Jordan Rogers

Telecom Data - Churn Analysis

In this analysis I will be using R as the main tool to help analyze the customer data for a telecommunications company who is concerned about their customer attrition. R is a great open-source statistical tool that has an expansive library that is always growing.

R runs on almost all systems and is very stable and reliable. R makes it easy to extract and load data. There are dozens of file formats from which one can choose.[1] In this analysis I will be using a .csv file, from which I will read in and extract the data using the read.csv function. This is an easy way to extract and load data into R, and is one of the many benefits of this software.[2]

The Goal

The goal of this analysis is to mitigate customer loss and reverse it by identifying indicators of customer churn and reporting back to the leadership team so they can make an informed business decision on what changes need to be made. This will be done by using a software called R that is designed to "provide a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible."[3]

The Method

Using Random forest as non-descriptive method to analyze the data is a great fit. I chose to use Random forest because it is a flexible and easy to use algorithm. Random forest does well in handling missing data while maintaining the accuracy of large proportions of the data. Because of the multitude of variables in this problem, random forest will compensate for any potential over-fitting issues that might occur.

PCA is what I will be using for a descriptive method. PCA fits well, the data meets most of the assumptions[4] needed when performing PCA:

- Sample size greater than 150, ratio 5:1 observations to features
- Variables exhibit a constant multivariate normal relationship
- No significant outliers

---

[1] "Import, Export, and Convert Data Files - CRAN." 25 Nov. 2018, https://cran.r-project.org/web/packages/rio/vignettes/rio.html. Accessed 21 Nov. 2019.
[2] "The Benefits of Using R - dummies." https://www.dummies.com/programming/r/the-benefits-of-using-r/. Accessed 21 Nov. 2019.
[3] "R: What is R? - The R Project for Statistical Computing." https://www.r-project.org/about.html. Accessed 21 Nov. 2019.
[4] Goonewardana, Harsha. "PCA: Application in Machine Learning - Apprentice Journal." Medium, 28 Feb. 2019, medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db.

- Large variance implying more structure

PCA is a great method for summarizing data. It is able to reduce the variables on which to focus while maintaining most of the variance in the data. It is a very efficient way to quickly make an educated analysis of the data set.

What does the data look like?

Dependent Variable

The target or response variable is our 'Churn' column. It is binary, yes/no, which can easily be replaced by a 0/1 data, if needed. In this case '0' and '1' are not numerical data, rather, they are representative of an assigned value or category. In our dataset, "No" equals "did not churn" and "Yes" equals "churned". Our churn data represents whether or not a customer has severed their contract and is no longer paying for services.[5] This is an incredibly important variable and being able to predict who will and will not churn to an accurate degree will allow the business to make educated decisions with high confidence. This is, as stated above, the goal of our analysis.

Independent Variable(s)

This dataset has multiple potential predictor variables aka independent variables. Any data related to the customer could influence the response variable. Most of our data is categorical, for example, are they a senior citizen? Or, do they use the streaming movie product? Yes or no? Also, the type of internet service they use and the contract type they have is also categorical.

There are some numeric and integer data types like tenure and monthly charges. Our analysis will determine which of all this data has statistical significance and if we should pay special attention to certain variables.

Data Prep

R software views certain data types such as our 'seniorcitizen' and 'churn' columns as categorical. Because of this we need to change the way the software interprets this data. We are able do this by using the factor function. This function coerces its arguments to a factor, also known as a 'category' or 'enumerated type'.[6]

---

[5] "What Is Customer Churn? [Definition] - HubSpot Blog." 6 Nov. 2018, https://blog.hubspot.com/service/what-is-customer-churn. Accessed 24 Nov. 2019.
[6] "factor function | R Documentation." https://www.rdocumentation.org/packages/base/versions/3.6.1/topics/factor. Accessed 25 Nov. 2019.

As stated above, our data categorical, ‘yes’ or ‘no’, when imported into R this data is interpreted as characters and is not friendly when using Random Forest and PCA analysis methods. Therefore the data must be changed so that R views it in a way it can interpret it in the analysis. The data is changed by using Factor(). In essence, “factors look (and often behave) like character vectors, they are actually integers under the hood, and you need to be careful when treating them like strings.”[7]

Here is how we are able to use Factor() in the code to manipulate the data. Below is a visual of the transformation of data required when using Random Forest and PCA.

```
> str(churn)
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr  "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr  "Yes" "No" "No" "No" ...
 $ Dependents      : chr  "No" "No" "No" "No" ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr  "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr  "No" "No" "No" "No" ...
 $ StreamingMovies : chr  "No" "No" "No" "No" ...
 $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr  "No" "No" "Yes" "No" ...
```

The goal here is to identify the variables that need to be converted to factors

```
#Wrangle the data, change the below columns to factors
churn$Churn <- as.factor(churn$Churn)
churn$gender <- as.factor(churn$gender)
churn$Partner <- as.factor(churn$Partner)
churn$Dependents <- as.factor(churn$Dependents)
churn$InternetService <- as.factor(churn$InternetService)
churn$Contract <- as.factor(churn$Contract)
churn$PaperlessBilling <- as.factor(churn$PaperlessBilling)
churn$PhoneService <- as.factor(churn$PhoneService)
churn$PaymentMethod <- as.factor(churn$PaymentMethod)
str(churn)
```

[7] “Understanding Factors - Programming with R.” Github.Com, 28 Sept. 2015, monashbioinformaticsplatform.github.io/2015-09-28-rbioinformatics-intro-r/01-supp-factors.html.

```
> str(churn)
'data.frame':   7032 obs. of  19 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 2 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 2 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
 $ tenure_group    : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
>
```

As can be seen here, the conversion for most of the variables that were character strings are now being view by R as factors

Data Cleansing

There is some missing data in the dataset. Before an analysis can be performed it must be determined what to do with that missing data. I am able to find the missing data using a function in R called Sapply(). It found 11 missing values in the 'TotalCharges' variable. Rather than take a median or average of the entire column and try to fit that number to these observations I thought it would be best to delete the 11 rows. 11 out of 7000+ variables would not have a significant effect on the final analysis

```
> sapply(churn, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen          Partner       Dependents          tenure     PhoneService
               0                0                0                0                0               0                0
   MultipleLines  InternetService   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV
               0                0                0                0                0               0                0
 StreamingMovies         Contract PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges            Churn
               0                0                0                0                0              11                0
>
```

```
#Remove missing values [rows]
churn <- churn[complete.cases(churn), ]
```

After running the above code, the return table shows that the missing data was deleted from the data.

```
> sapply(churn, function(x) sum(is.na(x)))
          gender    SeniorCitizen          Partner       Dependents     PhoneService    MultipleLines  InternetService
               0                0                0                0                0               0                0
  OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV  StreamingMovies         Contract
               0                0                0                0                0               0                0
PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges            Churn     tenure_group
               0                0                0                0                0               0
```

