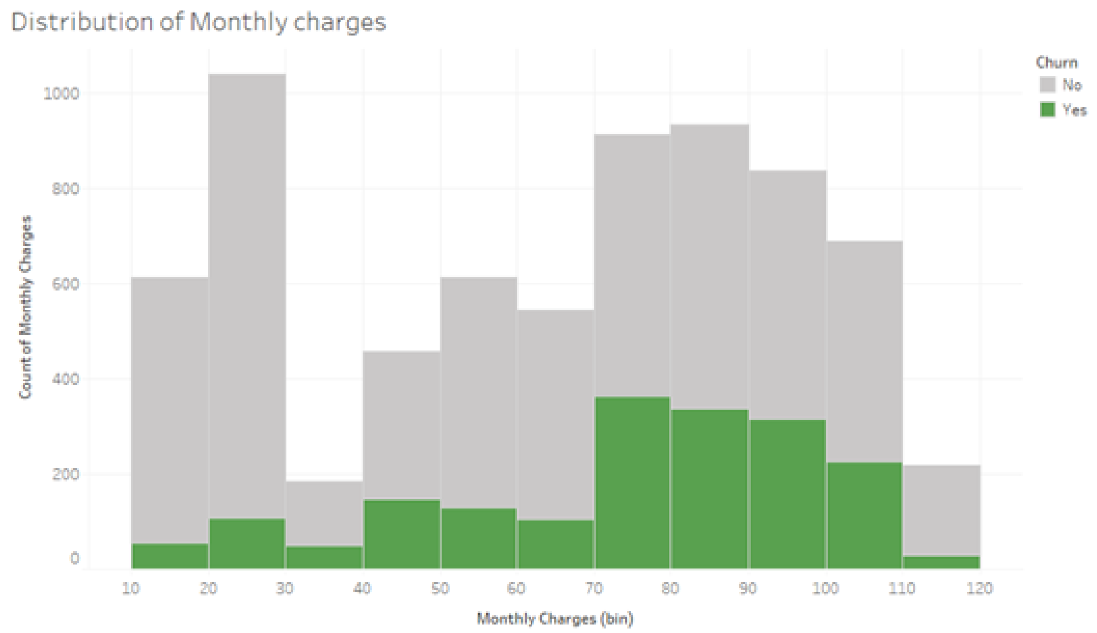


On a closer look to determine the relationship between the customers who churn the most by service type, we can see that people with Phone service as yes and internet service as Fiber Optic churn the most. Couple that with Tech Support and you can see that the ones that do not avail Tech support churn the most.This is illustrated in the graph below:

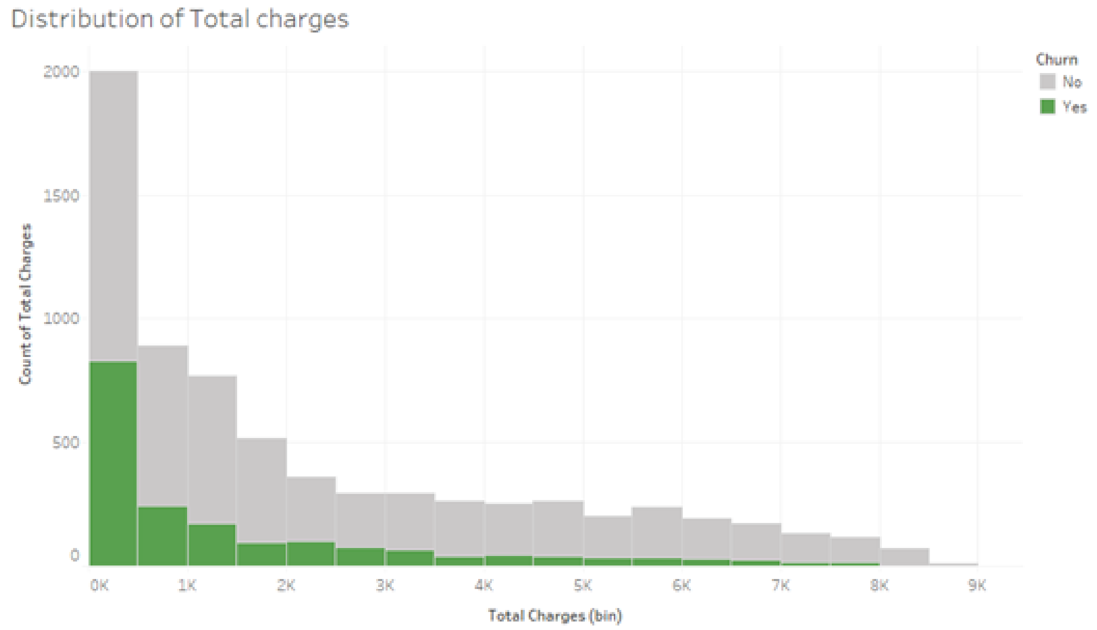


% of Total Number of Records for each Internet Service broken down by Tech Support and Phone Service. Color shows details about Tech Support. The data is filtered on Churn, which keeps Yes. Percents are based on the whole table.

