Data Mining & Analysis II - C744

Student Name: David Bailey

Student ID: #000658614

Date: 06/11/2021

Student Mentor Name: Kelly Smith

# Objectives and Initial Plans

The goal of this activity is to prepare, clean, and mine historical customer data to gain a solid understanding of:

1. The distribution of important variables for previous and current clients.
2. The variables that are most useful in predicting customer churn.
3. How accurately we can predict churn in current and future clients.

With a solid understanding of these statistical assumptions, more effective decisions can be made in operations, marketing, and sales to reduce churn and promote the packages and products that result in a higher likelihood of long-term customer loyalty.

The program R will be used to import, clean, and run statistical analysis on the customer churn data set. R is an open-source statistical language that has thousands of default and installable packages that make manipulating matrices and data frames, running code loops and statistical calculations, and creating graphics very simple. It also won’t add any cost to the company since it is open-source and available free of charge.

A quick glance at the initial data set indicates that there are many categorical variables. Because of this, multiple correspondence analysis will be used in place of principal component analysis, which would require replacing the categorical values with numeric values. The multiple correspondence analysis helps confirm which variables are most useful in predicting churn.

Because the target variable of customer churn is binary, I will be using logistic regression as the nondescriptive method. Logistic regression will allow a second look at the most influencing variables, which will be useful to compare with the results of the MCA. This regression paired with a receiver operating characteristic curve will show how accurately the model can predict which cases will churn.

# Initial Data Exploration & Wrangling

After cleaning, the data set has 7032 customer records, or observations, and 20 variables. The target variable is churn, which is dichotomous or binary data. There are two factors for the target variable, Yes and No. This indicates whether the customer had cancelled their landline service or not. This variable has been changed to the numerical values 1 and 0 for Yes and No, respectively.

The data set included categorical variables such as gender, payment method, and contract type. Contract type, for example, lists whether each customer is on a one-year, two-year, or month-to-month contract.

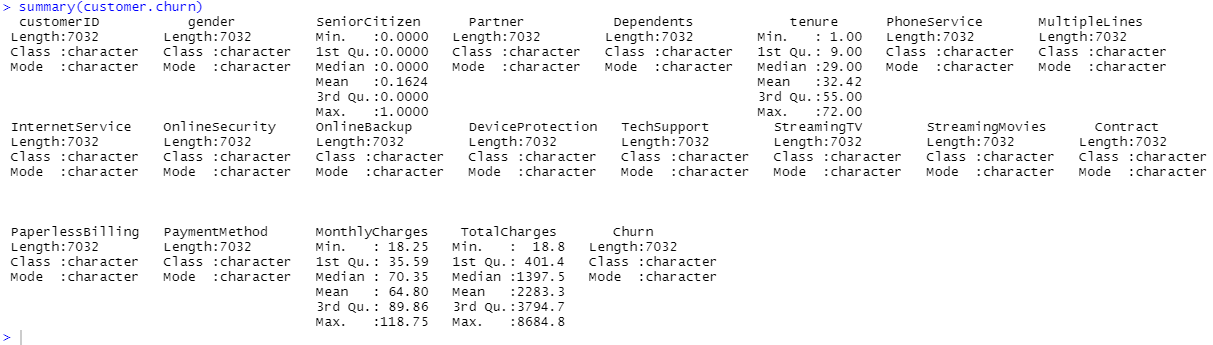
After importing the full data set into R and viewing a summary of the variables and observations, it could be seen that there were some missing values, some variables that wouldn’t be needed, and some variables that needed values to be standardized. This will help allow for the use of statistical software, such as R, for analysis. Missing values, disorganized factors, and non-standardized factor levels make accurate analysis more difficult by placing data in incorrect bins and being difficult to program around. To meet the objectives quicker and have more straight-forward and clear insights afterwards, the data set will need to be cleaned.

I removed the Customer ID variable from the table, since those ID numbers wouldn’t be useful in the analysis. I then used R to delete 11 observations that didn’t have an entry for the Total Charges variable. I also used the Yes/No factor labels for the Senior Citizen variable, and reassigned the values “No phone service” and “No Internet Service” to match the rest of the observations with the value “No.” Lastly, I changed the Tenure feature from a monthly integer to a factor with the following six labels:

|  |  |
| --- | --- |
| * First Year | * Fourth Year |
| * Second Year | * Fifth Year |
| * Third Year | * Sixth Year or More |

This will help narrow down the time frame that a customer is more likely to churn to a smaller number of periods, instead of a period for each month where the maximum value was 72.

To get a quick summary of the statistical values of the entire data set, the summary function is especially helpful. Below are those values.

