What factors cause a customer’s length of Tenure?

By:

Kevin Sandoval

**A1.**

The research questions I will be examining is: “What factors cause a customer’s length of Tenure?” This question has important implications for a business as retaining customers is a vital part of the company’s success. Looking at Tenure length is a key metric in determining what could be important factors to increase the length of Tenure. I will be using multiple linear regression for this model because Tenure is a quantitative variable and will be the interesting dependent variable to examine.

**A2.**

The goal(s) of the analysis will be to examine a variety of explanatory variables to see if there is a significant relationship between them and the dependent variable Tenure. If I can find a potential explanation for a customer’s Tenure length, then I can make important decisions to hopefully help the company decide what actions to take to improve Tenure length. Thus, the overall goal for the analysis will be to find a multiple linear regression equation that can help predict the Tenure length of a customer.

**B1.**

Per Dr. Sewell’s video lecture “D208 Predictive Modeling Episode 2” there are a few assumptions for multiple linear regression, and they are listed below.

* There is a linear relationship between the dependent/target variable and the independent/explanatory variables.
* The independent variables are not too highly correlated with one another
* All observations are selected independently and randomly from the population
* Residuals should be normally distributed with a mean of 0
* Explanatory power should increase with an increase in the number of explanatory variables.

The first assumption is straightforward as we need to assume that there is some relationship between the target and explanatory variables for there to be an equation to help predict. The second assumption is making sure that none of the independent variables are strongly correlated with one another as there becomes the issue of multicollinearity. The third assumption is good practice for all statistics as independent and random observations are key to limiting bias. The fourth assumption checks that the expected values and actual values are connected in a meaningful way. The last assumption is that the model should get better as we add more explanatory variables as that should help explain the target variable.

**B2.**

For this project I chose to use Python. I personally find working with databases much easier in Python, as the code and manipulation is easier. I also preferred Python when it comes to troubleshooting and fixing any errors in code. The Python error is easy to follow and easy to address any issues. Although R is likely easier for the statistical analysis code specifically, I find that using Python and Jupyter Notebook makes it much easier to get visuals of the graphs or displays as needed.

**B3.**

Multiple linear regression is a good technique for answering the question “What factors cause a customer’s length of tenure?” because it works with a continuous variable as the target variable. I chose “Tenure” as my target variable because I wanted to examine the length of time a customer stays with the company, which can have vital information for the company’s profits. According to Dr. Middleton’s video lecture, “D208-Webinar: Getting Started with D208 Part I (November 2022)”, task 1 requires “any logical continuous variable” as the target variable. “Tenure” is a continuous variable and a logical choice for the target variable since it measures a customer’s length of time at the company. Because multiple linear regression is used when the target variable is continuous, it is the appropriate method to use.

**C1.**

The first thing that I did when I pulled in the data set was to check the shape of the data frame. The shape was (10000 , 50) so I knew that there were 10,000 rows of data with 50 columns. I then checked if there were any duplicate values and got that there were no duplicates.

After that I checked if there were any missing values using the .isna().sum() function to determine if there were any missing values. I found that the data set did not contain any missing values so none were treated. I then decided to check every quantitative variable for outliers. I did find outliers for the columns “Population”, “Children”, “Income”, “Outage\_sec\_perweek”, “Email”, “Contacts”, and “Yearly\_equip\_failure.” I treated the outliers by excluding them to a new variable for each except for “Population” which I decided to retain. After excluding the initial outliers, “Income” and “Outage\_sec\_perweek” still showed that they had a few new outliers, but I decided to retain those as it would be diminishing returns on the exclusion of data. My new data frame shape is (9079, 50).

After I checked the outliers for the quantitative variables, I decided to drop all the columns of the data that I would not be using. I kept 5 categorical variables to examine and they are “Techie”, “Multiple”, “StreamingTV”, “TechSupport”, and “OnlineSecurity.”

