**LOGISTIC REGRESSION MODELING (NBM3 TASK 2)– D208**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

***Part I: Research Question***

**A-1**

**“What are the most important factors that contribute to a customer discontinuing service within the last month?”**

**A-2**

**The goal of my data analysis is to understand what factors are causing customers to discontinue service within the last month.**

***Part II: Method Justification***

**B-1**

**Below are four assumptions of a logistic regression and a brief explanation of them.**

* **The Response Variable is Binary. The response variable needs to have binary responses to be in a logistic regression.**
* **There is No Multicollinearity Among Explanatory Variables. The exploratory variables have no relationship with each other.**
* **No Extreme Outliers. Data points are not too far out of the inner quartile range.**
* **Sample Size is Sufficiently Large. Enough data points are needed to ensure the sample tells a similar thing to the whole of what they are representing.**

**B-2**

**Python is the programming language I will be using for my analysis. Python has many libraries and packages that can produce everything that is asked for in this analysis. For the analysis I will be using pandas, for data manipulation, matplotlib and seaborn, for graphical visualizations, SciPy and NumPy, for mathematical and statistical results. Python is an open-source language, this allows for public to use it without licenses leading it to have a large collaborative community. This allowed me to learn and use Python prior to enrolling in this degree path.**

**B-3**

My question requires using a logistic regression to analyze the relationship of independent variables have on a binary dependent variable. Churn is the binary response variable from my question. My question looks to see the relationship independent variables have on Churn.

***Part III: Data Preparation***

**C-1**

My goal in cleaning the given data is to detect and treat null values and outliers. I will be using “.info” function to see what columns have any null values present. I know this data should have 10,000 entries for each variable. Anything lower than this in the Non-Null Count will show nulls are in the dataset. After this, I will start looking into each column individually. Using “.value\_counts()” for qualitative columns will allow me to get and test the min and max for outliers using the rule of thumb of z-value is either greater than 3, or less than -3 to identify outliers(Larose & Larose, 2019). Please see uploaded Jupyter notebook for code.

**C-2**

The independent variables Children, Age, Income, Marital, Gender, Techie, Contract, Port\_modem, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Bandwidth\_GB\_Year, MonthlyCharge, and my dependent variable Churn are all the variables that will be in my initial logistic regression.

The qualitative variables include Marital, Gender, Techie, Contract, Port\_modem, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies. I will provide the breakdown of the percentage of each response.

* Marital
  + Number of Observations: 10,000
  + Married: 19.1%
  + Divorced: 20.9%
  + Widowed: 20.3%
  + Separated: 20.1%
  + Never Married: 19.6%

A pie chart with different colored circles with Crust in the background

Description automatically generated

* Gender
  + Number of Observations: 10,000
  + Male: 47.4%
  + Female: 50.2%
  + Nonbinary: 2.3%

A pie chart of a customer distribution

Description automatically generated

* Techie
  + Number of Observations: 10,000
  + Yes: 16.8%
  + No: 83.2%

A blue and orange pie chart

Description automatically generated

* Contract
  + Number of Observations: 10,000
  + Month-to-month: 54.6%
  + One year: 21.0%
  + Two year: 24.4%

A pie chart with numbers and a number of people

Description automatically generated

* Port\_modem
  + Number of Observations: 10,000
  + Yes: 48.3%
  + No: 51.7%

A pie chart with text

Description automatically generated

* InternetService
  + Number of Observations: 10,000
  + None: 21.3%
  + Fiber Optic: 44.1%
  + DSL: 34.6%

A pie chart of a customer service distribution

Description automatically generated

* Phone
  + Number of Observations: 10,000
  + Yes: 90.7%
  + No: 9.3%

A blue circle with orange triangle and black text

Description automatically generated

* Multiple
  + Number of Observations: 10,000
  + Yes: 46.1%
  + No: 53.9%

A pie chart with text

Description automatically generated

* OnlineSecurity
  + Number of Observations: 10,000
  + Yes: 35.8%
  + No: 64.2%

A blue and orange pie chart

Description automatically generated

* OnlineBackup
  + Number of Observations: 10,000
  + Yes: 45.1%
  + No: 54.9%

A blue and orange pie chart

Description automatically generated

* DeviceProtection
  + Number of Observations: 10,000
  + Yes: 43.9%
  + No: 56.1%