Summary Statistics

This summary shows important information for the numeric variables. Tenure has a maximum of 72 months, minimum of 1 month, and mean of 32.42 months. This variable will be converted to 6 categorical bins of years. I can also see the minimum of 18.25, maximum of 118.75, and mean of 64.80 for monthly charges.

# Initial Analysis

After cleaning the data set, a graphical representation of each of the variables gives a better understand of the variance and values for each feature. Bar graphs for each categorical variable and histograms for the two continuous variables are below.

Univariate Distributions of Categorical Variables*A screenshot of a computer

Description automatically generated with medium confidence*

*Chart, bar chart

Description automatically generated*

*Chart, bar chart

Description automatically generated*

Relative frequency is shown for each of the variables, which allows for a quick view at how often each factor occurs in relation to the others. With axis labels on the left it is easier to see a more exact percentage than a pie chart or histogram. For the quantitative variables, I used histograms, which help to visually see measures of central tendency and skewness, which are important for continuous data.

Univariate Histograms of Continuous Variables

*Chart, histogram

Description automatically generated*

Both continuous variables are right-skewed. It would also seem common for monthly charges to be correlated with total charges, so a correlation matrix is shown below.

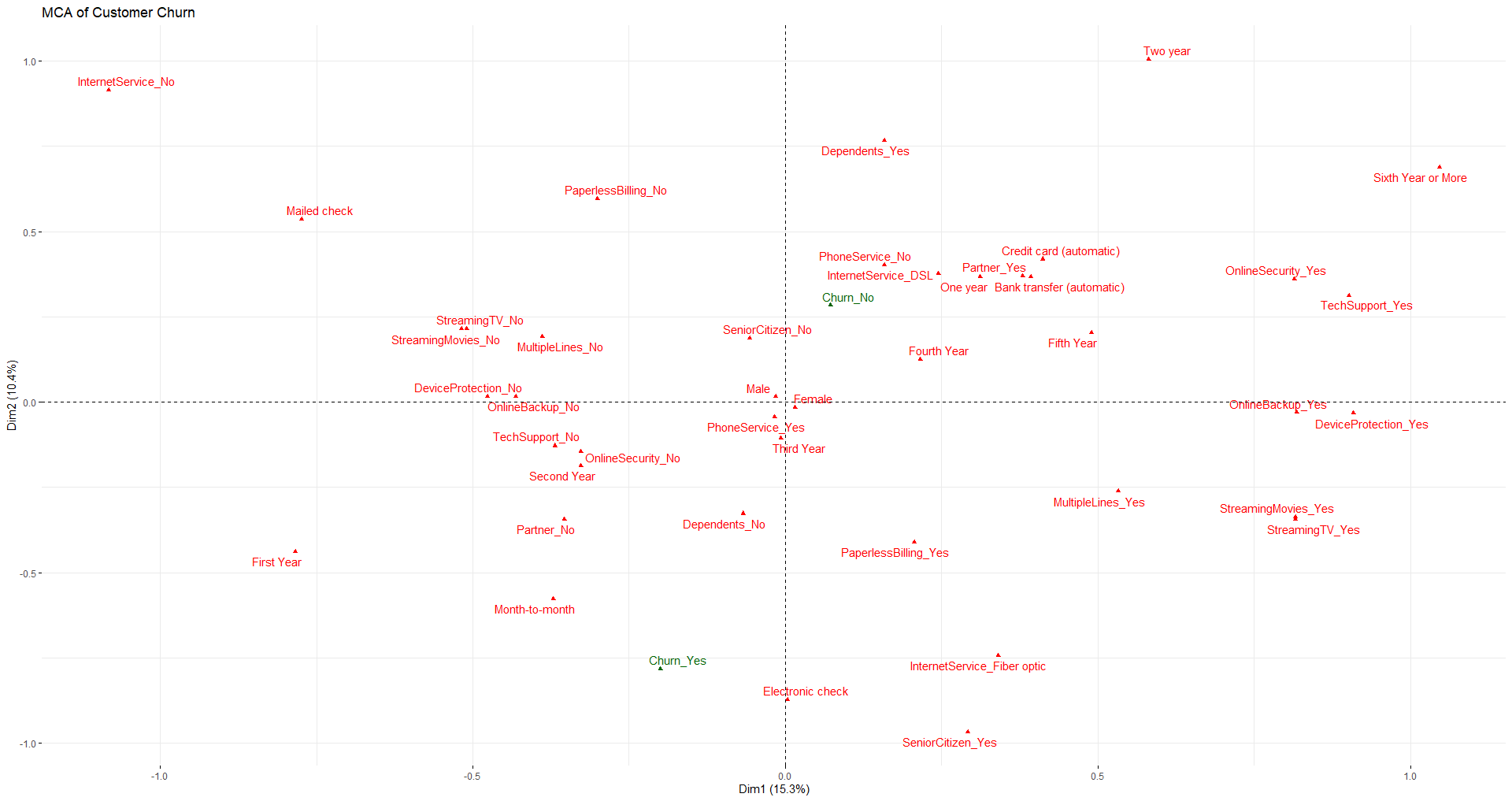
Bivariate Correlation Matrix – Continuous Variables