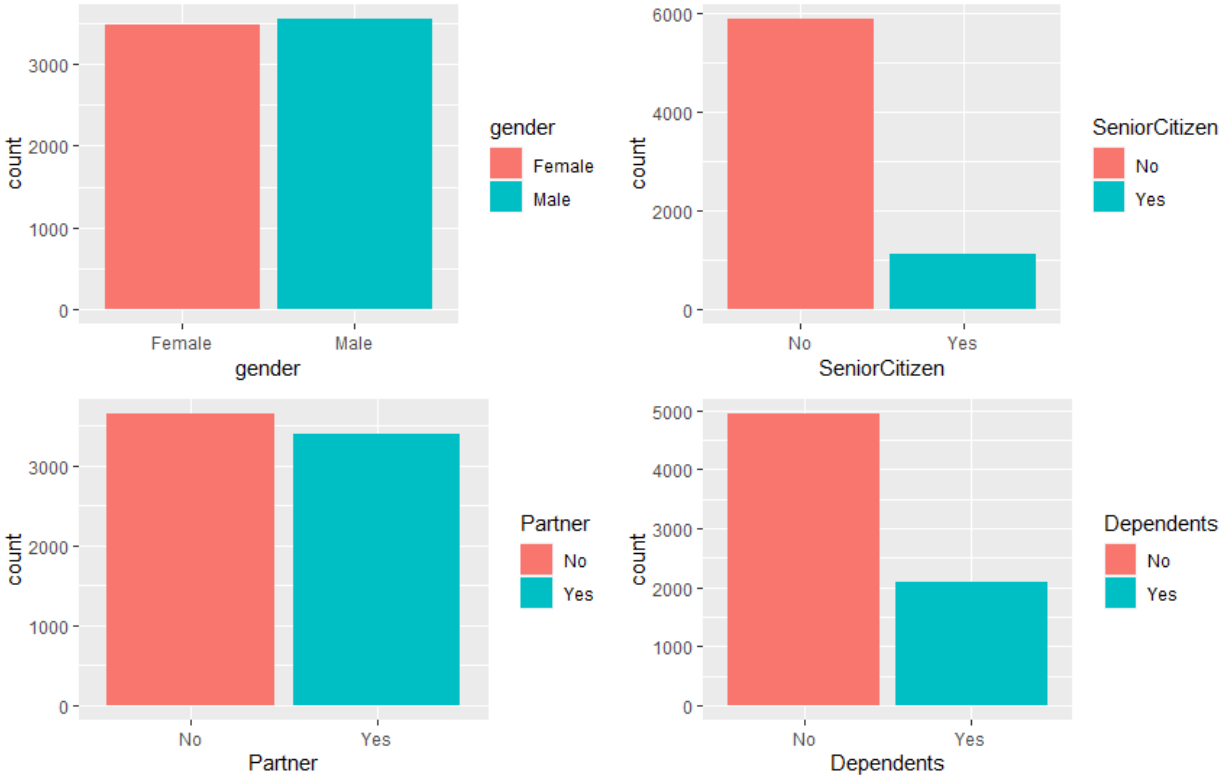
After additional analysis the variable 'Total Charges' was also removed completely. It has a high correlation with the 'Monthly Charges' variable and essentially would hurt the model.

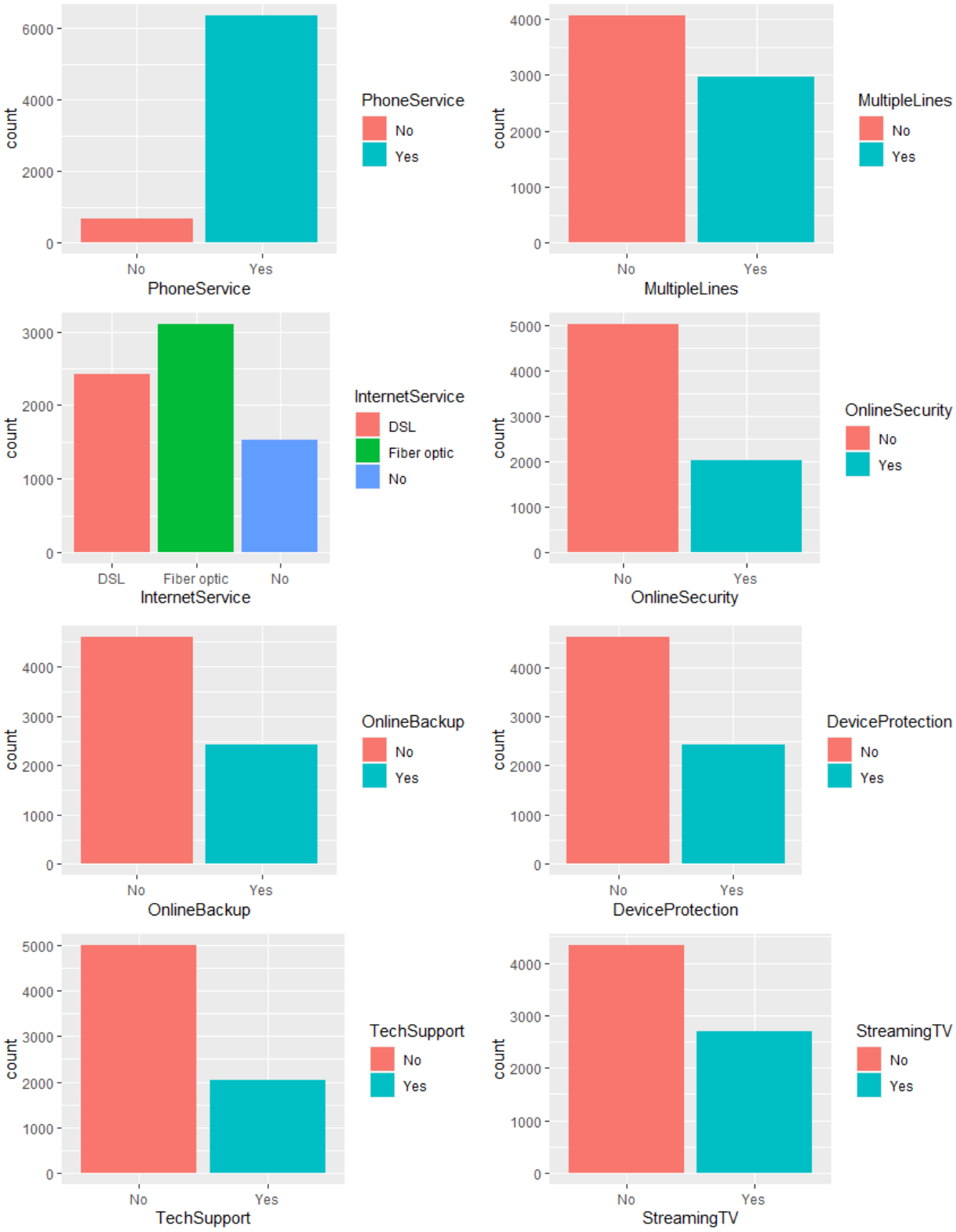
Graphing the Data

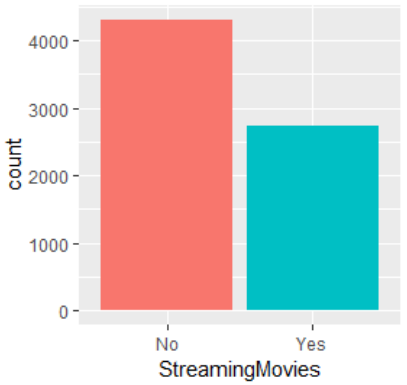
Below are 5 grids of graphs depicting the distribution of the variables using univariate statistics. There aren't any graphs with data that stands out as abnormal or extreme. The distributions between genders, seniors, family situations are all fairly normal.

