CLASSIFICATION ANALYSIS – D209

Gooden, Nina S. Gooden [ID #: 009823504]

Dr. Eric Straw

CLASSIFICATION ANALYSIS – D209

This analysis will explore the medical data of a theoretical real-world organization. By creating this logistic regression, I will address the business’ concern about readmission in a data-first evaluation. I will also provide visualizations to support my assessments, in tandem with code for my models. As previously requested, I will also discuss the limitations and potential course of action this data supports.

## **A. The Research Question:**

1. Research Question
   1. Using the *k*-nearest neighbor method, can we identify patients likely to contribute to readmission based on comorbidity factors such as age, gender, and noted health concerns?
2. Data Analysis Goals and Objective
   1. I will be reassessing the model I created in D208, with additional tools provided from this course. This evaluation is to further explore variables that were identified as having correlation with readmission. In this evaluation, we will look at variables that have been identified as patient notes—age, sex, and reported comorbidities—through a logistic regression in order to measure likelihood of readmission given said variables.

## **B. Justification for Method**

1. How KNN analyses the selected data set and hypothesized outcome:
   * 1. KNN finds the distance between my given parameter and all the available data, then selects the variables (k) closest to the query. Next, for classification, it votes for the most frequent label, as long as they are continuous.
     2. Once an algorithm is established, it can be trained and matched to the outcome of the primary variable—in this case, Yes/No for readmission. From this training step we gain a model accuracy score.
     3. It is my expectation that this classification will have an accuracy of at least 95%, as classification is a good fit for this question. I do not expect under- or overfitting.
2. KNN assumption summary:
   1. KNN assumes that each training data consists of variables and class levels that are associated with one another. “In the simplest case, it will be either + or – (for positive or negative classes).” (*A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm*, 2010)
3. Packages and libraries:
   1. I’ve opted to use Python for my analysis as this is the language, I am most comfortable with. Python offers packages and libraries that make visualizations and analysis easy and straightforward. I will be using pandas, Numpy, Matplotlib, and Seaborn in this evaluation, all of which are uniquely designed for data analysis and visualization. In addition, I will also be using scikit-learn, for splitting, training, testing, and fitting my data. This package is uniquely suited for predicting and classifying data in measurable models.
4. Why Logistic Regression is Appropriate:
   1. After going through the list of assumptions necessary to use this model, I have deemed my research question a good fit. We are looking at the ReAdmis variable as our constant for this evaluation and it is a categorical variable. In addition, the patient note variables are independent of one another and since we are not looking for a specific value, the model fits.

## **C. Data Preperation Process**

1. Data Preparation Goals and Manipulations
   1. A clean, workable data set is necessary in order to run an analysis that is not skewed. After importing my packages and data, I began by dropping the columns I wouldn’t need for my analysis.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Next, I needed to make sure there were no missing or null values to adjust.

A screenshot of a computer

Description automatically generated with medium confidence

A computer screen capture

Description automatically generated with medium confidence

* 1. Lastly, I ran a quick check to make sure my variables were classified correctly, since I’d identified the continuous variables already. I ran one more check to make sure I didn’t have any null values.

A screenshot of a computer

Description automatically generated

1. Summary Statistics:
   1. Next, I used .mean()—rounded to the 2nd place—and .describe() to look at some of the summary statistics for the continuous columns, since I was going to look at the categorical information via histograms.
      1. Target variable: ReAdmis
      2. Predictor variables: Age, Gender, VitD\_levels, HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, and Asthma.
   2. Looking at the mean data, we can see that the means aren’t entirely equal, though VitD\_levels are closer. Still, they’re not too far off so it’s worth continuing to evaluate the variables.
   3. From the .describe() data, we get a better idea of whether or not the variables are normally distributed. I wasn’t too concerned because the mean data is so close, but this looking at the STDs, we can see that Age and VitD\_Levels are normally distributed.

A screenshot of a computer

Description automatically generated with medium confidence

1. Data Preparation:
   1. Steps and code have been provided via screenshots in C1 and C2.
2. Univariate and bivariate visualizations:
   1. First, I looked at the distribution for the target variable. The distribution isn’t equal here, but we knew that would be the case—hence the need for the evaluation in the first place.

Graphical user interface

Description automatically generated

* 1. Both Categorical and Continuous variables were considered in my assessments, beginning with **the Univariate visualizations**:
  2. I created histograms for the continuous data.

Graphical user interface

Description automatically generated

|  |  |
| --- | --- |
|  |  |

* 1. Next, I created visualizations for the categorical data.