A pie chart with numbers and a number of text

Description automatically generated

* TechSupport
  + Number of Observations: 10,000
  + Yes: 37.5%
  + No: 62.5%

A pie chart with numbers and a number on it

Description automatically generated

* StreamingTV
  + Number of Observations: 10,000
  + Yes: 49.3%
  + No: 50.7%

A pie chart of tv distribution

Description automatically generated

* StreamingMovies
  + Number of Observations: 10,000
  + Yes: 48.9%
  + No: 51.1%

A pie chart of a movie distribution

Description automatically generated

The quantitative variables Children, Age, Income, Bandwidth\_GB\_Year, and MonthlyCharge will be getting the summary statistics pulled and investigated for outliers. The summary statistics will include the mean, standard deviation, the inner quartile range, median, minimum, and maximum.

Children has an outlier present as the max according to z-values but is valid because ten children is not unreasonable so I will not be doing any changes to the outliers in this variable to keep my results as true to the data as possible.

A screenshot of a computer

Description automatically generated

Age does not have an outlier present.

A screenshot of a computer

Description automatically generated

Income has an outlier present as the max according to z-values, but is valid because income of $258,900 while rare is not unreasonable so I will not be doing any changes to the outliers in this variable to keep my results as true to the data as possible.

A screenshot of a computer

Description automatically generated

Bandwidth\_GB\_Year does not have an outlier present.

A screenshot of a computer

Description automatically generated

MonthlyCharge does not have an outlier present.

A screenshot of a number of numbers

Description automatically generated

**C-3**

Below are my univariate visualization and then bivariate visualization relationship with my dependent variable.

A blue and orange pie chart

Description automatically generated

A graph of a number of numbers and a chart of a number of numbers

Description automatically generated with medium confidence

A graph with numbers and text

Description automatically generated

A blue graph with white text

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidence

A pie chart with text on it

Description automatically generated

A screen shot of a pie chart

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A pie chart with numbers and text

Description automatically generated

A screenshot of a graph

Description automatically generated

A pie chart with text on it

Description automatically generated

A screenshot of a phone distribution

Description automatically generated

A screenshot of a graph

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screenshot of a data server

Description automatically generated

A screen shot of a chart

Description automatically generated

A screen shot of a chart

Description automatically generated

A screen shot of a chart

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A graph of a person with a churn

Description automatically generated with medium confidence

A graph of a person and person

Description automatically generated

A graph with blue and orange squares

Description automatically generated

A graph of a bar chart

Description automatically generated with medium confidence

A graph with blue and orange bars

Description automatically generated

A graph of different colored bars

Description automatically generated

A graph of a customer response

Description automatically generated with medium confidence

A graph of a number of blue and orange bars

Description automatically generated

A graph of a customer

Description automatically generated with medium confidence

A graph of a customer backup response

Description automatically generated with medium confidence

A graph of a number of blue and orange bars

Description automatically generated

A graph of a customer support response

Description automatically generated

A graph of a number of blue and orange bars

Description automatically generated

A graph of a customer

Description automatically generated with medium confidence

A chart of a customer

Description automatically generated with medium confidence

A diagram of a relationship

Description automatically generated

A diagram of a customer age

Description automatically generated with medium confidence

A diagram of a relationship between a customer and a churn status

Description automatically generated

A diagram of a couple of blue squares

Description automatically generated

**C-4**

The goal for my data transformation steps is to prepare the data to be used in my initial logistic regression. First, I am changing the variables that have a yes or no response into Boolean true or false data type. Instead of having true or false, they will be 1 for yes and 0 for no. Next, I am re-expression of categorical variables, Marital, Gender, Contract, and InternetService using One-hot encoding. In doing this I will be creating dummy variables of one less than the given responses. In doing this, for example, Marital will produce two dummy variables since it has three choices. One of these dummy variables for Martial will be for the Male response and will result in 1 for customers that are male and 0 for the rest. This is done because the values in the variables that will be transformed using One-hot encoding do not have rank to their values, meaning one is not better than the others. After this, I will create a new dataset with just the columns of variables that I am using for my initial logistic regression. With the new dataset created, the dummy variables need to be added. Lastly, I will check visually and then export it as a csv file.