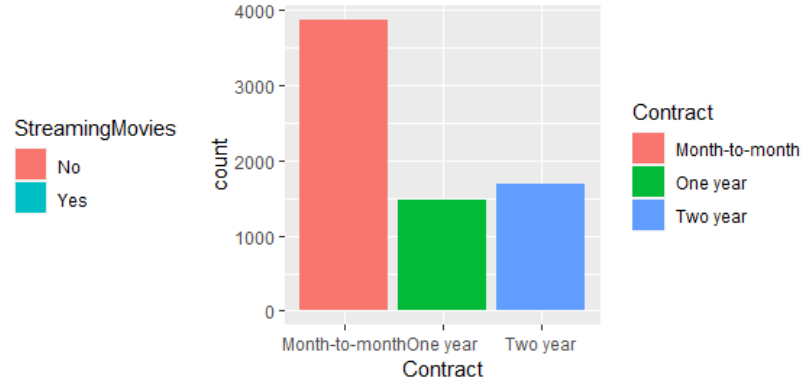
More interesting are the distributions of added services/technologies.The distribution of those who use online security, online backup, device protection, tech support, streaming tv and movies seem to be fairly similar. This leads one to believe if a customer has one of those services, they probably bundle them with most of the others.

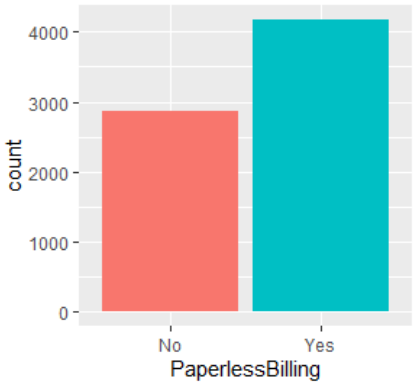
Lastly, but not as relevant to this analysis, I found it incredibly interesting that the total # of seniors was less than the total # of accounts that use a "mailed check" as their form of payment. That means there are a lot of non-seniors writing checks and mailing them to pay their bills.



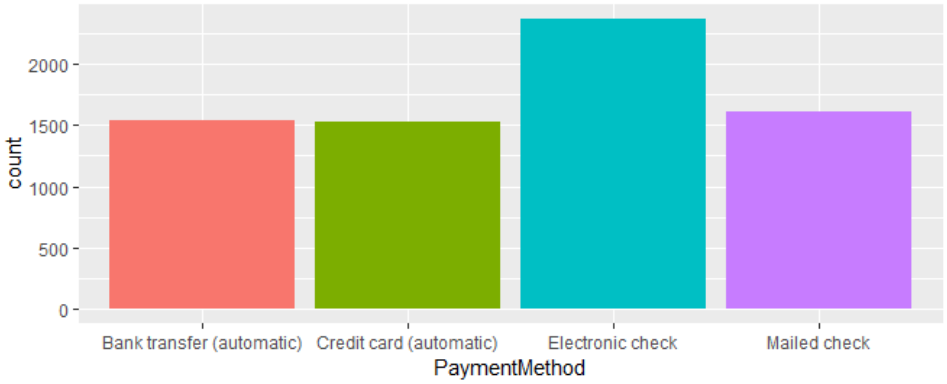

count
6000
4000
2000
0
No
Yes
PhoneService
PhoneService
No
Yes
count
4000
3000
2000
1000
0
No
Yes
MultipleLines
MultipleLines
No
Yes
count
3000
2000
1000
0
DSL
Fiber optic
No
InternetService
InternetService
DSL
Fiber optic
No
count
5000
4000
3000
2000
1000
0
No
Yes
OnlineSecurity
OnlineSecurity
No
Yes
count
4000
3000
2000
1000
0
No
Yes
OnlineBackup
OnlineBackup
No
Yes
count
4000
3000
2000
1000
0
No
Yes
DeviceProtection
DeviceProtection
No
Yes
count
5000
4000
3000
2000
1000
0
No
Yes
TechSupport
TechSupport
No
Yes
count
4000
3000
2000
1000
0
No
Yes
StreamingTV
StreamingTV
No
Yes

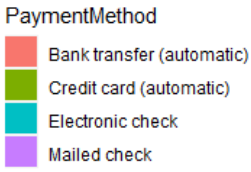

4000
3000
2000
1000
0
count
No
Yes
StreamingMovies

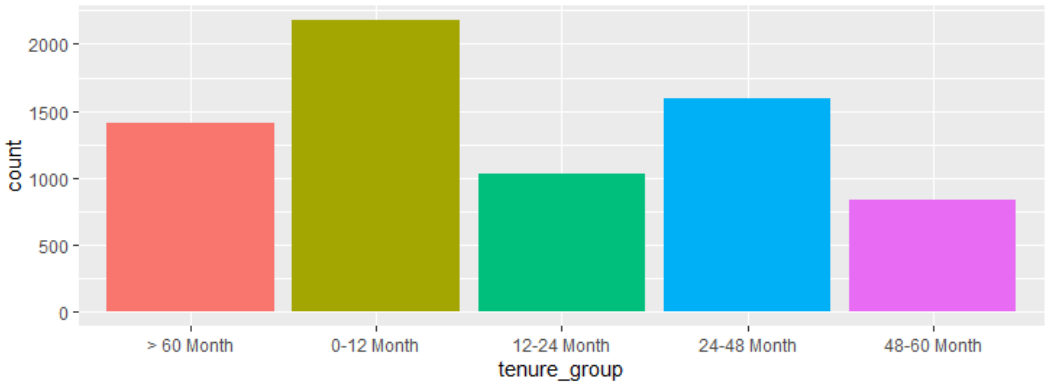

StreamingMovies
No
Yes
4000
3000
2000
1000
0
count
Month-to-month
One year
Two year
Contract
Contract
Month-to-month
One year
Two year

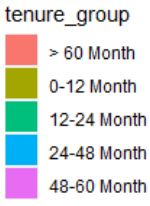

4000
3000
2000
1000
0
count
No
Yes
PaperlessBilling


PaperlessBilling
No
Yes


2000
1500
1000
500
0
count
Bank transfer (automatic)
Credit card (automatic)
Electronic check
Mailed check
PaymentMethod


PaymentMethod
Bank transfer (automatic)
Credit card (automatic)
Electronic check
Mailed check