**C2.**

The descriptive statistics for the dependent variable and all of the independent variables are shown in a screenshot below.

**Tenure:**

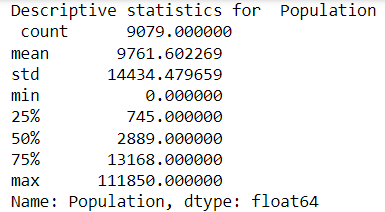
Tenure is the dependent variable that I will be analyzing in this report. The variable itself looks at the number of months the customer has stayed with the provider. The summary statistics for Tenure is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Population:**

Population is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the population count within a mile radius of the customer’s stated address. The summary statistics for Population is shown in the screenshot below.

****

**Children:**

Children is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of children a customer stated that they have. The summary statistics for Children is shown in the screenshot below.

**A screenshot of a computer code

Description automatically generated**

**Age:**

Age is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the customer’s age in years. The summary statistics for Age is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Income:**

Income is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the customer’s stated income per year. The summary statistics for Income is shown in the screenshot below.

**A white background with black text

Description automatically generated**

**Outage\_sec\_perweek:**

Outage\_sec\_perweek is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the average number of seconds per week of system outages in the customer’s neighborhood. The summary statistics for Outage\_sec\_perweek is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Contacts:**

Contacts is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of times a given customer contacted tech support. The summary statistics for Contacts is shown in the screenshot below.

**A white background with black text

Description automatically generated**

**Email:**

Email is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of times a given customer was emailed in the past year. The summary statistics for Email is shown in the screenshot below.

**A white background with black text

Description automatically generated**

**Yearly\_equip\_failure:**

Yearly\_equip\_failure is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of times a customer’s equipment failed and had to be reset or replaced in the past year. The summary statistics for Yearly\_equip\_failure is shown in the screenshot below.

**A screenshot of a computer code

Description automatically generated**

**Techie:**

Techie is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at if a customer considers themselves technically inclined. The summary statistics for Techie is shown in the screenshot below.

**A white background with black text

Description automatically generated**

We can see that the proportion of responses for the 9079 customers is as follows:

* Yes: 16.7%
* No: 83.3%

**Multiple:**

Multiple is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at if the customer has multiple lines. The summary statistics for Multiple is shown in the screenshot below.

**A white background with black text

Description automatically generated**

We can see that the proportion of responses for the 9079 customers is as follows:

* Yes: 46.24%
* No: 53.76%

**OnlineSecurity:**

OnlineSecurity is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at if the customer has an online security add-on. The summary statistics for OnlineSecurity is shown in the screenshot below.

**A white background with black text

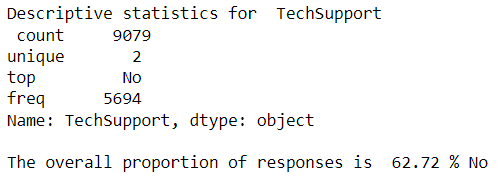
Description automatically generated**

We can see that the proportion of responses for the 9079 customers is as follows:

* Yes: 35.79%
* No: 64.21%

**TechSupport:**

TechSupport is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at whether the customer has a technical support add-on. The summary statistics for TechSupport is shown in the screenshot below.

****

We can see that the proportion of responses for the 9079 customers is as follows:

* Yes: 37.28%
* No: 62.72%

**MonthlyCharge:**

MonthlyCharge is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the average monthly charge that a customer receives. The summary statistics for MonthlyCharge is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**StreamingTV:**

StreamingTV is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at whether the customer has a streaming TV. The summary statistics for StreamingTV is shown in the screenshot below.

**A white background with black text

Description automatically generated**

We can see that the proportion of responses for the 9079 customers is as follows:

* Yes: 49.29%
* No: 50.71%

**Bandwidth\_GB\_Year:**

Bandwidth\_GB\_Year is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the average amount of data used, in GB, per year by the customer. The summary statistics for Bandwidth\_GB\_Year is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**C3.**

The univariate visualizations for each variable are shown on the left and the bivariate distributions for each independent variable with the dependent variable are shown on the right for each respective independent variable.