**C-5**

Please see uploaded CSV file, log\_churn.csv, for the dataset prepared for my initial logistic regression.

***Part IV: Model Comparison and Analysis***

**D-1**

**My initial logistic regression model was created with y being my dependent variable Churn and x being my multiple independent variables that were listed in C-2.** My model assigns a y-intercept a value of 1. Below are the results of running the initial model.

A screenshot of a computer

Description automatically generated

**D-2**

**After I ran my model, I decided to check for multicollinearity. I will look at the Variance Inflation Factor to check for the multicollinearity. I will be using a threshold of VIF greater than 5. I understand that using 5 vs 10 has its tradeoffs on what happens to the model. Using 5 will result in a more precise model with less multicollinearity but could have relevant variables being left off. My question is looking for the most important factors so I think the VIF greater than 5 will result in a more appropriate model for this question. I will be dropping the highest VIF that is over 5 and rerunning until I no longer get variables with a VIF over 5. Doing this resulted in variables MonthlyCharge, Phone, and Age being dropped. Now I run my model again with the variables taken out and start my process of Backward Stepwise Elimination. I am using an alpha value of 0.05 to do my elimination as any variable with a p-value higher than this indicates it is not statistically significant. I will be eliminating the variable with the highest p-value over 0.05 and run the model again without it until all variables have p-values under 0.05 meaning they will all be considered statistically significant. This resulted in the removal of Marital\_Never\_Married, OnlineSecurity, Income, Gender\_Nonbinary, Marital\_Married, Marital\_Separated, Port\_modem, Marital\_Widowed from the model. This leaves Children, InternetService\_Fiber\_Optic, InternetService\_None, Contract\_Two\_year,** **Contract\_One\_year, Gender\_Male, Techie, Multiple, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and** **Bandwidth\_GB\_Year as my independent variables in my model.**

**D-3**

**Please see the reduced model below. It will also be in my uploaded Jupyter notebook Christian LeBlanc D208 PA 2.**

A screenshot of a computer

Description automatically generated

**E-1**

After taking the steps in D-2 my reduced logistic regression model has less multicollinearity and all the remaining independent variables have p-values that indicate they are statistically significant. This should lend to the reduced model being a better fit than the initial model, but after checking five different model evaluation metrics it does not show a better fit. First, I checked the Log-Likelihood, and the initial model has a value of -2196.3 while the reduced has -2249.7 but I didn’t think much of this because they do not have the same number of predictor variables. So, next I compared the Pseudo R squared values. The initial model has 0.6202 while the reduced model has 0.6109. This would make the initial model the better model since it has the higher Pseudo R squared. Moving on to the LLR p-value for the reduced model of 0.000 shows that the model is “useful”. I also investigated Akaike’s Information Criteria (AIC) and Bayesian information criteria (BIC) to see if they could point to the reduced model being the better fit. For the AIC the lower value offers the best fit. The initial model has the lower value of AIC being approximately 4444 while the reduced model has approximately 4529. For the BIC the lower value offers the best fit. The initial model has the lower value of BIC being approximately 4632 while the reduced model has approximately 4637.

**E-2**

The confusion matrix created from the reduced model produced 6913 True Negatives(TN), 437 False Negatives(FN), 548 False Positive(FP), and 2102 True Positives(TP). The confusion matrix produced an accuracy of 0.9015, I got this from importing the accuracy\_score from sklearn.metrics library in python. This can be checked manually by adding the TP and TN which is 9015 and divide by the sum of TP, TN, FP, and FN, which is 10000 and results in the same accuracy that I got.

**E-3**

Please see **my uploaded Jupyter notebook Christian LeBlanc D208 PA 2 for a copy of the code for the** implementation of the logistic regression models.