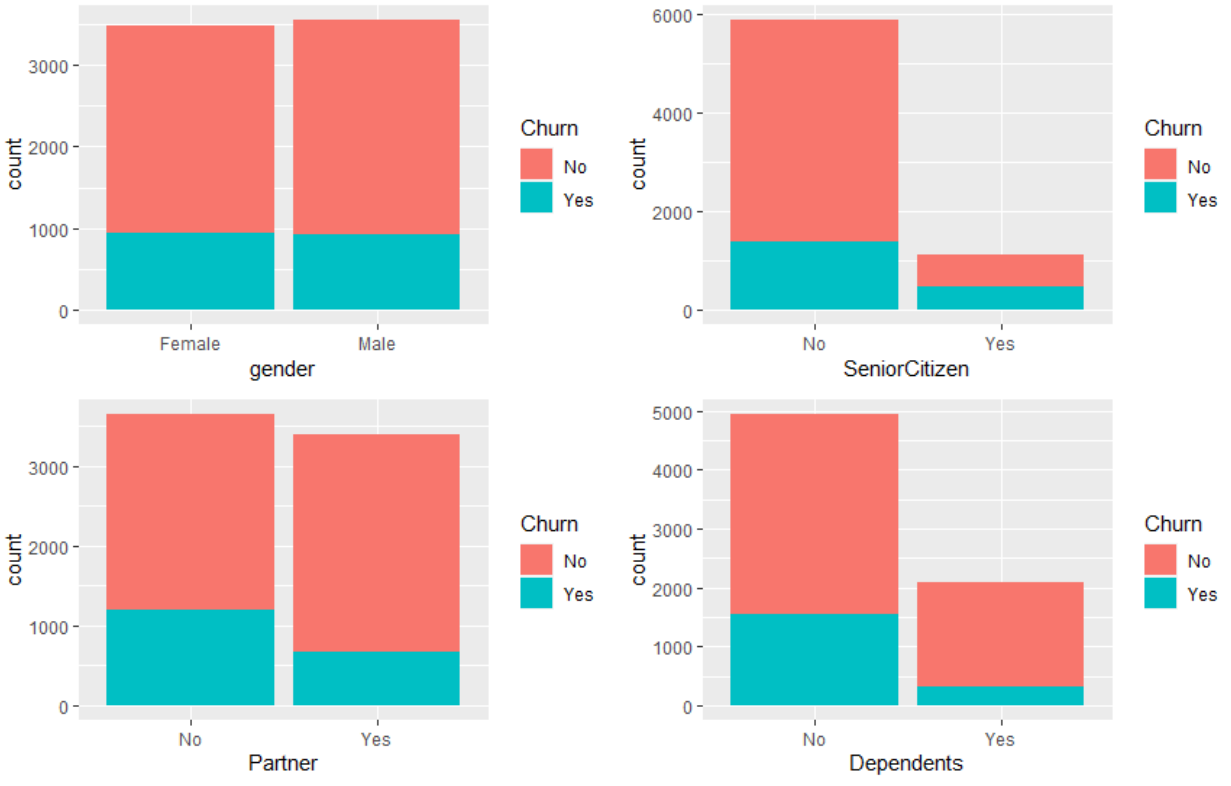

2000
1500
1000
500
0
count
> 60 Month
0-12 Month
12-24 Month
24-48 Month
48-60 Month
tenure_group