**Tenure:**

**A graph with blue bars

Description automatically generated**

**Population:**

**A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**Children:**

**A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**Age:**

**A graph of a number of people

Description automatically generatedA graph of a plot

Description automatically generated with medium confidence**

**Income:**

**A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**Outage\_sec\_perweek:**

**A graph with numbers and a number

Description automatically generated with medium confidenceA graph with blue dots

Description automatically generated**

**Contacts:**

**A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**Email:**

A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated with medium confidence

**Yearly\_equip\_failure:**

**A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**Techie:**

**A graph with a bar and a number

Description automatically generated with medium confidence** **A graph with blue and orange dots

Description automatically generated**

**Multiple:**

**A graph with blue bars

Description automatically generated** **A graph with multiple colored lines

Description automatically generated with medium confidence**

**OnlineSecurity:**

**A graph with blue rectangular bars

Description automatically generated** **A graph with blue and orange lines

Description automatically generated**

**TechSupport:**

**A graph with blue rectangular bars

Description automatically generated** **A graph with blue and orange squares

Description automatically generated**

**MonthlyCharge:**

**A graph of a graph

Description automatically generatedA graph of blue dots

Description automatically generated with medium confidence**

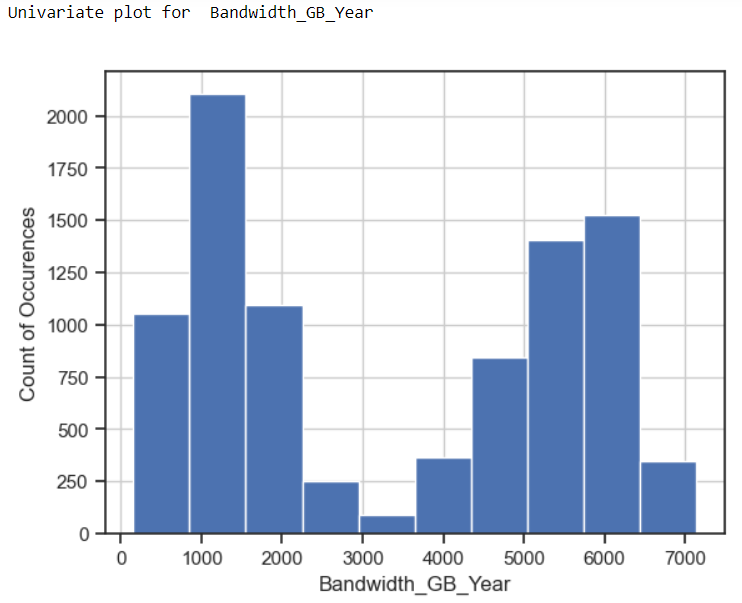
**StreamingTV:**

**A graph with blue bars

Description automatically generated** **A graph with blue and orange squares

Description automatically generated**

**Bandwidth\_GB\_Year:**

**A graph with dots on it

Description automatically generated**

**C4.**

As part of my data transformation goals I decided to change all of the categorical variables that are Yes/No responses to 1/0 responses which makes it easier for Python and the code to understand the variable. After those edits my data shape became (9079, 16). This helps align with my research question as it allows me to more efficiently check which factors will contribute to the Tenure length. Having categorical variables as 1’s and 0’s helps with the code to achieve that goal.

**C5.**

Prepared CSV is included in the submission.