4. We see the distribution of the monthly and total charges :

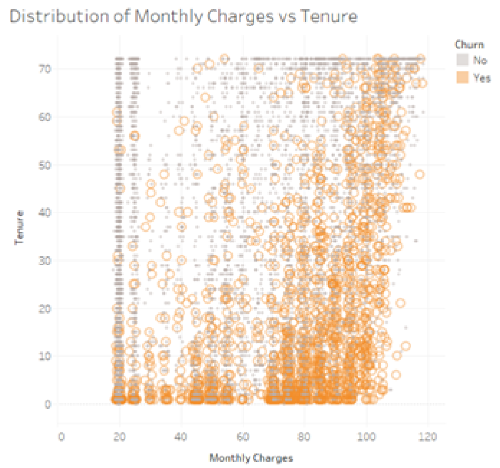


The trend of count of Monthly Charges for Monthly Charges (bin). Color shows details about Churn.

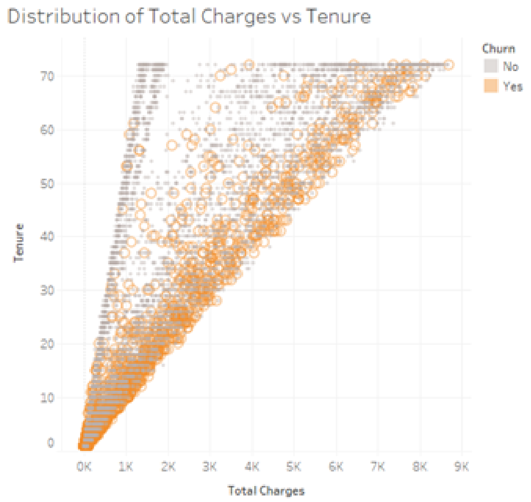


The trend of count of Total Charges for Total Charges (bin). Color shows details about Churn.

It looks like people with low total charges and high monthly charges churn the most. We need to further analyse this.

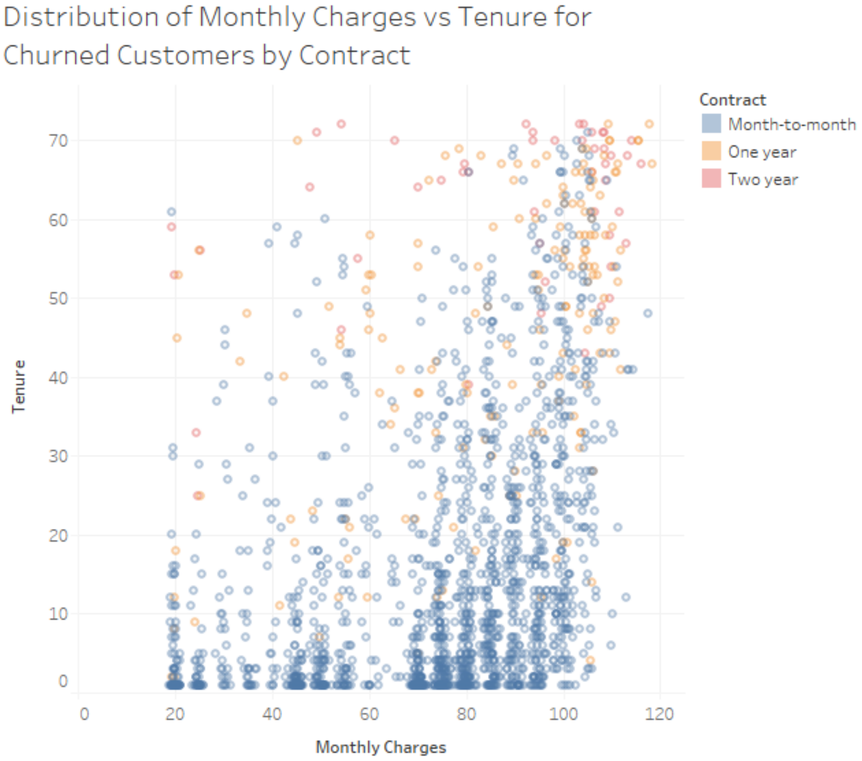


Monthly Charges vs. Tenure. Color shows details about Churn. Size shows details about Churn.



