Jordan Tompkins

D208: Predictive modeling task 1

A1, Research Question: The research question looked at for this project was, What variables impact MonthlyCharge? Essentially, which variables of the churn data set cause the amount charged to the customer every month to be that amount. This question can help the company determine which variables are most impactful when analyzing and even predicting a customer’s bill for the month.

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 MonthlyCharge and then create a regression model using these variables.

B1, Summary of Assumptions: There are a number of key assumptions when it comes to multiple regression. A few of those assumptions are that there must be a linear relationship between the response and predictor variables, the residuals are normally distributed, the independent variables, or predictor variables, are not highly correlated with each other. There is also the assumption of homoscedasticity which means “… that the variance of error terms are similar across the values of the independent variables” (“Assumption of linear regression”, 2021). Conducting multiple linear 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; Multiple Regression was the appropriate technique for this question for a couple reasons. The reason why we chose linear regression was the fact that the response variable, MonthlyCharge, was a continuous variable. Multiple regression was used, rather than simple linear regression which has one predictor and one response variable, because a multiple regression model allows for the use of several explanatory variables (Hayes, 2023). Because the question wanted to look at more than just one predictor variable, multiple regression, also known as multiple linear regression, was needed.

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’ and our response variable of ‘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. 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 the MonthlyCharge. These variables all relate to customer usage and their impact on MonthlyCharge, 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. Using multiple linear regression does not allow for predictor variables to be non-numeric. 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 as well as the .map method on “InternetService” since this variable had a couple different responses. Using pd.get\_dummies(df, columns =[‘Gender’, ‘Port\_modem’, ‘Tablet’, ‘Multiple’, ‘StreamingTV’, ‘StreamingMovies’), 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, waterfall chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Bandwidth\_GB\_Year:

Chart, histogram

Description automatically generated

Bivariate distribution of Children and MonthlyCharge:

Chart

Description automatically generated

Age and Monthly Charge:

Chart, scatter chart

Description automatically generated

Income and MonthlyCharge:

Chart, scatter chart

Description automatically generated

Outage\_sec\_perwekk and MonthlyCharge:

Chart, scatter chart

Description automatically generated

Bandwidth\_GB\_Year and MonthlyCharge:  
Chart, scatter chart

Description automatically generated

Gender and MonthlyCharge:

Chart, box and whisker chart

Description automatically generated

Port\_modem and MonthlyCharge:

Chart, box and whisker chart

Description automatically generated

Tablet and MonthlyCharge:

Chart, box and whisker chart

Description automatically generated

Multiple and MonthlyCharge:

Chart, box and whisker chart

Description automatically generated

StreamingTV and MonthlyCharge:

Chart, box and whisker chart

Description automatically generated

StreamingMovies and MonthlyCharge:

Chart, box and whisker chart

Description automatically generated

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

D1, Initial Model: The predictor variables we used for this data set were ‘Gender\_Male’, ‘Gender\_Nonbinary’, ‘InternetService’, ‘Bandwidth\_GB\_Year’, ‘Children’, ‘Age’, ‘Income’, ‘Outage\_sec\_perweek’, ‘Tablet\_Yes’, ‘Port\_modem\_Yes’, ‘Multiple\_Yes’, ‘StreamingTV\_Yes’, and ‘StreamingMovies\_Yes’. All the categorical variables that were chosen, dummy variables needed to be created in order for the multiple linear regression 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 and checking the VIF for multicollinearity, we were able to create our initial model. The initial model was: 'MonthlyCharge ~ intercept + InternetService + Bandwidth\_GB\_Year + Children + Age + Income + Outage\_sec\_perweek + Gender\_Male + Gender\_Nonbinary + Port\_modem\_Yes + Tablet\_Yes + Multiple\_Yes + StreamingTV\_Yes + StreamingMovies\_Yes'

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 8 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 0.913, 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 0.111. After that, the next model gave us “Income” with a p-value of 0.247. 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. Both ‘Gender\_Male’ and ‘Gender\_Nonbinary’ had p-values of 0.581 and 0.368 in the next two models created, respectively. There were still a couple variables with high p-values, so the next varible to be dropped was ‘Port\_modem\_Yes’ due to the p-value being 0.437. After this, the final variable, ‘Tablet\_Yes’ was dropped with a p-value of 0.198. 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 'MonthlyCharge ~ intercept + InternetService + Bandwidth\_GB\_Year + Multiple\_Yes + StreamingTV\_Yes + StreamingMovies\_Yes'. 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

