What independent variables can help predict the probability a customer will churn?

By:

Kevin Sandoval

**A1.**

The research questions I will be examining is: “What independent variables can help predict the probability a customer will churn?” This question has important implications for a business as retaining customers is a vital part of the company’s success. Examining the variables that might help predict the probability a given customer is likely to churn could either aide in retaining the customer, or focusing more on other customers. Acquiring new customers is much more expensive than retaining old customers. I will use logistic regression for this prediction as I am looking at the “Churn” variable which is a Yes/No response.

**A2.**

The goal(s) of the analysis will be to examine a variety of explanatory variables to see if they can help predict if a customer will churn. The model will examine the probability that a customer will leave the company and a good model will help with customer retention within the company.

**B1.**

Per Dr. Sewell’s video lecture “D208 Predictive Modeling Episode 3” there are a few assumptions for logistic regression, and they are listed below.

* The predicted values are restricted to a range of nominal values like “Yes”, “No”
* It predicts the probability of the outcome rather than the outcome itself
* Requires the observations to be independent of each other
* Assumes linearity of independent variables and log odds

The first assumption is straightforward as we need to assume the dependent variable has a binary outcome to predict the probability of. The second assumption is clarifying that logistic regression predicts the probability of something rather than that specific outcome. The third assumption is checking that the independent variables are independent of one another. This is crucial as we want as little multicollinearity as possible. The fourth assumption assumes that the independent variables are linearly related to the log odds of the outcome.

**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.**

Logistic regression is a good technique for answering the question “What independent variables can help predict the probability a customer will churn?” because it works with a categorical, two outcome variable as the target variable. I chose “Churn” as my target variable because I wanted to examine the probability of a customer leaving 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 categorical variable” as the target variable. “Churn” is a categorical variable and a logical choice for the target variable since it measures if a customer left the company in the last month. Because logistic regression is used when the target variable is categorical, 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.

**Churn:**

Churn is the dependent variable that I will be analyzing in this report. The variable itself looks at if the customer left the company in the last month. The summary statistics for Churn is shown in the screenshot below.

**A white background with black text

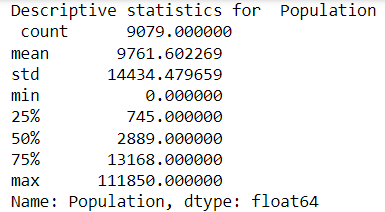
Description automatically generated**

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

* Yes: 26.58%
* No: 73.42%

**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

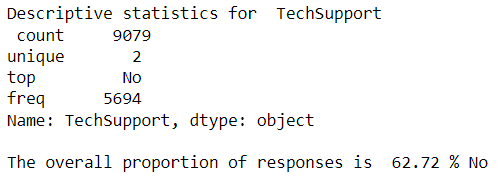
Description automatically generated**

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.