Total Charges vs. Tenure. Color shows details about Churn. Size shows details about Churn.

We can see that customers with less total charges also had a smaller tenure which explains the distribution. And these are the customers that are often seen leaving the company earlier.

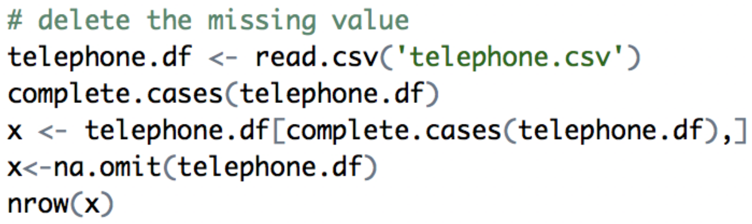


Further analysing this shows that irrespective of the monthly charges, customers with month to month contracts are the ones to leave within the first 10 months. Customers churn lesser if their monthly bills are lower.

Data Cleaning

The process of data cleaning is not as hard as we initially thought since the dataset itself is in a good condition. We use the “ sum(is. na()) “ function to check whether there is missing value in every column.

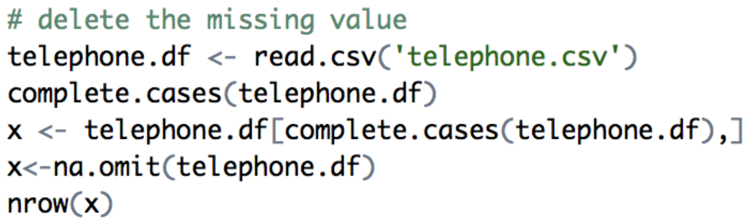
Only 11 missing values are found in “ Total Charges” column. Compared with 7043 rows in total, we decided to delete those 11 rows using “ omit “ function. Even though the cleaning process seems easy, we became proficient in those two essential functions in R.



```
# delete the missing value
telephone.df <- read.csv('telephone.csv')
complete.cases(telephone.df)
x <- telephone.df[complete.cases(telephone.df),]
x<-na.omit(telephone.df)
nrow(x)
```

Normalization

Based on initial analysis of the dataset, more than 10 variables are categorical in 21 columns. Therefore, we decided to apply the logistic regression and classification tree. For those variables who have only 2 categories, we just make sure they have been set as categorical data using “factor” functions. For those who owns 3 categories, we set one of them as the base using “relevel” function. In addition, one tricky part during our normalization is we found some columns include ‘No, No Internet/Phone Service and Yes’. After the discussion, we think ‘No Internet Service’ can be considered as ‘NO’. So, instead replacing ‘No Internet Service’ with ‘NO’, we set the both levels to be ‘NO’.



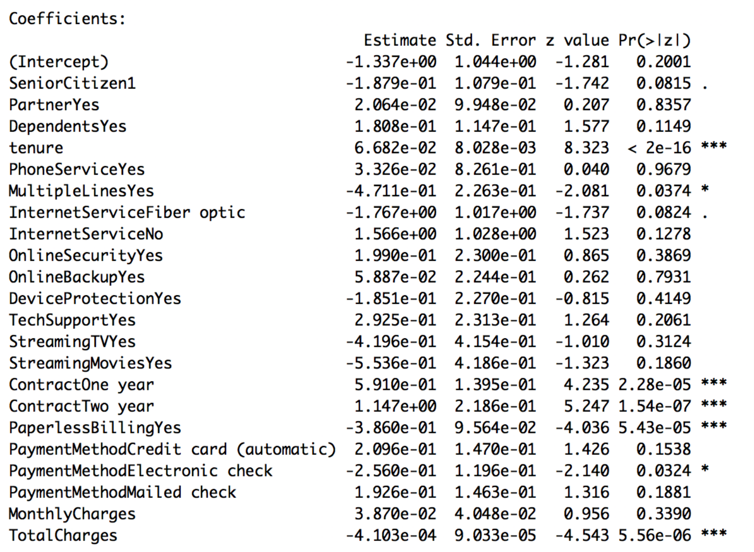
```
# delete the missing value
telephone.df <- read.csv('telephone.csv')
complete.cases(telephone.df)
x <- telephone.df[complete.cases(telephone.df),]
x<-na.omit(telephone.df)
nrow(x)
```

During the normalization process, we found the “str” is an useful function. Every time after we change or set the levels of variables, it is always a good idea to double check the attribute of data. We forgot to set the target value as categorical data when we implement the logistic regression, it’s easy and convenient to find which data are wrong by using “str” function especially when you are handling with a number of variables.