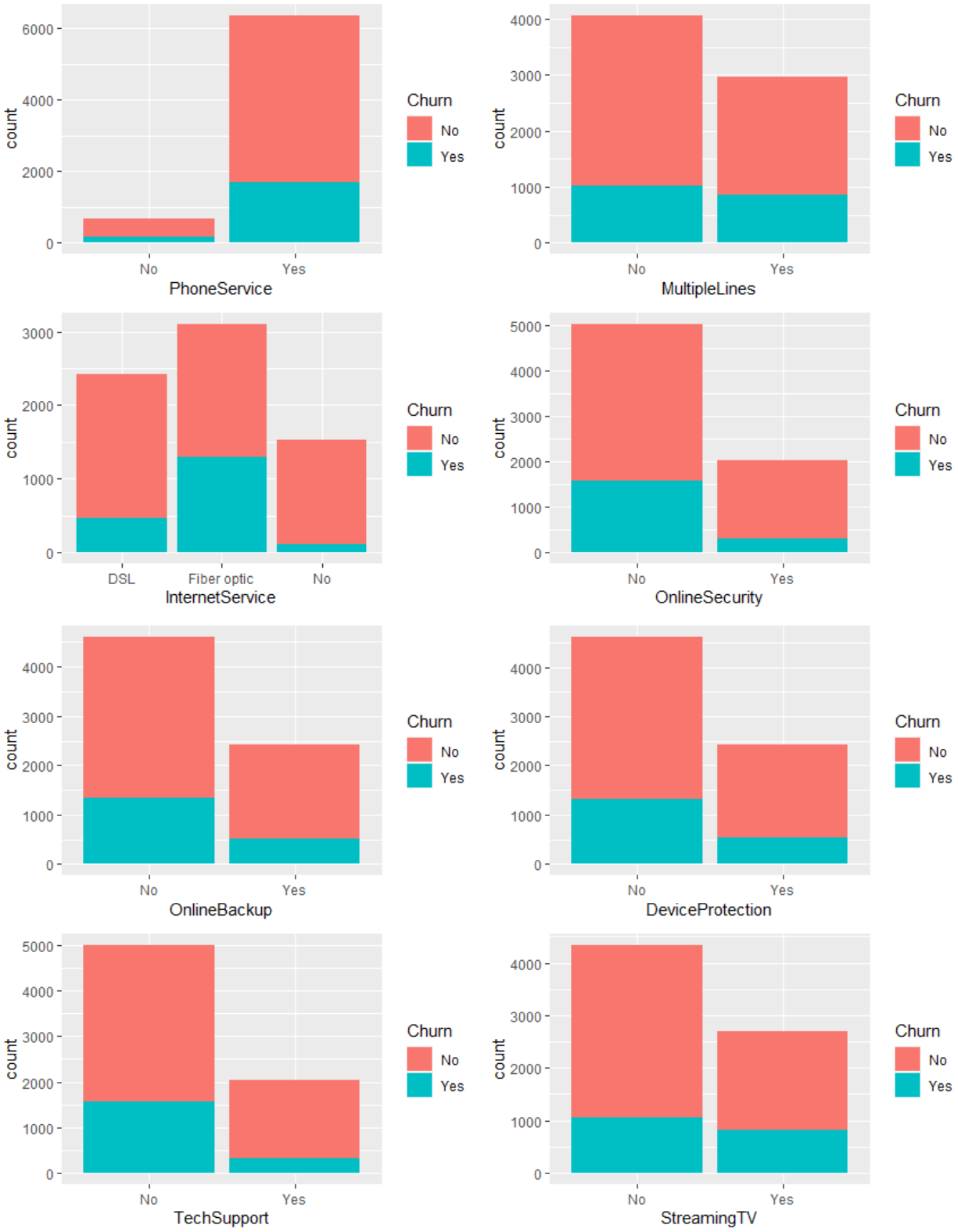

tenure_group
> 60 Month
0-12 Month
12-24 Month
24-48 Month
48-60 Month

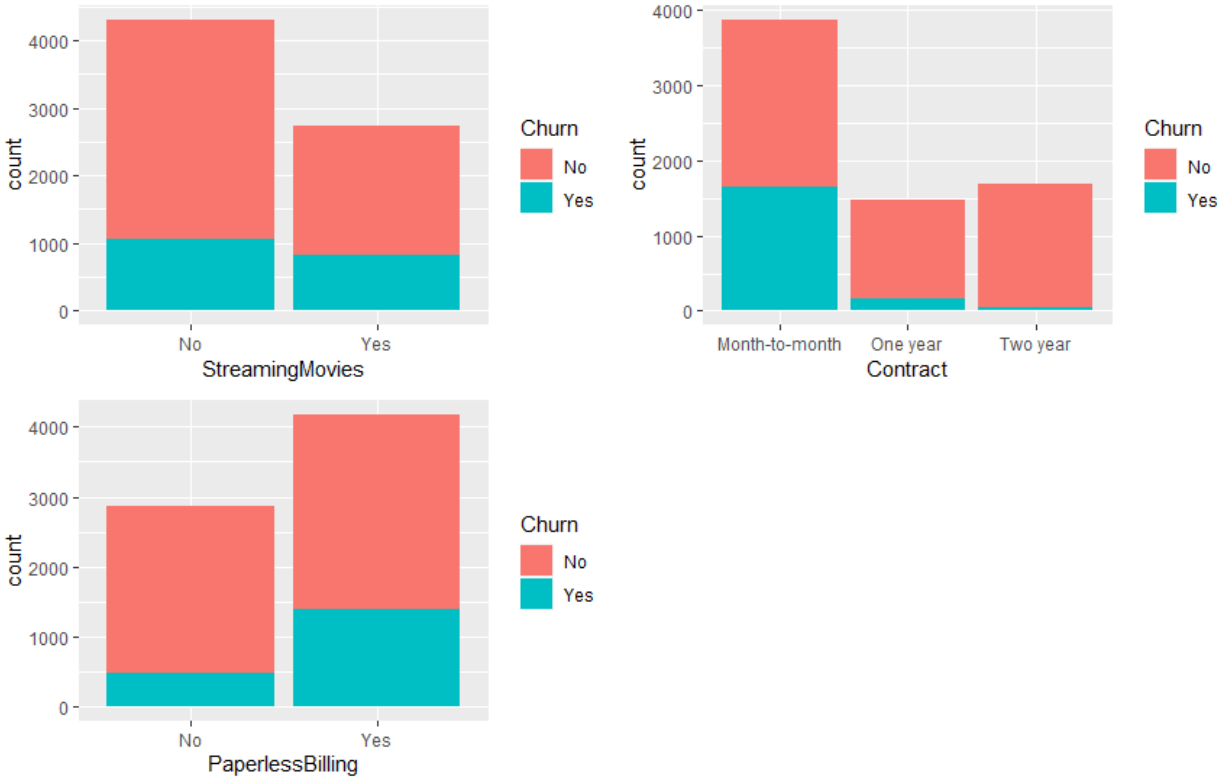
The bivariate statistical analysis shows the data of each variable vs our **Churn** variable. Viewing the data in the manner allows us to see some correlation between certain independent variables and our dependent variable.

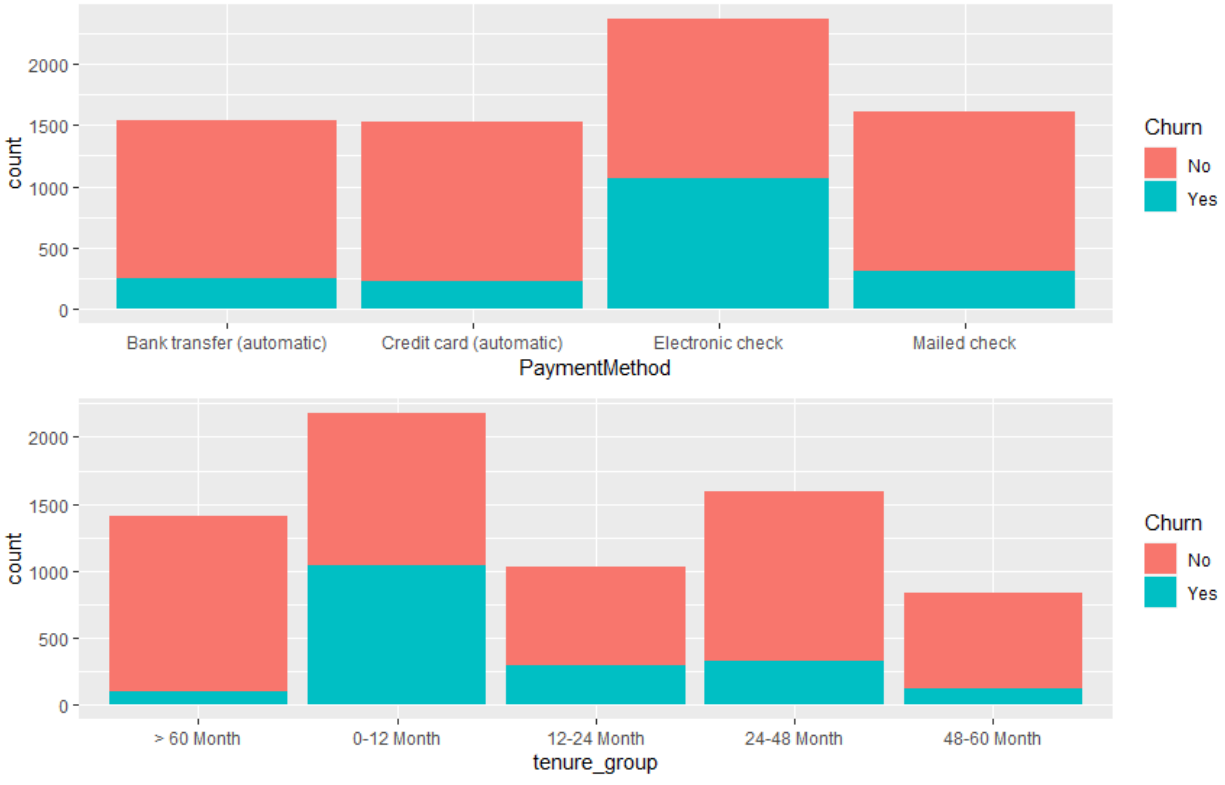
For example, there are two rather interesting graphs that draw my eye. **Contract** and **Tenure Group.** Contract refers to the type of contract the customer has. From our univariate graphs we can easily see that most customers are on a month to month contract. Here in our bivariate graphs we are able to see the effect on Churn that having the month to month versus a longer contract type. Although there are far fewer customers with one and two year contracts, their churn rate is significantly reduced.

Related to this data is the Tenure Group data. While looking at this data we can deduce that the longer a customer stays the less likely they are to churn. This also means, although many customers are on the month to month contract, if they hit the one-year mark as a customer their likelihood of churning significantly decreases.