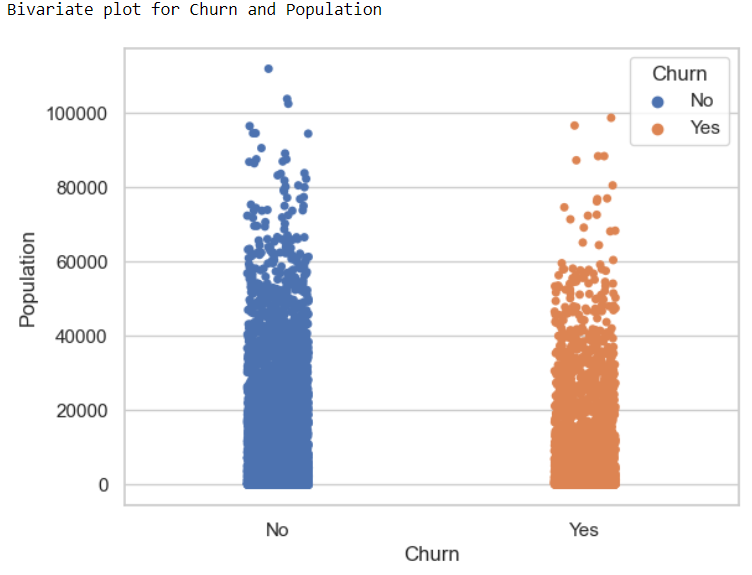
**Churn:**

**A graph with blue squares

Description automatically generated**

**Population:**

**A graph with blue bars

Description automatically generated** ****

**Children:**

**A graph with blue bars

Description automatically generated** **A graph of different colored bars

Description automatically generated**

Although “Children” is a quantitative variable, I felt that the bivariate plot for it with “Churn” would be better represented as a stacked bar chart as opposed to the stripplot like “Population.”

**Age:**

**A graph of a number of people

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

Description automatically generated**

**Income:**

**A graph with blue bars

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

Description automatically generated**

**Outage\_sec\_perweek:**

**A graph with numbers and a number

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

Description automatically generated**

**Contacts:**

**A graph with blue bars

Description automatically generated** **A graph of different colored squares

Description automatically generated**

Although “Contacts” is a quantitative variable, I felt that the bivariate plot for it with “Churn” would be better represented as a stacked bar chart as opposed to the stripplot like “Population.”

**Email:**

A graph with blue bars

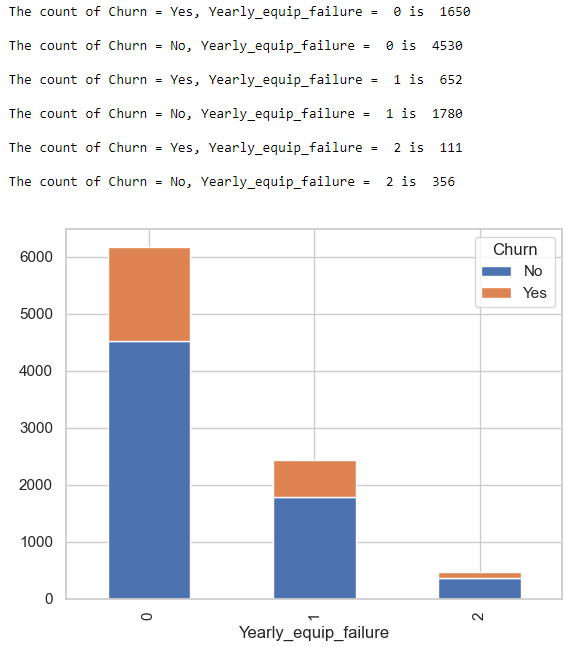
Description automatically generated A graph with numbers and a chart

Description automatically generated

Although “Email” is a quantitative variable, I felt that the bivariate plot for it with “Churn” would be better represented as a stacked bar chart as opposed to the stripplot like “Population.”

**Yearly\_equip\_failure:**

**A graph with blue bars

Description automatically generated** ****

**Techie:**

**A graph with a bar and a number

Description automatically generated with medium confidence** A screenshot of a graph

