**LINEAR REGRESSION MODELING (NBM3 TASK 1)– 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’s monthly charge?”**

**A-2**

**The goal of my data analysis is to understand what factors contribute to a customer’s monthly charge.**

***Part II: Method Justification***

**B-1**

**Using a multiple linear regression model requires some assumptions to be true. First is the relationship between the independent and dependent variables are assumed to have a linear relationship. This is shown when the independent variable has a change in value the dependent variable has a proportional change with it. Next, is an assumption of residuals are to be normally distributed. So, the variance from the line of best fit, residuals, should form a bell-shaped curve when observed. Also, the assumption of Homoscedasticity. This is assuming that the variance of the residuals is consistent for all of the predicted variable. Lastly, the residuals should be independent of each other.**

**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 multiple linear regression to analyze the relationship of independent variables have on a dependent variable, MonthlyCharge. MonthlyCharge is a continuous variable and I am looking to see the impact of multiple independent variables have on MonthlyCharge.

***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, as I know this data should have 10,000 entries for each variable anything lower than this in the Non-Null Count will show if any 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, and my dependent variable MonthlyCharge are all the variables that will be in my initial multiple linear 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 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 screenshot of a graph

Description automatically generated

A chart with blue squares and text

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

**C-4**

The goal for my data transformation steps is to prepare the data to be used in my initial multiple linear 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. This allows use to use the same value for all the response and to be used in the multiple linear regression. After this, I will create a new dataset with just the columns of variables that I am using for my initial multiple linear regression. With the new dataset created, the dummy variables need to be added. Lastly, I check visually and then export it as a csv file.

**C-5**

Please see uploaded CSV file, mlr\_churn.csv, for the dataset prepared for my initial multiple linear regression.

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

**D-1**

**My initial multiple linear regression model was created with y being my dependent variable MonthlyCharge 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, the notes from the model indicated that it has a large condition number which might indicate a strong multicollinearity. This is the first thing I will investigate for my reduction. I will look at the Variance Inflation Factor. 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 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\_Married, Marital\_Separated, Children, Marital\_Widowed, Contract\_Two\_year,** **Contract\_One\_year,** **Marital\_Never\_Married,** **Bandwidth\_GB\_Year,** **Income,** **Port\_modem,** **Gender\_Male,** **Gender\_Nonbinary, and Techie from the model. This leaves InternetService\_Fiber\_Optic, InternetService\_None, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies 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 1.**

A screenshot of a computer

Description automatically generated

**E-1**

After taking the steps in D-2 my reduced multiple linear regression model has less multicollinearity and all the remaining independent variables have p-values that indicate they are statistically significant. Using Akaike’s Information Criteria (AIC) as a model evaluation metric shows that the reductions made resulted in a better model than the initial. The initial multiple linear regression gave an AIC value of 7.189e+04 while the reduced model gave a lower AIC value of 7.187e+04.

**E-2**

Please see the histogram of the residuals from the reduced multiple linear regression model. It shows the values of the residuals on the x-axis and the frequency of that value of residual on the y-axis. The reduced model has a residual standard error of 8.794851152426581.

A graph of a number of blue lines

Description automatically generated with medium confidence

A close up of a computer screen

Description automatically generated

**E-3**

Please see **my uploaded Jupyter notebook Christian LeBlanc D208 PA 1 for a copy of the code for the** implementation of the linear regression models.