E1, Model Comparison: The two models that were created were very similar in terms of certain model evaluation metrics. Both models had an R-squared of 0.836 with an adjusted R-squared of 0.835. The R-Squared metrics measures the variation of the predictor variable that is explained by the independent variables. So 83.6% of variation in MonthlyCharge was explained by the predictor variables, for both models. The F-Statistic was different in the two models. The F-Statistic was higher in the reduced model than in the initial model. The adjusted R-Squared was also important to look at. The reduced model had an Adj. R-Squared of 1.015e4 which is much higher than the initial Adj. R-Squared of 3905. The Adj R-Squared increased in the reduced model. AIC or Akaike’s Information Criteria was also looked at. The reduced model had an AIC of 8.554e4 while the initial model had an AIC of 8.555e4. The lower the AIC, the better the fit. So in this case, the reduced model would have the better fit when looking at AIC. Another important metric looked at was the residual standard error. This metric measures how well the model will fit the data set. It measures the standard deviation of the residuals (Zach, 2021). For both models, the residual standard error was 17.42, so very similar to each other. 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. The residual plot was also looked at for the model, noting that they were a little scattered, rather than a fitted linear model. See plot below.

Chart, scatter chart

Description automatically generated

Histogram of the residuals:

Chart, histogram

Description automatically generated

E2, Output and Calculations: Graphical user interface, text, application

Description automatically generated

Table

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 multiple linear regression analysis. For starters, a regression equation was created. This equation is: MonthlyCharge = 99.013 + 24.6994(InternetService) + 0.0004(Bandwidth\_GB\_Year) + 32.6353(Multiple\_Yes) + 41.9028(StreamingTV\_Yes) + 52.3894(StreamingMovies\_Yes) **.** There were 5 variables in our reduced model. The coefficients for these variables shows how much the Monthly Charge will either increase or decrease with a one unit increase in that variable. So in the reduced model, for a one unit increase in fiber optic for internet service, the monthly charge would increase by 24. Bandwidth only increased monthly charge by 0.0004, while if the customer had more than one line of service, the monthly charge increased by 32.6. If the customer had the TV streaming service or movie streaming service, the monthly charge increased by 41.9 and 52.4 respectively. One thing to look at in terms of important metrics in the regression model was the F-Statistic. Our reduced model had a F-Statistic of 1.015e4 which means that it was very high. Due to this, the model was 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 multiple linear regression, there are a few limitations using this analysis. A big one is multiple linear regression is sensitive to outliers (Flom, 2019). Due to this, if the data is not fully prepared and cleaned and there are extreme outliers in the data used, there might be issues with the models. Another is that the predictor variables need to be independent. One variable must not rely or be impacted by another. These are just a few of the limitations that multiple linear regression has. However, with proper variable usage, and proper data cleaning and preparation, the limitations could be mitigated.

F2, Recommendations: Using multiple linear regression, we were able to create a reduced linear regression model that could help the company answer the question of what impacts Monthly Charge the most. Using this reduced model, the company can more easily predict how much a customer’s monthly charge will be given the 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, such as getting customers who stream TV and/or movies, etc.

G, Demonstration: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ef2f2085-dd99-441b-ac3f-afbe018001d4>

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

I, Sources:

*Assumptions of multiple linear regression*. Statistics Solutions. (2021, August 11). Retrieved March 4, 2023, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/

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

Flom, P. (2019, March 2). *The disadvantages of linear regression*. Sciencing. Retrieved March 4, 2023, from https://sciencing.com/disadvantages-linear-regression-8562780.html

Hayes, A. (2023, January 11). *Multiple linear regression (MLR) definition, formula, and example*. Investopedia. Retrieved March 4, 2023, from https://www.investopedia.com/terms/m/mlr.asp

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.

Zach. (2021, May 11). *How to interpret residual standard error*. Statology. Retrieved March 6, 2023, from https://www.statology.org/how-to-interpret-residual-standard-error/