Description automatically generated

**Multiple:**

**A graph with blue bars

Description automatically generated** A screenshot of a graph

Description automatically generated

**OnlineSecurity:**

**A graph with blue rectangular bars

Description automatically generated** A screenshot of a graph

Description automatically generated

**TechSupport:**

**A graph with blue rectangular bars

Description automatically generated** A screenshot of a graph

Description automatically generated

**MonthlyCharge:**

**A graph of a graph

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

Description automatically generated**

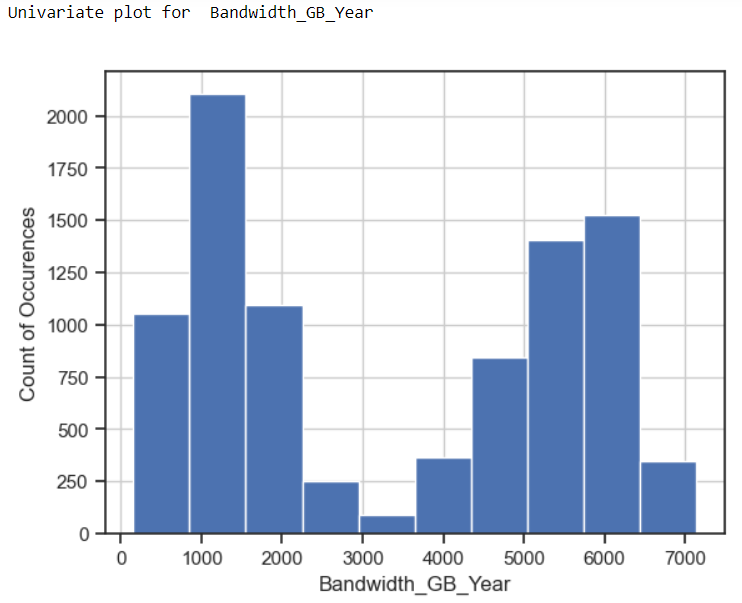
**StreamingTV:**

**A graph with blue bars

Description automatically generated** A screenshot of a graph

Description automatically generated

**Bandwidth\_GB\_Year:**

**** **A graph with blue and orange dots

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 effectively check which independent variables are impactful in helping predict the probability a customer will churn. 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 = 0.0 - 0.000005\* (Population) – 0.028464\*(Children) - 0.011484\*(Age) – 0.000005\*(Income) - 0.101245\*(Outage\_sec\_perweek) – 0.041012\*(Contacts) – 0.100992\*(Email) - 0.118986\*(Yearly\_equip\_failure) + 0.393319\*(Techie) + 0.112338\*(Multiple) – 0.274485\*(OnlineSecurity) - 0.253504\*(TechSupport) + 0.022592\*(MonthlyCharge) +0.816317\*(StreamingTV) - 0.000821\*(Bandwidth\_GB\_Year)

The y-intercept is 0.5 as we are looking at the increase in log odds that a variable can have on the dependent variable. The coefficients for each variable are difficult to interpret as they are, but if we use e as the base and the coefficients as the power, we can obtain the log odds change in the odds of a customer churning. I then did the result, minus 1, times 100 to obtain the percentage change for each variable. The percentage change is shown below.



If we match the corresponding variable names to the log odds shown above, we can find an interpretation for the effect. For example, responding “Yes” to have a StreamingTV (second to last in the list), will result in a 126.22% increase in the odds that someone will churn.

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

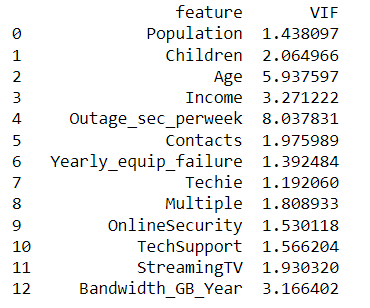
Description automatically generated

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.



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 “Income.” The output is shown below.

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

According to the Logit output here, the last variable that has a P-value greater than 0.05 is “Yearly\_equip\_failure”. The output after removing is shown below.

A screenshot of a computer

Description automatically generated

The final, reduced model equation is as shown below.

y = 0.5– 0.0043 \* Age – 0.0312\* Outage\_sec\_perweek + 0.4263 \* Techie + 0.8378 \* Multiple + 1.6232 \* StreamingTV – 0.0007 \* Bandwidth\_GB\_Year

The percentage of change for the reduced model is shown below.



Thus we can now see that a unit increase in Outage\_sec\_perweek has a 3.08% decrease in the odds that someone will churn, while answering “Yes” to StreamingTV is a 406.94% increase in the odds that they will churn.

**E1.**

From the initial model’s Logit 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**

* **Pseudo R-squared (McFadden):** 0.359. Per Dr. Sewell’s video lecture “D208.Ep.6.v” at around minute 14 he states that a McFadden number greater than 0.30 (30%) is an indicator of a good model. We can see from the logit results output that the initial model passes this threshold.
* **AIC:** 5429.8517. This is a large number for AIC which could suggest a bad fit for the data. It is useful to compare to the reduced model.

