**Data Mining Report: Customer Churn**

Heidi L. Petersen

College of Information Technology, Western Governors University

C744: Data Mining and Analytics II

Dr. Daniel Smith

March 7, 2021

The main goals of this analysis will be to discover reasons customers are leaving the landline company for cable and find specific indicators whether a customer is more likely to leave in the future. Churn is the target variable within this dataset. Its class is character. In this case, “yes” or “no”.

I chose R to process this data. R is a good resource for statistical analysis. It has the features needed to achieve the goals of this particular analysis quickly and efficiently. I mostly chose R for its ease. SAS would have been my choice if the data set had been much larger and I needed to worry about RAM. Python is a more of a generalized program where R is specifically designed for statistics.

**Cleaning the Data Set:**

I imported the dataset as a csv file into R. I then ran a summary on the data to get an idea of the variable types and get an overview of the data. There are 7043 observations and 21 features to begin with. I used the sapply command to find any empty cells. The only variable with missing data was the TotalCharges column. MonthlyCharges and Tenure already capture all the information needed in Total Charges so I deleted the entire column. I also removed the column CustomerID because it was not relevant to the analysis. SeniorCitizen is classified as 0 or 1. It would be more uniform and make more sense to change to factors of “yes” and “no” so I made that change. MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTv, StreamingMovies all had values of “No Internet Service” or “No Phone Service”. To simplify those columns, I changed all of those entries to “No” instead. After that, I also converted Gender, Partner, Dependents, PhoneService, InternetService, Contract, PaperlessBilling, PaymentMethod, and Churn to factors as well. I wanted to be sure that each observation had similar classification.

**Descriptive Analysis:**

When considering a descriptive method, I ruled out primary component analysis because it lends itself better to numerical data. Most of the data for this analysis is categorical. In the end, I decided to use factor analysis. It is suited to categorical data and would give me a detailed view of the data.

It is imperative to understand the data set that I am working with to be able to do a detailed analysis. The descriptive analysis of this data was comprised of obtaining basic statistics, creating histograms, and making crosstables to be able to visualize and dissect the data. Histograms are a simple way to visualize each variable where there were not many choices for each variable. The histogram makes it simple to see the proportions in each category. The crosstables allowed me to see how often each option occurred with churn and also gave the correlation of each variable with my target variable.

For a univariate analysis, I again summarized the data again with the changes after cleaning. This gave me the descriptive statistics for my numeric variables and the frequency of the factors in my categorical variables.

A screenshot of a computer

Description automatically generated with low confidence

The summary gave me a general impression of each of the variables. It provides the mean for the numeric variables. To better visualize the information contained in the summary, I created histograms of the information. It makes it easier to see the distribution and proportions of each of the factors.

The breakdown for gender was almost equal. A majority of the customers are not senior citizens. The customers are almost split equally between those who have a partner and those who do not with slightly more not having a partner.

Square

Description automatically generated with medium confidenceChart

Description automatically generatedShape, square

Description automatically generated

About twice as many customers did not have dependents compared to those who did. Tenure was skewed left. The mean is 32.37 and median is 29.

Chart, shape

Description automatically generatedChart, histogram

Description automatically generated

Most customers subscribe to phone service and they are about evenly split on having multiple lines. Nearly all customers have internet service where the most common is fiber optic.

Chart

Description automatically generatedChart, shape, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Customers were not signed up for online security, online backup, device protection, tech support, and streaming compared to those who were signed up for those services.

Chart, bar chart

Description automatically generatedChart, shape, bar chart

Description automatically generatedChart, shape, square

Description automatically generatedChart, shape

Description automatically generatedChart, shape, bar chart, square

Description automatically generatedChart, shape, square

Description automatically generated

The most common contract length was a month-to-month agreement. Almost twice as many customers were month-to-month than each of the other two options. Paperless billing was the more popular choice with customers as well.

Chart, bar chart

Description automatically generatedChart

Description automatically generated

The most common way to pay was with electronic check with the other three options about equal in popularity. Monthly charges ranged from $18.25 to $119.75. The mean was $64.76 and median of $70.35. About 26.5% (1869/7043) of customers left the company.