Models

Logistic Regression

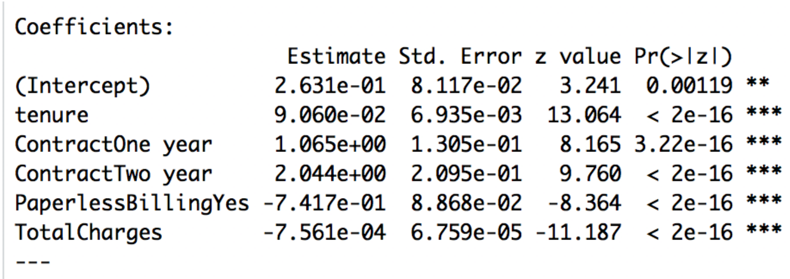
The dataset now is ready to do the logistic regression since the data cleaning and normalization parts have been totally done. Because of the special attribute of churn, we recode the target value “churn” to be class 1. We’ve tried 2 regression models. The first one is including all the variables except the “customerID” and “gender”. Apparently, “customerID” has no impact on the target value. As for gender, it shows a weak relationship with the target based on the visualization mentioned above. By running the “glm” function, we got our first logistic regression model. For the validation part, we’ve tried different cutoff such as 0.3, 0.5 and 0.6. It turns out that the 0.5 cutoff has the highest accuracy.



```
Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         -1.337e+00  1.044e+00  -1.281   0.2001    
SeniorCitizen1                      -1.879e-01  1.079e-01  -1.742   0.0815 .  
PartnerYes                           2.064e-02  9.948e-02   0.207   0.8357    
DependentsYes                        1.808e-01  1.147e-01   1.577   0.1149    
tenure                               6.682e-02  8.028e-03   8.323  < 2e-16 ***
PhoneServiceYes                      3.326e-02  8.261e-01   0.040   0.9679    
MultipleLinesYes                    -4.711e-01  2.263e-01  -2.081   0.0374 *  
InternetServiceFiber optic          -1.767e+00  1.017e+00  -1.737   0.0824 .  
InternetServiceNo                    1.566e+00  1.028e+00   1.523   0.1278    
OnlineSecurityYes                    1.990e-01  2.300e-01   0.865   0.3869    
OnlineBackupYes                      5.887e-02  2.244e-01   0.262   0.7931    
DeviceProtectionYes                 -1.851e-01  2.270e-01  -0.815   0.4149    
TechSupportYes                       2.925e-01  2.313e-01   1.264   0.2061    
StreamingTVYes                      -4.196e-01  4.154e-01  -1.010   0.3124    
StreamingMoviesYes                  -5.536e-01  4.186e-01  -1.323   0.1860    
ContractOne year                     5.910e-01  1.395e-01   4.235 2.28e-05 ***
ContractTwo year                     1.147e+00  2.186e-01   5.247 1.54e-07 ***
PaperlessBillingYes                 -3.860e-01  9.564e-02  -4.036 5.43e-05 ***
PaymentMethodCredit card (automatic)  2.096e-01  1.470e-01   1.426   0.1538    
PaymentMethodElectronic check       -2.560e-01  1.196e-01  -2.140   0.0324 *  
PaymentMethodMailed check            1.926e-01  1.463e-01   1.316   0.1881    
MonthlyCharges                       3.870e-02  4.048e-02   0.956   0.3390    
TotalCharges                        -4.103e-04  9.033e-05  -4.543 5.56e-06 ***
```

Confusion Matrix and Statistics of Regression 1		
	Reference	
Prediction	0	1
0	408	233
1	334	1848
Accuracy	0.802	

Based on the regression 1, we've noticed that 4 variables have high z-values, which are used to tell us which values are the most statistically significant, independent of other variable movement. We chose "tenure", "Contract", "PaperlessBilling", and "TotalCharges to run a second regression with less variables.



```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         2.631e-01  8.117e-02   3.241  0.00119 ** 
tenure              9.060e-02  6.935e-03  13.064  < 2e-16 ***
ContractOne year    1.065e+00  1.305e-01   8.165 3.22e-16 ***
ContractTwo year    2.044e+00  2.095e-01   9.760  < 2e-16 ***
PaperlessBillingYes -7.417e-01  8.868e-02  -8.364  < 2e-16 ***
TotalCharges        -7.561e-04  6.759e-05 -11.187  < 2e-16 ***
---
```