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





Applying the Methods

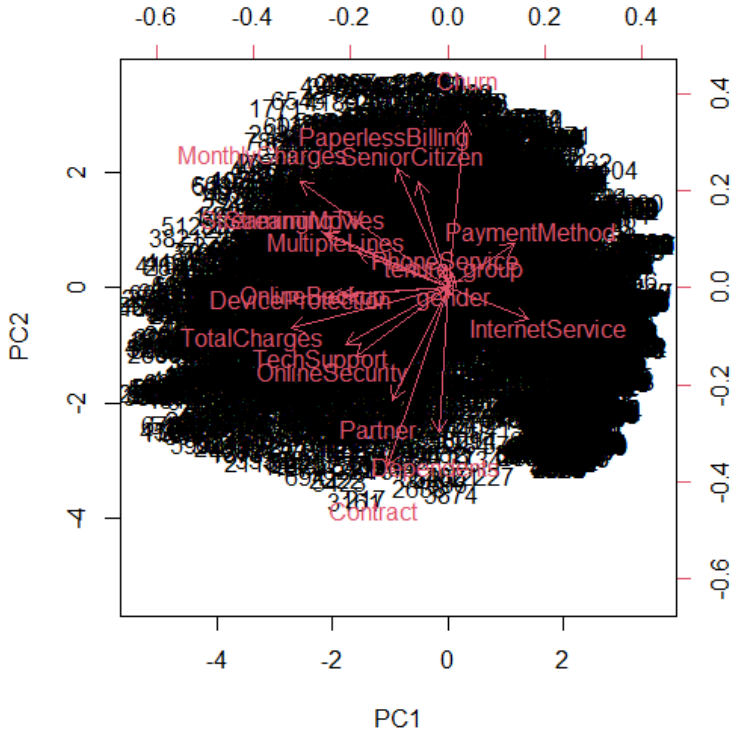
The summary of the Principal components tells us that the first 11 components explain 80% variance and hence we select the first 11 PCs.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     2.140 1.5093 1.32778 1.07129 1.01720 1.00119 0.97891 0.95251 0.92237 0.87271 0.83839 0.79509
Proportion of Variance 0.229 0.1139 0.08815 0.05738 0.05173 0.05012 0.04791 0.04536 0.04254 0.03808 0.03515 0.03161
Cumulative Proportion  0.229 0.3429 0.43107 0.48845 0.54019 0.59030 0.63822 0.68358 0.72612 0.76420 0.79935 0.83096
                          PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20
Standard deviation     0.78372 0.77664 0.74912 0.68794 0.67637 0.58868 0.45740 0.34033
Proportion of Variance 0.03071 0.03016 0.02806 0.02366 0.02287 0.01733 0.01046 0.00579
Cumulative Proportion  0.86167 0.89182 0.91988 0.94355 0.96642 0.98375 0.99421 1.00000
>
```

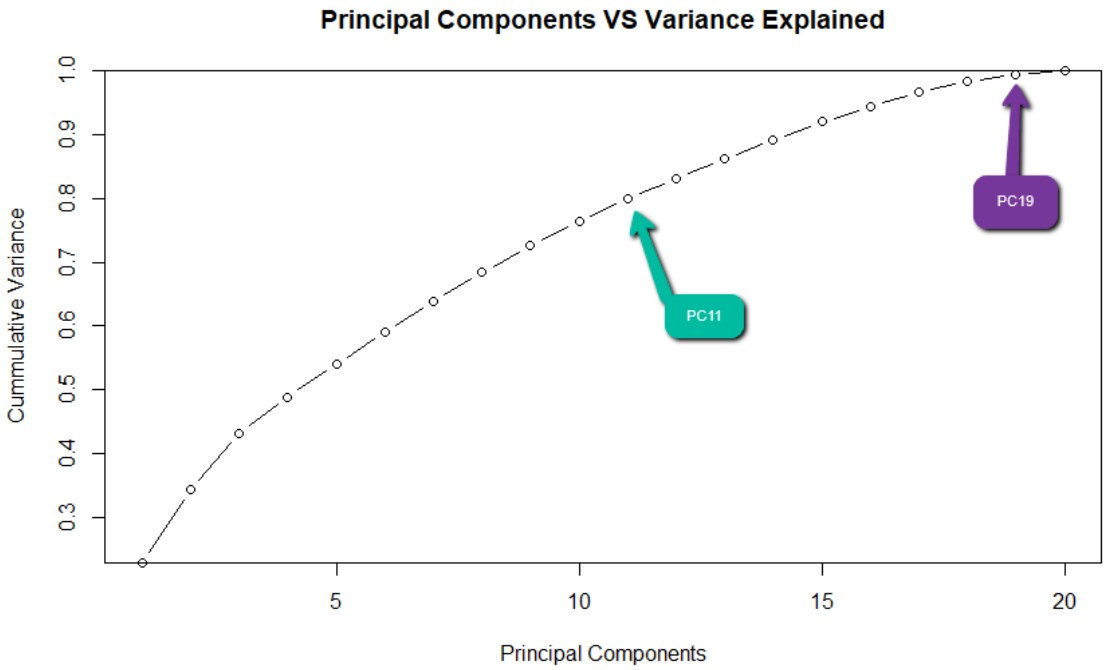
We feed these 11 PCs to the RPART model to predict whether the customer will be churned or not.

The biplot suggests the directions of all the variables. It states that **Dependent, Contract, Partner** etc convey the same data value and explain the same variances as these vectors are in the same direction. They are also negatively correlated to **Churn**.

On the other end of the biplot we can see that **PaperlessBilling SeniorCitizen, MonthlyCharges,** etc have a positive correlation with one another and **Churn**.



The Cumulative variance plot states that the first 11 PCs explain almost 80% variance and PC's after 19 are redundant and should be dropped as 19 PC's explain ~ 100% of the variance.



Our PCA confusion matrix lacked in **Accuracy** with a total of **0.6833** compared to our Random Forest model described below. However, the **Sensitivity** is **0.6310** and the **Specificity** is **0.8218** showing some promise.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 966 103
         1 565 475

               Accuracy : 0.6833
                 95% CI : (0.6629, 0.7031)
    No Information Rate : 0.7259
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3626
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6310
            Specificity : 0.8218
         Pos Pred Value : 0.9036
         Neg Pred Value : 0.4567
             Prevalence : 0.7259
         Detection Rate : 0.4580
   Detection Prevalence : 0.5069
      Balanced Accuracy : 0.7264

       'Positive' Class : 0
```

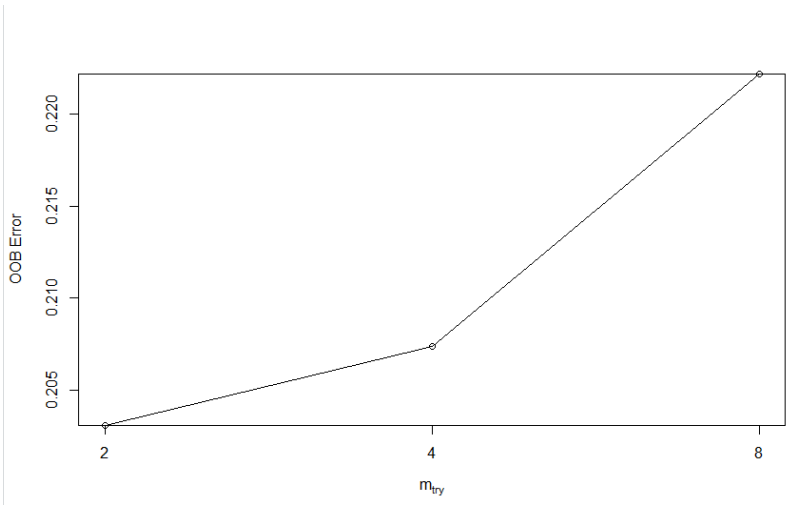
From the confusion matrix from our Random Forest we find that the **Accuracy** of the model is **0.7932**, **Sensitivity** is **0.9167**, & **Specificity** is 0.**4518**

The accuracy and sensitivity numbers are good and expected. The specificity value is much lower than expected, but can most likely be attributed to the imbalance of our dependent variable data (5163 = No vs 1869 = Yes).