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

**F-1**

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

**The coefficients from this equation have the following effects on the dependent variable:**

* **83.8996 is the y-intercept and is the average expected value for the dependent variable MonthlyCharge when all of the independent variables are equal to zero.**
* **52.3185 is the coefficient for StreamingMovies. This tells us that keeping all things constant, customers that have streaming movies have on average a 52.3185 higher MonthlyCharge than customers that do not.**
* **42.1836 is the coefficient for StreamingTV. This tells us that keeping all things constant, customers that have streaming TV have on average a 42.1836 higher MonthlyCharge than customers that do not.**
* **12.5622 is the coefficient for TechSupport. This tells us that keeping all things constant, customers that have TechSupport have on average a 12.5622 higher MonthlyCharge than customers that do not.**
* **12.5241 is the coefficient for DeviceProtection. This tells us that keeping all things constant, customers that have DeviceProtection have on average a 12.5241 higher MonthlyCharge than customers that do not.**
* **22.5623 is the coefficient for OnlineBackup. This tells us that keeping all things constant, customers that have OnlineBackup have on average a 22.5623 higher MonthlyCharge than customers that do not.**
* **2.6711 is the coefficient for** **OnlineSecurity. This tells us that keeping all things constant, customers that have OnlineSecurity have on average a 22.5623 higher MonthlyCharge than customers that do not.**
* **32.5984 is the coefficient for Multiple.** **This tells us that keeping all things constant, customers that have Multiple have on average a 32.5984 higher MonthlyCharge than customers that do not.**
* **The following coefficient are for Indicator (Dummy) variables for the categorical variable InternetService. -12.8822 is the coefficient for InternetService\_None This tells us that keeping all things constant, customers that have InternetService\_None have on average a 12.8822 lower MonthlyCharge than customers that have DSL(Omitted reference category).**
* **The following coefficient are for Indicator (Dummy) variables for the categorical variable InternetService. 19.8385 is the coefficient for InternetService\_Fiber\_Optic. This tells us that keeping all things constant, customers that have InternetService\_Fiber\_Optic have on average a 19.8385 higher MonthlyCharge than customers that have DSL(Omitted reference category).**

**The reduced model is statistically significant. This is evident in having R-squared value of 0.958 and the p-value associated with the F-Statistic shows 0.00. The reduced model is also practically significant. The reason I am saying that it has practical significance is of the coefficient associated with the streaming products StreamingTV and StreamingMovies are the two biggest coefficients. This is very useful information to give to the sales team because you can have them pushing StreamingTV and StreamingMovies knowing that customers that have these two correlates to the two biggest jumps in MonthlyCharge. So they can be pushing StreamingMovies to customers even though the price to get may not be around the 52.3185 monthly the customer having it on average has a MonthlyCharge increase of 52.3185. Selling strategies like this is a good way to increase the amount of money a customer spends.**

**While I do believe the reduced model gave good insight, like all models it is not perfect. The linearity of the model is not ideal, with many variables having values close to 0. I feel this is not surprising with the variable that were remaining in the reduced model. The “yes” or “no” nature of the remaining variables does not lend itself to linearity. Another concern I have with my model is the Residual Standard Error results. 8.794851152426581 for the reduced model is greater than the 8.79420895844156 for the initial model. This means after I reduced by doing the VIF and then the Backward Stepwise Elimination the spread of the residuals around the regression line were greater. Selection of the independent variables is always a concern for me when making a model like this. While I tried to cast a wide net to get all the things that I thought would have an impact on MonthlyCharge sometimes things that are not obvious could show up in analysis that could’ve had an impact on my reduced model. I did not include a variable like Longitude, because I personally don’t think it would have any correlation with MonthlyCharge, but if there was a correlation with the larger Longitude line came a higher MonthlyCharge we wouldn’t know.**

**F-2**

**The top two most important factors to a customer’s MonthlyCharge are StreamingMovies and StreamingTV. As mentioned before, I believe that the results from my reduced model are not only statistically significant but also practically significant. I would recommend that the sales team be informed of this and advised to try to promote StreamingMovies and StreamingTV. It is unclear if these two things are inhouse services that are offered or a general question they asked customers and they said yes if they had a Netflix or YoutubeTV. Either way it should a topic of conversation in all sales going forward because on average with everything else constant customers that have StreamingMovies have a greater MonthlyCharge of 52.3185.**

***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=c2d6c573-032d-4e0f-87e5-b0dc018098d3>

**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.