Confusion Matrix and Statistics of Regression 2		
	Reference	
Prediction	0	1
0	394	241

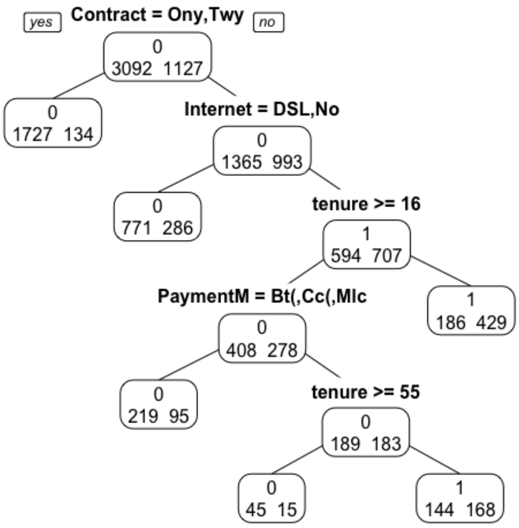
1	393	1830
Accuracy	0.7746	

The accuracy of this predictive model is slightly lower, about 2%; however, the standard error of all the variables and the intercept is much lower. This is useful for removing useless noise from the data, indicating that some of our variables are likely dependant on other factors, and insignificant on their own. Both of these regressions provide useful insights, but greater specificity can give us more confidence in the future predictive power of our regression 2.

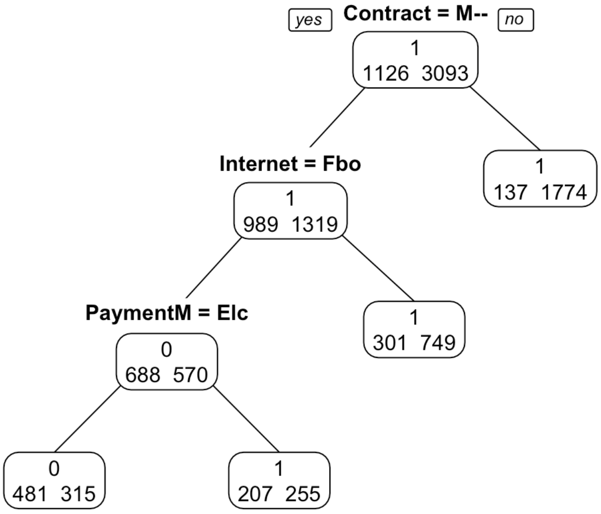
Classification Tree

The classification tree is implemented smoothly after data cleaning and normalization. Also, we've tried different columns mix to get a higher accuracy (Tree2). In the end, we found the first one which contains most columns has the highest accuracy.

Confusion Matrix and Statistics of Classification Tree1		
	Reference	
Prediction	0	1
0	382	233
1	360	1838
Accuracy	0.7892	



Classification Tree 2



Confusion Matrix and Statistics of Classification Tree2		
	Reference	
Prediction	0	1
0	146	469
1	597	1601
Accuracy	0.621	

Random Forest

The dataset is already clean and we have implemented classification tree and logistic regression models till now. We now try to implement Random Forest classification algorithm. This algorithm will help us predict the target variable “Churn” value. The beginning of random forest algorithm starts with selecting random features. Then these randomly selected features are used to find root node by using the best split approach. Next, the daughter nodes are calculated using the same best split approach. These steps are repeated to generate n number of random trees. These randomly created trees form the random forest.

To predict the target, the total votes are calculated for each predicted target. The highest voted predicted target is considered as the final prediction from the random forest algorithm. We implemented this Random Forest model on our dataset and obtained an accuracy of 79.09%. Following is the confusion matrix of the same:

Confusion Matrix and Statistics of Classification Tree1		
	Reference	
Prediction	0	1
0	2779	568
1	314	558
Accuracy	0.7909	

We get this confusion matrix with an Out-of-the-bag error of 20.91%

Number of trees created = 500

Number of variables tried at each split = 4

The above accuracy of 79.09% is obtained using votes of the 500 trees that are created randomly using random combination of features at every split using best split approach.

The accuracy is higher than the classification trees and logistic regression models implemented before.