Chart

Description automatically generatedChart, histogram

Description automatically generatedShape

Description automatically generated

After I had examined the variables on their own, I needed to see how they interact with the target variable, churn. I created a crosstable with each variable against churn to understand the breakdown and significance of each of the variables. This combined with the summary, gives me a good description of the data. The tables easily enable me to see the breakdown within each variable related to churn. I also graphed the results as bar graphs to better see the connection between each variable and my target variable.

Contract:

Table

Description automatically generated Chart, bar chart

Description automatically generated

Tenure:

A picture containing table

Description automatically generatedChart, bar chart

Description automatically generated

Internet Service:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Paperless Billing:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Online Security:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Tech Support:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Dependents:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Payment Method:

Table

Description automatically generated



Chart, bar chart

Description automatically generated

Senior Citizen:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Partner:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Monthly Charges:

A picture containing text

Description automatically generated

Online Backup:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Device Protection:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Streaming TV:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Streaming Movies:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Multiple Lines:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Phone Service:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Gender:

Table

Description automatically generatedChart, bar chart

Description automatically generated

Factor Analysis on Variables:

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

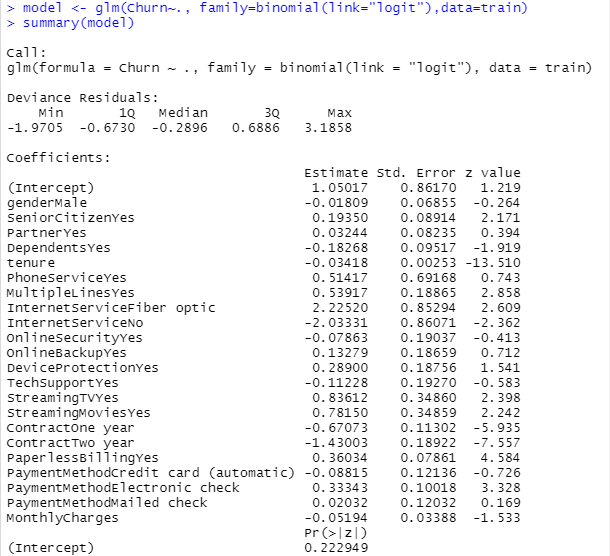
Description automatically generated

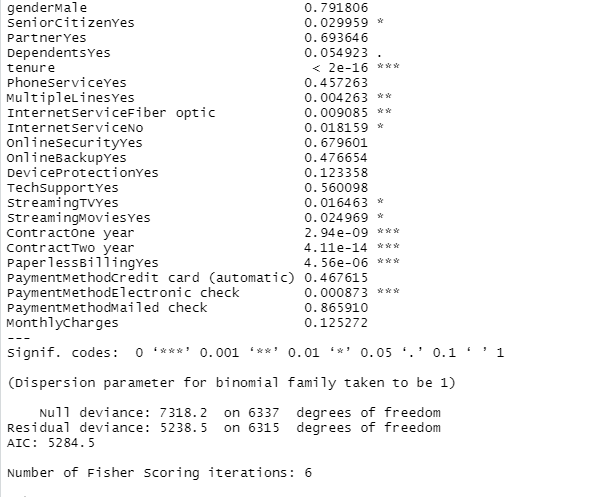
From these results, I could see that Contract, Tenure, and InternetService were the variables with the most impact on churn. Gender and PhoneService did not have a significant influence on churn.

**Predictive Analysis**

For the predictive method, I chose to use logistic regression. This method is especially good in a case where we have a binary outcome. In this case, whether the customer will leave the telecommunications company.

I separated the data into a training set (6338 observations) and a test set (705 observations). Next I created the logistic regression mode using the training datal:





This indicates that customers with a one or two year contract, and no internet service are less likely to churn. Customers who have multiple lines, use fiber optic internet and pay by electronic check are more likely to churn.

I evaluated the logistic regression model’s efficacy. First, I changed “No” and “Yes” to “0” and “1” respectively. I ran the test data through the model. At this point, my fitted.results had probabilities of each observation leaving the company, so I needed to change the results to either a “0” or “1”. If the probability was above 0.5, then I changed the result to 1. Otherwise, it became 0. The next step was to check how the predications compared to the actual data. I found that the accuracy of my model is 80%.

Text

Description automatically generated