Jordan Tompkins

D208: Predictive modeling task 2

A1, Research Question: The research question looked at for this project was, What variables impact Churn? Essentially, which variables of the churn data set have the most impact in predicting if someone will discontinue their service. This question can help the company determine which variables are most impactful when analyzing and even predicting the likelihood of whether a customer is going to discontinue their service or not.

A2, Objectives and goals: The objective and goals of this data analysis is to determine the variables needed to create the best prediction model for our research question. We want to be able to figure out which variables have the most statistical impact on Churn and then create a regression model using these variables.

B1, Summary of Assumptions: There are a number of key assumptions when it comes to logistic regression. A few of those assumptions are that the variables must be independent of each other, there should be little to no multicollinearity between the independent variables. It also “…assumes linearity of independent variables and log odds” (“Assumptions of logistic regression”, 2021). Conducting logistic regression, these were a few of the key assumptions that we looked at to get the initial and the eventual reduced model.

B2, Tool Benefits: A number of different tools can be used to conduct multiple regression analysis. Two big ones are programs such as R or Python. For this question we used the programming language of Python. The reason being is that Python is “…a robust tool to handle, process, and model data. It has an array of packages for linear regression modelling” (Li, 2019). With the many packages such as “ols” from statsmodels, or something like matplotlib for visualization, we are able to perform the multiple regression analysis very smoothly.

B3, Appropriate technique; Logistic regression was an appropriate technique to use for analyzing the research question because we wanted to look into predicting the probability of whether someone discontinued their service or not. Logistic regression allows for that because it Logistic Regression is used to predict the probability of a binary outcome (“What is logistic regression?”, n.d). The Churn variable is the binary outcome we were looking at since it was a “Yes” or “No” response.

C1, Data Goals: The goals of data preparation goals for this research question to was ensure that the data being used, the Churn Data set, was without any missing or null data, and make sure that there were no large outliers that would potentially impact the regression model that was created. The categorical data also had to be re-expressed to allow for the regression model to be created. Some responses were “yes” or “no” responses, or “male”, “female” or “nonbinary” responses, to name a couple. In order to correctly prepare the data for the multiple regression analysis being performed, the categorical variables needed to be re-expressed from their initial responses to either “1” or “0”.

C2, Summary Statistics: When looking at the data set, an important factor we needed to look at was the summary statistics of the data. Using the df.describe() code, we obtained the summary statistics located in the accompanying jupyter notebook uploaded. The summary statistics that were obtained were for all the numerical variables in the data set, after dropping the unnecessary columns. Those variables were ‘Children’, ‘Age’, ‘Income’, ‘Outage\_sec\_perweek’, ‘Bandwidth\_GB\_Year’, ‘Yearly\_equip\_failure’, ‘Tenure’ and ‘MonthlyCharge’. In this summary statistics, the mean, standard deviation and the quartiles of the data, as well as the maximum and minimum were obtained. For ‘Children’, there was a mean of 2.1 kids, with a standard deviation of 2.14. The minimum amount of kids was 0 and the max was 10. The average age of the customers was 53.7 years old, with a standard deviation of 20.7. The youngest was 18 and the oldest was 89. For income, the average was $39.8K while the standard deviation was $28.2K with a minimum of $348 and a maximum of $259K. Outage seconds per week had an average of 10 seconds, a minimum of 0.1 seconds and a maximum of 21.2 seconds. The average bandwidth used per year was 3,392 gigs per year with the lowest being 155 gigs and the highest being 7,159 gigs. ‘Yearly\_equip\_failure’ had a mean of .40 failures, with a minimum of zero failures and a maximum of the equipment failing 6 times. The average length of being a customer, “Tenure”, was 34.5 months with a minimum of 1 for new customers and a maximum of 72.0 months. The lowest ‘MonthlyCharge’ was $80 with an average charge of $172.62 and a maximum of $290.16. Looking at all these statistics, we can tell that there weren’t any extreme values in terms of the minimums or maximums. Do to this, there were no rows dropped.