**D1.**

Using all of the variables identified in C2, my initial multiple linear regression equation is as follows.

y = -1.1831 + 0.00000038 \*(Population) – 0.3712\*(Children) + 0.0393\*(Age) – 0.000002189\*(Income) + 0.021\*(Outage\_sec\_perweek) – 0.0201\*(Contacts) – 0.0018\*(Email) + 0.0114\*(Yearly\_equip\_failure) – 0.0308 \*(Techie) + 0.1338\*(Multiple) – 0.8407\*(OnlineSecurity) + 0.2504\*(TechSupport) – 0.0325\*(MonthlyCharge) – 1.4429\*(StreamingTV) + 0.0121\*(Bandwidth\_GB\_Year)

The y-intercept is -1.1831 and each coefficient shows the ‘weight’ of each variable. For example, for every 1 child that a customer has, their tenure length would decrease by 0.3712 months. Another example would be, for each GB used by the customer per year, their tenure length would increase by 0.0121 months. The OLS data output is shown in a screenshot below.

A screenshot of a computer

Description automatically generated

**D2.**

One of the notes given to me when I ran the initial model is that the condition number is large, which might indicate strong multicollinearity between some of the variables. To check for this, I will use the first model reduction method and I will check for the Variance Inflation Factor (VIF) and remove any variables that have a VIF greater than 10. The second step to improve the model through reduction that I decided to enact was backwards stepwise elimination. This would remove the variable with the largest p-value greater than 0.05 and then re-check the model and remove the next largest p-value. Since the p-values are large, it implies that the variable is not significant in explaining the target variable.

**D3.**

Through my initial look, I found that there were 3 variables that had a VIF greater than 10 as seen in the screenshot below.

A screenshot of a computer

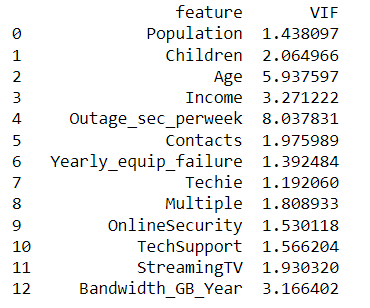
Description automatically generated

Since “MonthlyCharge” has the highest VIF, I will remove and test for the VIF again. The output is shown below.

A screenshot of a computer

Description automatically generated

As we can see, “MonthlyCharge” is no longer a feature in the analysis but “Email” still has a VIF over 10, so I will choose to remove it as well. The output is shown below.

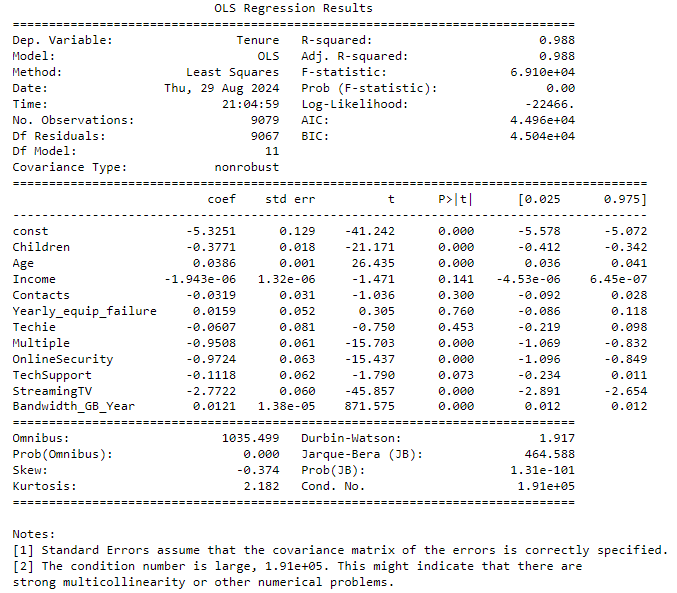


Although I initially was checking for a VIF greater than 10, “Outage\_sec\_perweek” is a clear max for this data. I decided to remove it from the analysis as well, as there still suggest some multicollinearity issues with that variable. The output is shown below.