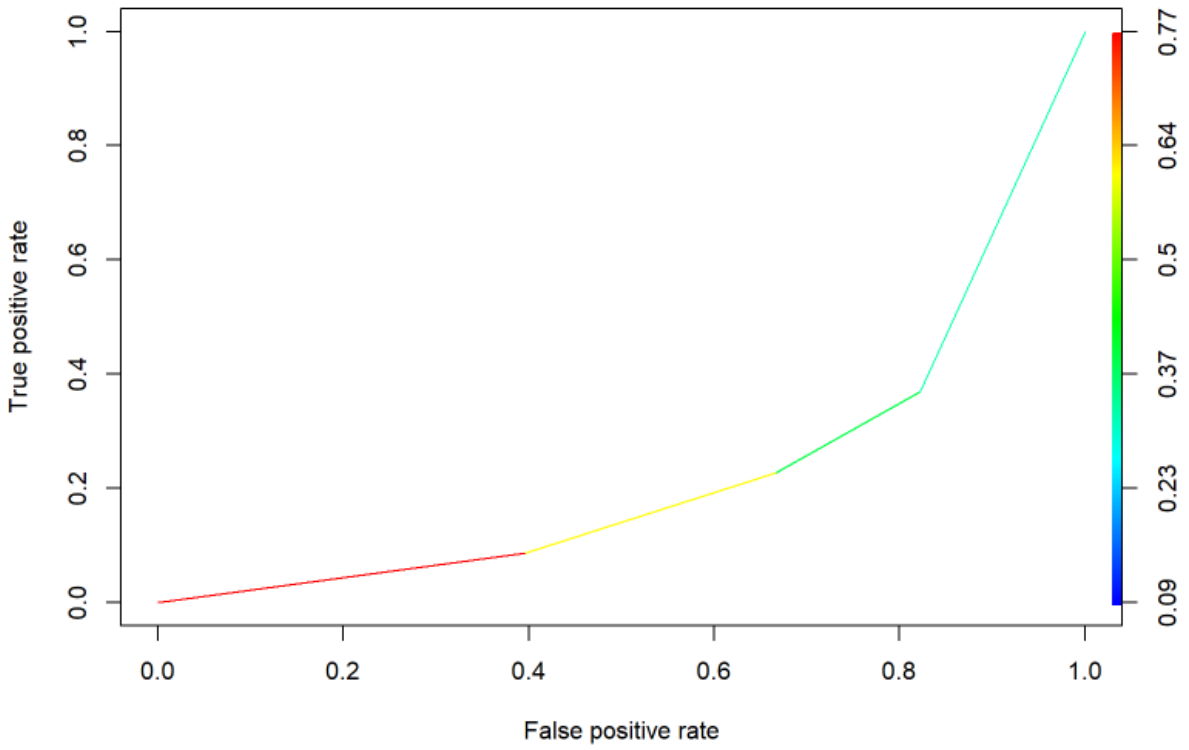
```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1419  307
       Yes  129  253

               Accuracy : 0.7932
                 95% CI : (0.7752, 0.8103)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 2.074e-10

                  Kappa : 0.41

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

            Sensitivity : 0.9167
            Specificity : 0.4518
         Pos Pred Value : 0.8221
         Neg Pred Value : 0.6623
             Prevalence : 0.7343
         Detection Rate : 0.6731
   Detection Prevalence : 0.8188
      Balanced Accuracy : 0.6842

       'Positive' Class : No
```

Justify the Analysis Methods

Random forest with an accuracy of 80% outperformed the other models I looked at, including PCA which had an accuracy of more than 10 points lower. Having used Random Forest in our model allows us to say with confidence that the discovery is due to the robust decisions made by this algorithm when looking at the data. Especially since we were able to increase our accuracy through our ‘ntree’ and ‘mtry’ until we saw the most accurate predictions from our data.



PCA had much higher sensitivity and specificity values than other models. These numbers justify our using this approach when analyzing the data. Below represents our True Positive v False Positive rates directly influencing our high sensitivity and specificity values.

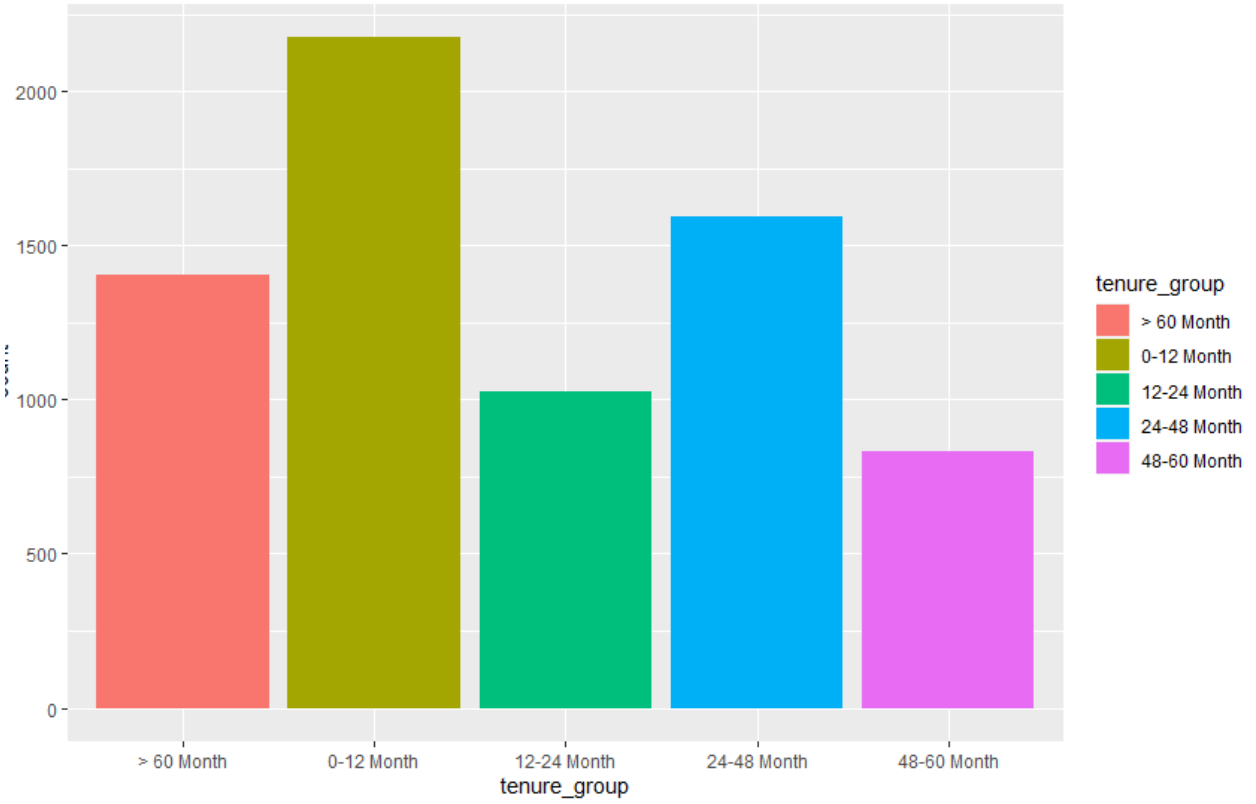


Justify the Visualization Methods

R has several packages for graphing and visually representing data. I have chosen to visualize the data by using a package called ggplot2. This is an especially powerful tool within R because you can use functions within ggplot2 that allow you to make customization to the visuals. For example, the theme() function allows for 3 main types of components[8]:

- Axis: controls the title, label, line and ticks
- Background: controls the background color and the major and minor grid lines
- Legend: controls position, text, symbols and more.