C3, Steps to Prepare the Data: The Churn data set used contained 10,000 rows of customers with many different variables and information surrounding these customers. Because of this, it was important to check that the data was cleaned and prepared for usage. Once the csv file was loaded into the workbook, the columns deemed unnecessary to the question were dropped. To determine which columns were unnecessary, domain knowledge was the main driver. All columns that involved geographic information, so city, state, latitude, longitude, etc. were dropped because they were specific information to each individual customer. All of that information was unique to each customer and was deemed unnecessary to the research question. It was also determined that any variables that were survey responses, so Item’s 1 through 8, were to be dropped too. These variables were rankings of importance for each individual customer and again, was deemed unnecessary to the research question due to their nature. Columns such as “Techie”, “Contract” and the different types of online security, backup and protection add-ons, etc. were dropped as well. The columns that were left, after dropping other columns, were the columns that were deemed necessary for creating the model using the variables that impacted Churn. These variables all relate to different aspects of the service provided, which was the main focus of the research question. After dropping those columns, .isnull().sum() was called onto the dataframe in order to check for any nulls. We wanted to make sure that there was no missing data, or any variables that were labeled ‘Null’ or “N/A”. The .isnull() allows for us to check this. Using the summary statistics and the distribution visualizations created of the data, we were also able to look for any outliers. Using the distributions and summary statistics, we noted that there were no outliers that needed to be dropped in the data set. See C4, Visualizations, for the distribution’s visualizations utilized for this decision. There was no data that we considered extreme that would be needed to be dropped. The final step in the data preparation phase for creating the multiple regression model was making sure that all the non-numeric variables were changed to numeric values. To use logistic regression, we weren’t able to keep categorical responses. So there can’t be any “Yes” or “No” answers, “gender” had to be reclassified, etc. To do this, we use the pd.get\_dummies method. Using pd.get\_dummies(df, columns =[‘Gender’, ‘Churn’, ‘InternetService’, ‘Multiple’), drop\_first=True) allowed us to “recode” these responses from their categorical response into a numeric response of 0 for no or 1 for yes. Setting the data equal to the data frame and calling the columns we wanted to get the dummy variables of, as well as setting drop\_first equal to “True” to make sure we didn’t create too many dummy variables since a “Yes” variable will be 1 for yes or 0 for no, our data was re-expressed, and ready for use. After we made sure there were no nulls, no extreme outliers, the data was prepared for usage in regression analysis.

C4, Visualizations: The following visualizations are the univariate and bivariate distributions of the predictor variables. For the bivariate analysis and distributions, the distribution is on the predictor and the response variable. See the visualizations below as well as the jupyter notebook attached separately for the plots:

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Tenure:

Chart, histogram

Description automatically generated

Bivariate distribution of Children and Churn:

Chart, box and whisker chart

Description automatically generated

Age and Churn:

Chart, box and whisker chart

Description automatically generated

Income and Churn:

Chart, box and whisker chart

Description automatically generated

Outage\_sec\_perwekk and Churn:

Chart, box and whisker chart

Description automatically generated

Yearly\_equip\_failure and Churn:

Chart, box and whisker chart

Description automatically generated

Tenure and Churn:

Chart, box and whisker chart

Description automatically generated

Gender and Churn:

Chart, bar chart

Description automatically generated

InternetService and Churn:

Chart, bar chart

Description automatically generated

Multiple and Churn:

Chart, bar chart

Description automatically generated

C5, Prepared Data Set: See CSV file uploaded in conjunction with the word document.

D1, Initial Model: The planned predictor variables for this model were ‘Children’, ‘Age’, ‘Income’, ‘Gender\_Male’, ‘Gender\_Nonbinary’, ‘InternetService\_FiberOptic’, ‘InternetService\_None’, ‘Outage\_sec\_perweek’, ‘Yearly\_equip\_failure’, ‘Multiple\_Yes’, and ‘Tenure’. All the categorical variables that were chosen, dummy variables needed to be created in order for the logistic to be completed. The variables could not stay as ‘Yes’ or ‘No’, or ‘Male’ or ‘Female’, etc. responses, so using pd.get\_dummies and entering in the columns categorical columns we were looking at, dummy variables were used to created the initial model. After creating the dummy variables the VIF, or variance inflation factor, was checked for multicollinearity. It was noted that no variables had a VIF of larger than the cutoff of 10 that was used. That left the initial model being ‘Churn\_Yes ~ Children + Age + Income + Gender\_Male + Gender\_Nonbinary + Outage\_sec\_perweek + Yearly\_equip\_failure + InternetService\_FiberOptic + InternetService\_None + Multiple\_Yes + Tenure’.

D2, Justification of Model Reduction: The method used for the reduction of the model was “Backward Stepwise Elimination”. This method allows for us to use all the initial variables in the initial model and then we eliminate the least significant variables (Choueiry, 2022). Using Backward Stepwise Elimination, we look for the variables that have high p-values and eliminate them until there are no significant p-values for variables left. When using Backward Stepwise Elimination, you start with the initial model discussed in section D1, and you eliminate variables, one at a time, until you reach the proper reduced model. The factor that determined elimination was statistical significance, so looking at p-values. Any p-value greater than a predetermined alpha level of 0.05 was dropped. Our initial model had 6 variables that had significant p-values. So going in order based on the OLS Regression Results located in the jupyter notebook uploaded with the code, the first variable to be dropped was “Children” since the p-value was 0.732, well above the 0.05 cutoff. We created a new regression model with the remaining variables and saw that ‘Age’ was the next variable that needed to be dropped. It had a p-value of 0.082. After that, the next model gave us “Income” with a p-value of 0.507. This variable was dropped accordingly. The model after this had ‘Outage\_sec\_perweek’ with a p-value of 0.841, so this variable was also dropped. ‘Gender\_Nonbinary’ had p-values of 0.5 in the next model created. After dropping ‘Gender\_Nonbinary’, ‘Outage\_sec\_perweek’ had a p-value of 0.895, so that variable got dropped. After this, the final variable, ‘Yearly\_equip\_failure’ was dropped with a p-value of 0.262. Using Backward Stepwise elimination, and eliminating each variable one by one until there were no variables with p-values greater than 0.05, we were left with all variables that were to be used in the final reduced model.