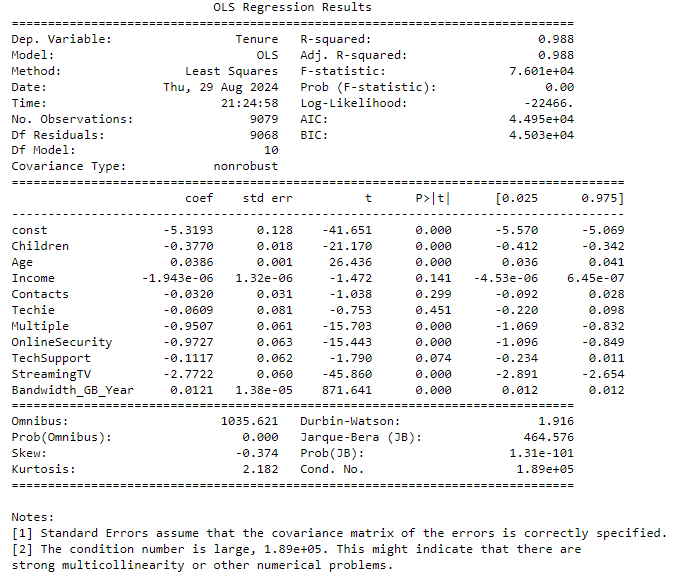
A screenshot of a computer

Description automatically generated

Another reason I was willing to omit “Outage\_sec\_perweek” is that the P-value shows as not statistically significant. This segways into another model reduction that I chose to do, which was Backwards Stepwise Elimination. I removed each variable with the highest P-value and then re-ran the model calculation. The first variable to be removed was “Population.” The output is shown below.



According to the OLS output here, the next variable to remove is “Yearly\_equip\_failure.” The output after removing is shown below.



According to the OLS output here, the next variable to remove is “Techie”. The output after removing is shown below.

A screenshot of a computer

Description automatically generated

According to the OLS output here, the next variable to remove is “Contacts”. The output after removing is shown below.

A screenshot of a computer

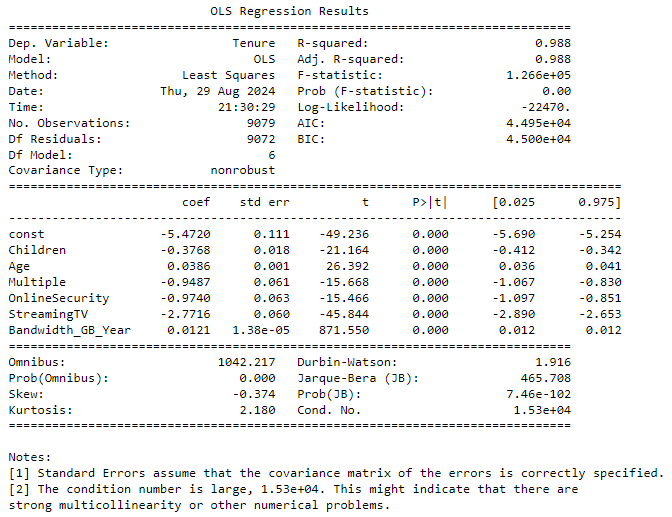
Description automatically generated

According to the OLS output here, the next variable to remove is “Income”. The output after removing is shown below.

A screenshot of a computer

Description automatically generated

According to the OLS output here, the last variable that has a P-value greater than 0.05 is “TechSupport”. The output after removing “TechSupport” is shown below.



The final, reduced model equation is as shown below.

y = -5.472 – 0.3768\*(Children) + 0.0386\*(Age) – 0.9487\*(Multiple) – 0.974\*(OnlineSecurity) – 2.7716 \*(StreamingTV) + 0.0121\*(Bandwidth\_GB\_Year)

The coefficients for each variable show the ‘weight’ that each variable has. For example, for every year older a customer is, their tenure length will increase by 0.0386 months. Since there are categorical variables still in the reduced model, an example of their interpretation is as follows. If a person answers “Yes” to have a streaming TV, their tenure will be 2.7716 months less on average than someone who answered “No”.