***Part V: Data Summary and Implications***

**F-1**

**The equation for the reduced multiple logistic regression is below.**

**The coefficients from this equation have the following effects**

* Children has a coefficient of 0.0573. Keeping all things constant, for one unit increase in Children, changes log odds of Churn by 0.0573.
* InternetService\_Fiber\_Optic has a coefficient of -1.9079. Keeping all things constant, customers that have InternetService\_Fiber\_Optic, changes log odds of Churn by -1.9079 than customers with DSL.
* InternetService\_None has a coefficient of -1.9778. Keeping all things constant, customers that have InternetService\_None, changes log odds of Churn by -1.9778 than customers with DSL.
* Contract\_Two\_year has a coefficient of -3.3599. Keeping all things constant, customers that have Contract\_Two\_year, changes log odds of Churn by -3.3599 than customers with month-to-month contract.
* Contract\_One\_year has a coefficient of -3.2844. Keeping all things constant, customers that have Contract\_One\_year, changes log odds of Churn by -3.2844 than customers with month-to-month contract.
* Gender\_Male has a coefficient of 0.3597. Keeping all things constant, customers that are Gender\_Male, changes log odds of Churn by 0.3597 than customers with the gender female.
* Techie has a coefficient of 1.0874. Keeping all things constant, customers that are techie, changes log odds of Churn by 1.0874.
* Multiple has a coefficient of 1.7307. Keeping all things constant, customers that have multiple lines, changes log odds of Churn by 1.7307.
* OnlineBackup has a coefficient of 0.9131. Keeping all things constant, customers that have OnlineBackup, changes log odds of Churn by 0.9131.
* DeviceProtection has a coefficient of 0.5568. Keeping all things constant, customers that have DeviceProtection, changes log odds of Churn by 0.5568.
* TechSupport has a coefficient of 0.2771. Keeping all things constant, customers that have TechSupport, changes log odds of Churn by 0.2771.
* StreamingTV has a coefficient of 3.2333. Keeping all things constant, customers that have StreamingTV, changes log odds of Churn by 3.2333.
* StreamingMovies has a coefficient of 3.7233. Keeping all things constant, customers that have StreamingMovies, changes log odds of Churn by 3.7233.
* Bandwidth\_GB\_Year has a coefficient of -0.0013. Keeping all things constant, for one unit increase in Bandwidth\_GB\_Year, changes log odds of Churn by -0.0013.
* The constant is -1.2602. This shows that when all the variables in the model are 0, then the log odds of Churn are -1.2602.

The reduced logistic regression model is statistically significant. The given value of 0.000 for the LLR p-value is an indicator of this. The reduced model is also practically significant. The reason I believe in the practical significance of the results is twofold. The first is dealing with the contracts. The one- and two-year contracts both decrease the odds of Churn when compared with the month-to-month reference category. This is something that seems logical, that customers with longer contracts are less likely to leave, but having the data confirm it is good to know and will be a factor in my recommended course of action. The other part that I find practically significant is the coefficients on both the streaming variables. This shows that StreamingTV and StreamingMovies increase the odds of churn with higher coefficients than the other variables. This becomes interesting when I put it together with the results from my linear regression model from task 1, as the same two variables were the biggest drivers of monthly charges.

One of my biggest concerns about the limitations of the analysis on this data is it feels like it is a look at data at one chosen point, because churn in the data dictionary is if a customer discontinued service in the last month. So, at best, it is data from a single month and when trying to determine what factors that could be lending to customers leaving, I feel like looking over a longer period would give better results. In a real world setting I would think accessing something like customer’s signup date and date of discontinue would be easy to obtain and would give better results for my question. Another concern that I have is the results from the model evaluation metrics. None of the metrics have the reduced model as the better fitting model over the initial.

**F-2**

My recommendation on a course of action considers my results from my linear regression results, because not only trying to keep customers but also those that are profitable. I would recommend for sales to push the longer contracts to keep customers longer than the month-to-month. The streaming tv and movies is a little more complex because they are tied to the highest monthly charges even though they increase the log odds of churn. I think to keep customers that are paying the most monthly, customers with streaming tv and movies from my linear regression, I would suggest possibly include a slight discount on the streaming tv and movies for customers with contracts that are one and two years. I believe this would help both keep customers and promote things that would get the most from those customers.

***Part VI: Demonstration***

**G**

Uploaded it to the Panopto drop box titled “Regression Modeling – NBM3 | D208.” Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=278ef3dc-8676-4d5a-b5a8-b0ed01417302>

**H**

I did not use any web sources to acquire data or segments of third-party code in this Performance Assessment.

**I**

Chantal D. Larose, & Daniel T. Larose. (2019). *Data Science Using Python and R*. Wiley.

**J**

The content in this Performance Assessment is set up and presented with the highest professional standards.