*Chart, box and whisker chart

Description automatically generated*

The two variables are positively correlated, so Total Charges will be removed to avoid collinearity between predictor variables. A correlation matrix is a great way to see correlations between variables quickly and can involve many more than two predictors, when needed.

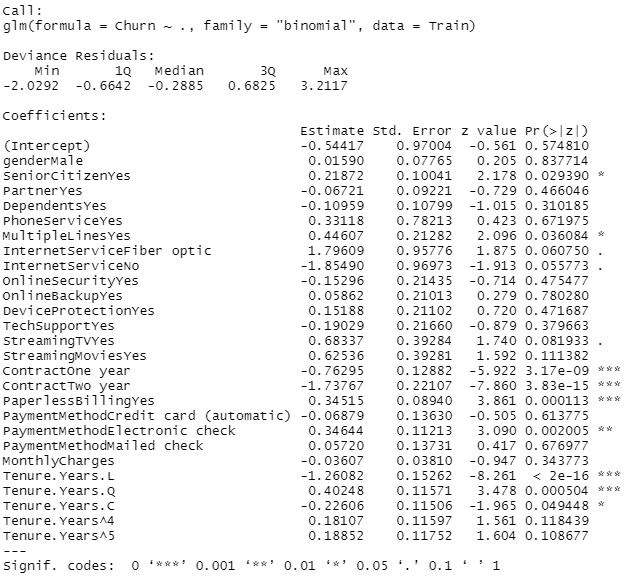
# Descriptive Analysis with Multiple Correspondence Analysis (MCA)

Now that summary statistics and visualizations have been created for each variable, the next objective is to determine which of those variables are most important in predicting churn. After the data cleaning, all but one of the predictor variables are categorical. Because of this, multiple correspondence analysis is used in place of principle component analysis or even factor analysis of mixed data. By mapping the variable correspondence, we can see which variables are more correlated to a customer churning or not.

This biplot of the MCA plots the variables in the data set, which can be compared to the two response variable factors, Churn\_No and Churn\_Yes. Churn\_Yes appears to be correlated with customers in the first year of their tenure, who pay by electronic check on a month-to-month contract. Churn\_No appears to be correlated with customers that aren’t senior citizens, who pay by automatic payment methods, have DSL internet service, and have a longer customer tenure.

# Analytic Analysis with Logistic Regression

Logistic regression is a smart choice for analyzing a binary variable like customer churn. It is used to calculate the probability that an event occurs or not, which fits the scenario much better than linear regression or other analytic methods that predict a value and not a probability of one of two conditions.



A picture containing text, receipt

Description automatically generatedThe logistic regression data above confirms what was shown by the MCA plot that contract, tenure, and paperless billing are the variables with the lowest p-values and most helpful to include in our model to predict churn. Adjusting the model to include those three predictors and predicting on the validation data gives the confidence matrix below.

The logistic model predicts with an accuracy of 77.7% on the validation data. A better way to visually see the usefulness of the model is with an ROC curve, which plots the model’s true positive vs. false positive rates. A model with strong predictive power will lift away from the diagonal line towards the top-left corner, which is the true positive axis. The model does indeed pull more towards that corner and is much higher than the diagonal line that represents random chance prediction.

Chart, line chart

Description automatically generated

# Summary

By using data mining techniques, the answers to our three objectives can be found. First, the distributions for each variable can be viewed and collinearity can be cleared out with bivariate visualizations like the correlation matrix. This helps to remove discrimination in our model, like removing Total Charges which was highly correlated with Monthly Charges.

Next, we can determine which predictor variables are most important to include in the model. Multiple Correspondence Analysis helps to see those variables interacting visually by plotting them against each other and then the logistic regression confirmed which were most influential on model performance. The three predictor variables that showed up on the MCA and had the lowest p-values on the logistic regress were contract type, tenure length, and paperless billing.

Last, the model can be tested against a validation data set to assess performance. By using a logistic regression model with contract type, tenure length, and paperless billing status, churn can be predicted in current and future clients with an accuracy of 77.7%.