**E1.**

From the initial model’s OLS output, we can see a few things that can help decide the effectiveness of the model. I will bullet point all the important statistics that we have for each model and then compare.

**Initial Model**

* **R2:** 0.99. This means that 99% of the variation in y is explained by the explanatory variables. This number increases with the number of predictor variables so it has a significant drawback.
* **Adjusted R2:** 0.99 This is similar to the R2 value but it only increases when an additional carriable adds to the explanatory power
* **F-statistic:** 59080 This shows that the group of variables are jointly significant
* **Prob(F-statistic):** 0.00 This shows the overall significance of the regression. Since the value is below 0.05, it implies that overall the regression is meaningful.
* **AIC:** 43580 This is a large AIC and it implies that the model can definitely be fitted much better.
* **BIC:** 43700 Similar to AIC, a lower number is preferred and suggests a better fit. Since this is a large number, it implies that there could be a model with a better fit.
* **Residual Standard Error:** 2.66536 This shows the error associated with the prediction. This number is rather large for the residual error so there should be a model that can fit better.

**Reduced Model**

* **R2:** 0.988 This means that 98.8% of the variation in y is explained by the explanatory variables. This number increases with the number of predictor variables so it has a significant drawback.
* **Adjusted R2:** 0.988 This is similar to the R2 value but it only increases when an additional carriable adds to the explanatory power
* **F-statistic:** 126600 This shows that the group of variables are jointly significant
* **Prob(F-statistic):** 0.00 This shows the overall significance of the regression. Since the value is below 0.05, it implies that overall the regression is meaningful.
* **AIC:** 44950 This is a large AIC and it implies that the model can definitely be fitted much better.
* **BIC:** 45000 Similar to AIC, a lower number is preferred and suggests a better fit. Since this is a large number, it implies that there could be a model with a better fit.
* **Residual Standard Error:** 2.87589 This shows the error associated with the prediction. This number is still large so the model may not fit the data well

Upon seeing these values we can see that the reduced model performs roughly as well, if not worse than the initial model in every comparison. Although the R2 values for both the initial and reduced model are very high, every other metric suggests that the model does not fit the data set well. The reduced model only succeeded the initial model in the condition number reducing. That is to be expected as I checked the VIF scores to help with multicollinearity. Overall there is not much significant difference between the two models.

**E2.**

The residual for the reduced model is shown in the screenshot below.

A white box with black text

Description automatically generated

The residual plots for each variable are shown below respectively.

**Children:**

**A graph of different sizes and colors

Description automatically generated with medium confidence**

**Age:**

**A group of blue and red graphs

Description automatically generated**

**Multiple:**

**A group of graphs on a white background

Description automatically generated**

**Onlinesecurity:**

**A graph of different colored lines

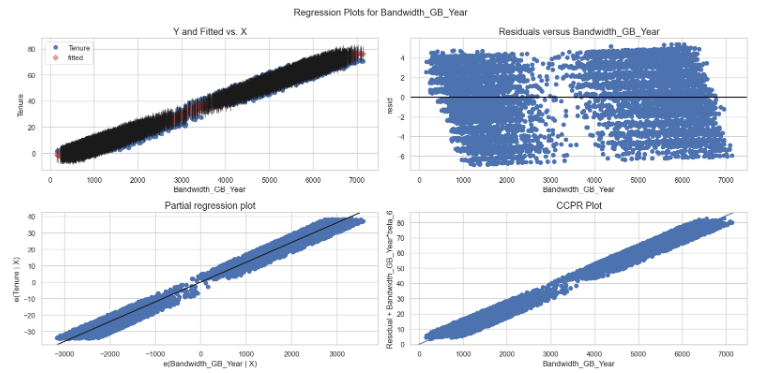
Description automatically generated with medium confidence**

**StreamingTV:**

**A graph of different types of graphs

Description automatically generated with medium confidence**

**Bandwidth\_GB\_Year:**

****

**E3.**

The Jupyter Notebook file used to perform the analysis is provided in the submission.