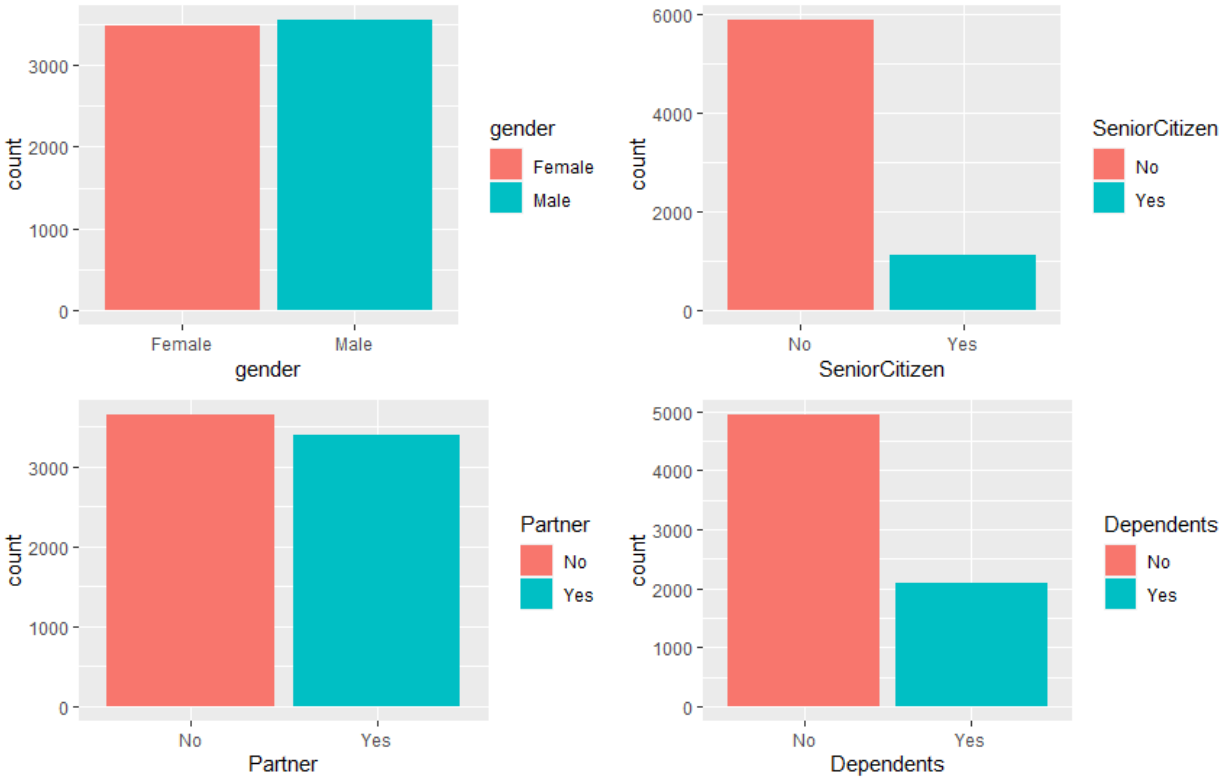
Below is an example of a single graph (displayed previously) that shows some of the above aforementioned capabilities of ggplot2.



[8] "Ggplot2." The R Graph Gallery, 2018, www.r-graph-gallery.com/ggplot2-package.html.

Explain Discrimination and Phenomenons

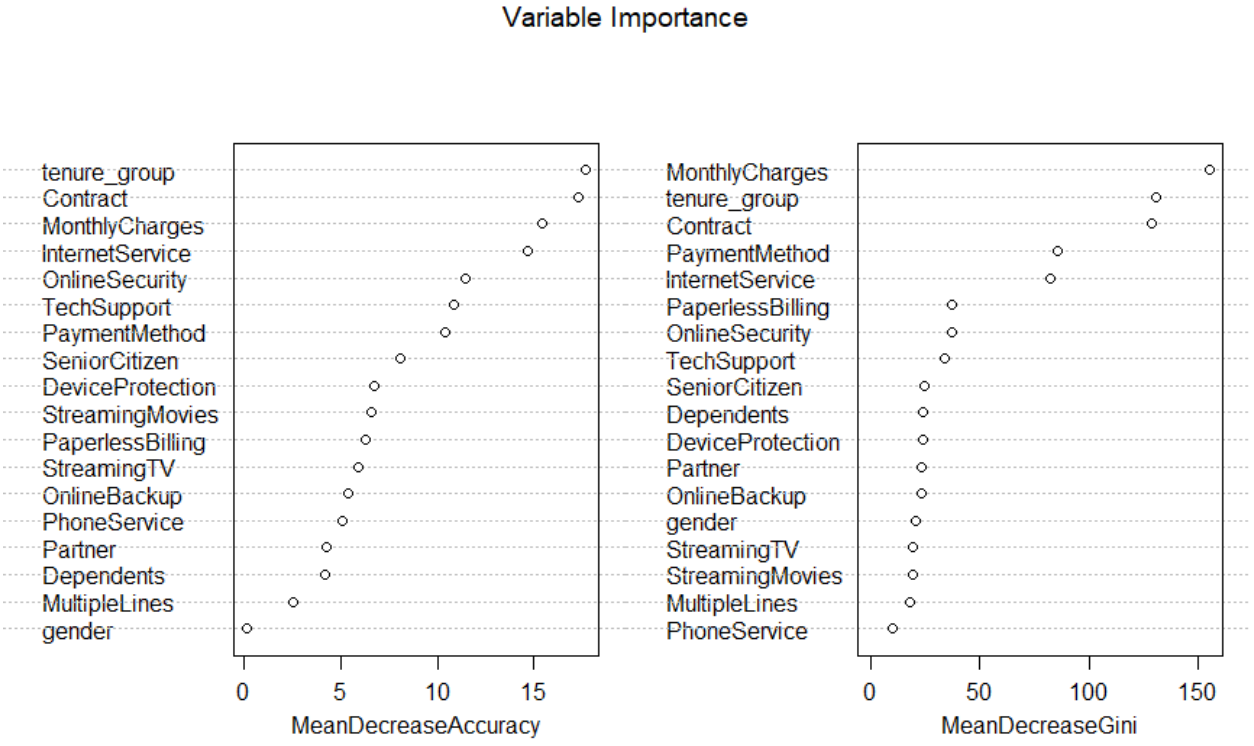
As it pertains to discrimination in the data, this grid I share below speaks volumes when combined with the “Variable Importance” graph from my analysis.



Most of the data is related to how the customer interacts with the company, the grid above specifically relates to the customers themselves. Gender, age group, family situation… These are all potential risks of discrimination in the data. None of these attributes rank highly on the most important features in this model and I have concluded that there is no evidence to point to any discrimination.

In the analysis of the graph data I suspected that Tenure Group and Contract would have significant roles in whether or not someone would churn. Based on our biplot the Tenure Group seemed not as relevant but as seen above it listed as one of the most important or influential attributes in determining churn.

Additionally, the biplot showed that Contract was highly correlated to churn, what I didn’t expect is that it would be negatively correlated. I expect that the month to month contract, although beneficial to the customer, is detrimental to the attrition of the company.



Description of Data Usage

In reviewing my data I found that very little of the data would not be useful. As for ranking the importance of that data, as seen in the "Variable Importance" graph. This is a function within Random Forest that allows the variables to be measured by importance and plotted on a dotchart[9]. I was able to customize the chart to include all applicable variables as well as add titles, etc to the dotchart.

Conclusion

With the models in place accurate predictions can now be made as they are updated on a scheduled cadence with new/updated data. The company can now, with strong confidence, take action to mitigate the customer churn.

[9] "VarImpPlot Function | R Documentation." R Documentation, 2020, www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/varImpPlot.