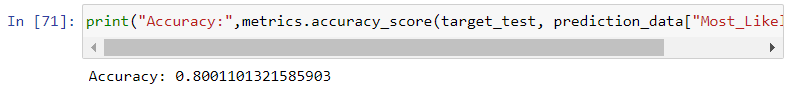
**Reduced Model**

* **Pseudo R-squared (McFadden):** 0.284. We can see that this is a significant step down from the initial model. It also does not pass the 0.3 threshold which indicates the reduced model is not as good of a fit as the initial.
* **AIC:** 6044.8018. This is a large number for AIC which could suggest a bad fit for the data. It is larger than the reduced model which could indicate a slightly worse fit for the data

Upon seeing these values we can see that the reduced model performs worse than the initial model. The AIC is higher for the reduced model and the McFadden number is significantly lower. I think that since the reduced model does not pass the 0.3 threshold for a good model per Dr. Sewell, the initial model would be a better fit for the data.

**E2.**

The accuracy score for the reduced model is shown below.



The accuracy rating for the reduced model is 0.8001 compared to 0.8216 for the initial. The confusion matrix for the reduced model is shown below.

A yellow and purple squares with numbers

Description automatically generated

**E3.**

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

**F1.**

The final, reduced model equation with the adjusted coefficients is as shown below.

y = 0.5– 0.0043 \* Age – 0.0312\* Outage\_sec\_perweek + 0.4263 \* Techie + 0.8378 \* Multiple + 1.6232 \* StreamingTV – 0.0007 \* Bandwidth\_GB\_Year



Thus we can now see that a unit increase in Age will result in a 0.43% decrease in the odds of a customer churning, a unit increase in Outage\_sec\_perweek has a 3.08% decrease in the odds, answering “Yes” to Techie has a 53.15% increase in the odds someone will churn, answering “Yes” to Multiple has a 131.12% increase in the odds someone will churn, answering “Yes” to StreamingTV is a 406.94% increase in the odds that they will churn, and a unit increase in their Bandwidth\_GB\_Year will have a 0.07% decrease in the odds that a customer will churn.

The initial model has a lower AIC than the reduced model, which could imply that the initial model is a better fit for the data. There is not much statistical significance in either model, but the initial model at least passes the 0.30 threshold for the McFadden number, which the reduced model does not. 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 response for the StreamingTV variable. In both the initial and reduced model, StreamingTV had a significant impact on the increase in odds of a customer churning. Techie and Multiple also have a large increase in the odds as well.

There are some limitations to the analysis. For one, there could be more variables in the data set that could help contribute to determining Churn. 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 customers who responded “No” towards Yes/No questions aimed at figuring out their confidence in technology. From the models, we can see the “Techie”, “Multiple”, and “StreamingTV” all see significant increases in the log odds of someone churning. Since all of these variables are indicative of how technically literate a customer is, it could be important to examine these responses more closely. 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. 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:**

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

Rkabacoff (n.d) *Chapter 5 Bivariate Graphs* Retrieved September 12th, 2024, From  
<https://rkabacoff.github.io/datavis/Bivariate.html#categorical-vs.-quantitative>

Scikit-learn. (n.d). *ConfusionMatrixDisplay.* Retrieved September 13th, 2024,From <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html>

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

Statsmodels. (Sept. 12th, 2024) *statsmodels.discrete.discrete\_model.Logit.predict*Retrieved September 9th, 2024 From:

<https://www.statsmodels.org/dev/generated/statsmodels.discrete.discrete_model.Logit.predict.html#statsmodels.discrete.discrete_model.Logit.predict-parameters>

Van den Broeck, M. (n.d). *D208 – Predictive Modeling.* Multiple logistic regression From: <https://campus.datacamp.com/courses/intermediate-regression-with-statsmodels-in-python/multiple-logistic-regression-4?ex=1>