**F1.**

The final, reduced model equation is as shown below.

y = -5.472 – 0.3768\*(Children) + 0.0386\*(Age) – 0.9487\*(Multiple) – 0.974\*(OnlineSecurity) – 2.7716 \*(StreamingTV) + 0.0121\*(Bandwidth\_GB\_Year)

The coefficients for each variable show the ‘weight’ that each variable has on the model. For example, for every year older a customer is, their tenure length will increase by 0.0386 months. “Multiple”, “OnlineSecurity”, and “StreamingTV” are all Yes/No responses re-valued to be 1/0. So for the interpretation of their respective coefficients, a 0 would add no change and a 1 would add whatever the coefficient was respectively.

The only statistically significant outcome of this analysis are the R2 values received in the OLS report. Both models showed that the R2 values are high enough to help explain the variability in the outcome of Tenure. However, the AIC, BIC, and residual standard error all suggest issues with the model. The model is not statistically significant at accurately predicting a customer’s tenure length and thus can be difficult to make assumptions from. Despite the lack of statistical significance in the model, I do believe there are some practical benefits to the model analysis. One practical significance could be looking at the graphs and correlation for “Bandwidth\_GB\_Year” and “Tenure”. Clearly from the graphs there is a large correlation between the two variables, and thus it could provide practical benefit to look more closely at the large bandwidth users. Although the predictive power of the model is limited, it also suggests that customers that answer “Yes” to a few of the questions imply that they are more likely to not stay as long as their counterpart “No” responses. These practical benefits would need more analysis to confirm, but there is a clear direction that could be taken.

There are some limitations to the analysis. For one, there could be more variables in the data set that could help contribute to determining Tenure. I chose all the quantitative variables and 5 categorical variables as I used predominantly categorical variables in D207 and wanted to change what I was looking for. This is somewhat arbitrary and thus more robust or better predictive models could potentially be made by checking all independent variables and then reducing the model.

**F2.**

A course of action that I would recommend to the prospective company would be to focus heavily on the retention of the high bandwidth users. From the model, we can see that “Bandwidth\_GB\_Year” was by far the strongest variable when it comes to predicting the Tenure of a customer, so if there are customers that have high bandwidth usage, it is likely that they will stay in their current plan. Another interesting note is that the people who have an online security background, multiple lines, and a streaming TV are more likely to leave. I believe that this implies that more technically literate people are more likely to try and find other options around them, as opposed to less technically inclined people who are more likely to stay as their usage doesn’t affect them as much personally. Older customers also tend to stay in their plan longer, so focusing on them could help increase the retention rate as well. This is all under the assumption that the model could help explain the variability in “Tenure” as shown by the significant value of R2. Another course of action would be to re-evaluate the model with different independent variables to try and find a model that might be more statistically significant for the data set.

**G.**

Panopto video provided in the submission.

**H/I:**

Bobbitt, Zach. (July 21, 2020). *How to Create a Residual Plot in Python.* Retrieved August 27th, 2024,From <https://www.statology.org/residual-plot-python/>

Bobbitt, Zach. (July 24, 2020). *A Complete Guide to Linear Regression in Python.* Retrieved August 27th, 2024,From <https://www.statology.org/linear-regression-python/>

Data Viz Talk. (February 11th, 2016). *How to add jitter to a plot using Python’s matplotlib and seaborn.* Retrieved August 30th, 2024 From <https://dataviztalk.blogspot.com/2016/02/how-to-add-jitter-to-plot-using-pythons.html>

matplotlib. (n.d). *matplotlib.pyplot.hist.* Retrieved August 30th, 2024 From <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html>

Dr. Middleton, W (n.d). *Getting Started with D208 Part 1.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Middleton, W (n.d). *Getting Started with D208 Part 2.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 1.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 2.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 3.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 4.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 5.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 6.* Retrieved August 17th, 2024,From D208 Student Resources Folder