D3, Reduced Multiple Regression Model; The reduced model was 'Churn\_Yes ~ Gender\_Male + InternetService\_FiberOptic + InternetService\_None + Multiple\_Yes + Tenure'. See below for the output of both the initial and the reduced models:

Initial model: Table

Description automatically generated

Reduced model:

**Table

Description automatically generated with medium confidence**

E1, Model Comparison: The two models that were created were very similar in terms of certain model evaluation metrics, but also different in terms of others. A few key evaluation metrics used to compare the models were Psuedo R-Squared, LL-p-values and Akaike’s Information Criteria, or AIC. The initial model had a Psuedo R-squared of 0.2632 while the reduced models Psuedo R-Squared was 0.2627. This metric can help determine which model is better. The higher the value, the better the fit. Both models had LL-p-values of 0.000, which when compared to the typical threshold of 0.05, means that the two models were overall, good models to use for our regression model. AIC or Akaike’s Information Criteria was also looked at. This was one of the main metrics looked at for the model comparison. The reduced model had an AIC of 8538.25, while the initial model had an AIC of 8544.94. The lower the AIC, the better the fit. So in this case, the reduced model would have the better fit when looking at AIC. As previously stated, the model was reduced using backward stepwise elimination. The initial model had a few variables that had p-values greater than 0.05 which caused them to be dropped. The reduced model did not have any variables with a p-value greater than 0.05. So for the reduced model, no more variables were dropped.

E2, Output and Calculations:

**Table

Description automatically generated with medium confidence**

Graphical user interface

Description automatically generated with low confidence

A picture containing graphical user interface

Description automatically generated

Chart, treemap chart

Description automatically generated

E3, Code: See jupyter notebook attached for all code related to the implementation of the regression model.

F1, Results: There is a lot that can be discussed about the reduced model created in this logistic regression analysis. For starters, a regression equation was created. This equation is: ln(Churn\_Yes/(1-Churn\_Yes)) = 0.4053 + .1310(Gender\_Male) -0.6275(InternetService\_FiberOptic) – 0.6457(InternetService\_None) +7811(Multiple\_Yes) - 0.0578(Tenure)**.** There were 5 variables in our reduced model. The coefficients for these variables shows how much the likelihood that a customer will Churn will either increase or decrease with a one unit increase in that variable. So for a one unit increase in the Gender\_Male variable, the likelihood that the customer would churn would increase by 0.130. For every increase in InternetService\_FiberOptic, the likelihood of the customer churning would decrease by 0.6275. An increase in InternetService\_None, the likelihood of churning would decrease by 0.6457, while an increase in Multiple\_Yes would increase the likelihood by 0.7811. Finally, an increase in Tenure would decrease the likelihood of the customer churning by 0.0578. One thing to look at in terms of important metrics in the regression model was the LLR p-value. Our reduced model had a LL p-value of 0.000 which means that the model is statistically significant. It is practically significant due to the fact that the results of this analysis will be helpful to the company in future business operations. While the model was able to be reduced down using logistic regression, there are a few limitations using this analysis. A big one is that the number of observations needs to be more than the number of features. (GeeksforGeeks, 2023). Due to this, smaller data sets with a large number of variables might run into issues of overfitting. It also requires average or zero multicollinearity between the predictor variables. This is why it is essential to check for multicollinearity using something like variance inflation factor, or VIF. Having a multicollinearity problem will lead to issues with the model and needs to be explored before using logistic regression. These are just a few of the limitations that logistic regression has. However, with proper variable usage, and proper data cleaning and preparation, the limitations could be mitigated.

F2, Recommendations: Using logistic regression, we were able to create a reduced regression model that could help the company answer the question of what variables will impact whether or not a customer churns. Using this reduced model, the company can more easily predict how likely a customer will discontinue their service with certain, statistically significant variables that were used in the final reduced model. If they wanted to use this model to help with the business, they can use it to help justify a focus on certain areas of business to try and keep their existing customers.

G, Demonstration: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1d5a7d6e-084d-4675-bc5c-afc70109bee0>

H, Sources of Third-Party Code: No third-party code utilized

I, Sources:

*Assumptions of logistic regression*. Statistics Solutions. (2021, August 11). Retrieved March 11, 2023, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

Choueiry, G. (2022, November 25). *Quantifying health*. QUANTIFYING HEALTH. Retrieved March 4, 2023, from https://quantifyinghealth.com/stepwise-selection/

GeeksforGeeks. (2023, January 10). *Advantages and disadvantages of logistic regression*. GeeksforGeeks. Retrieved March 11, 2023, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Li, L. (2019, February 5). *Introduction to linear regression in python*. Medium. Retrieved March 4, 2023, from https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0#:~:text=Understanding%20how%20to%20implement%20linear,packages%20for%20linear%20regression%20modelling.

*What is logistic regression? - definition from Techopedia*. Techopedia.com. (n.d.). Retrieved March 11, 2023, from https://www.techopedia.com/definition/32065/logistic-regression