Graphical user interface

Description automatically generated

Diagram, polygon

Description automatically generated

* 1. Followed by bivariate analyses to compare the categorical variables and the target variable. Both Categorical and Continuous variables were considered in my assessments, concluding with **the Bivariate visualizations**:

A screenshot of a computer

Description automatically generated

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart, bar chart  Description automatically generated |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

* 1. Lastly, I encoded the ReAdmis variable in order to run scatterplots for the continuous variables.

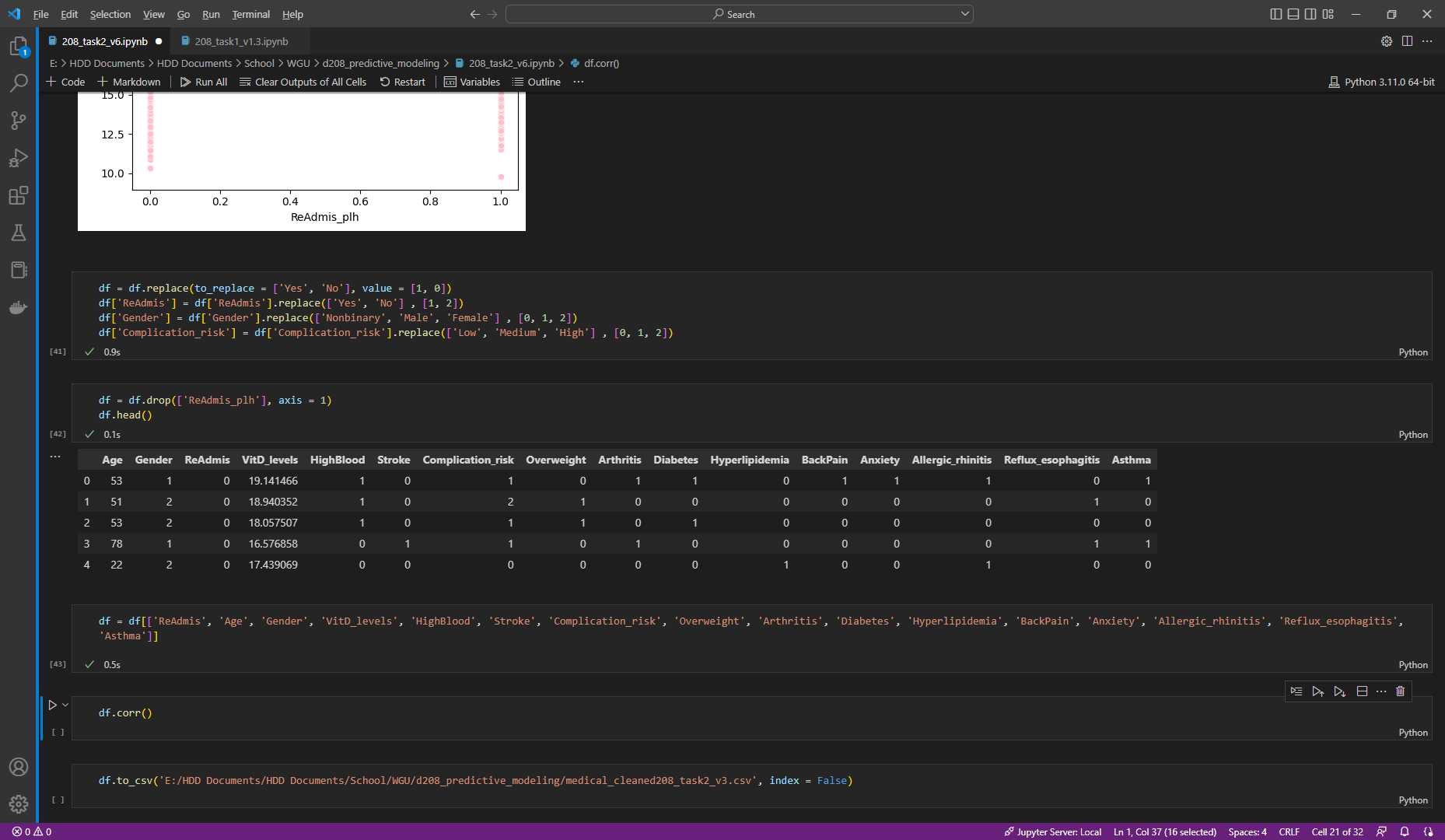
A screenshot of a computer

Description automatically generated

Graphical user interface

Description automatically generated

* 1. As I was looking at the univariate and bivariate visualizations, my primary concern was how uneven the data was looking. I encoded the last of the variables in order to run the logit and created a correlation matrix on the cleaned and reorganized data.



Graphical user interface, application

Description automatically generated

1. Provide prepared data set:

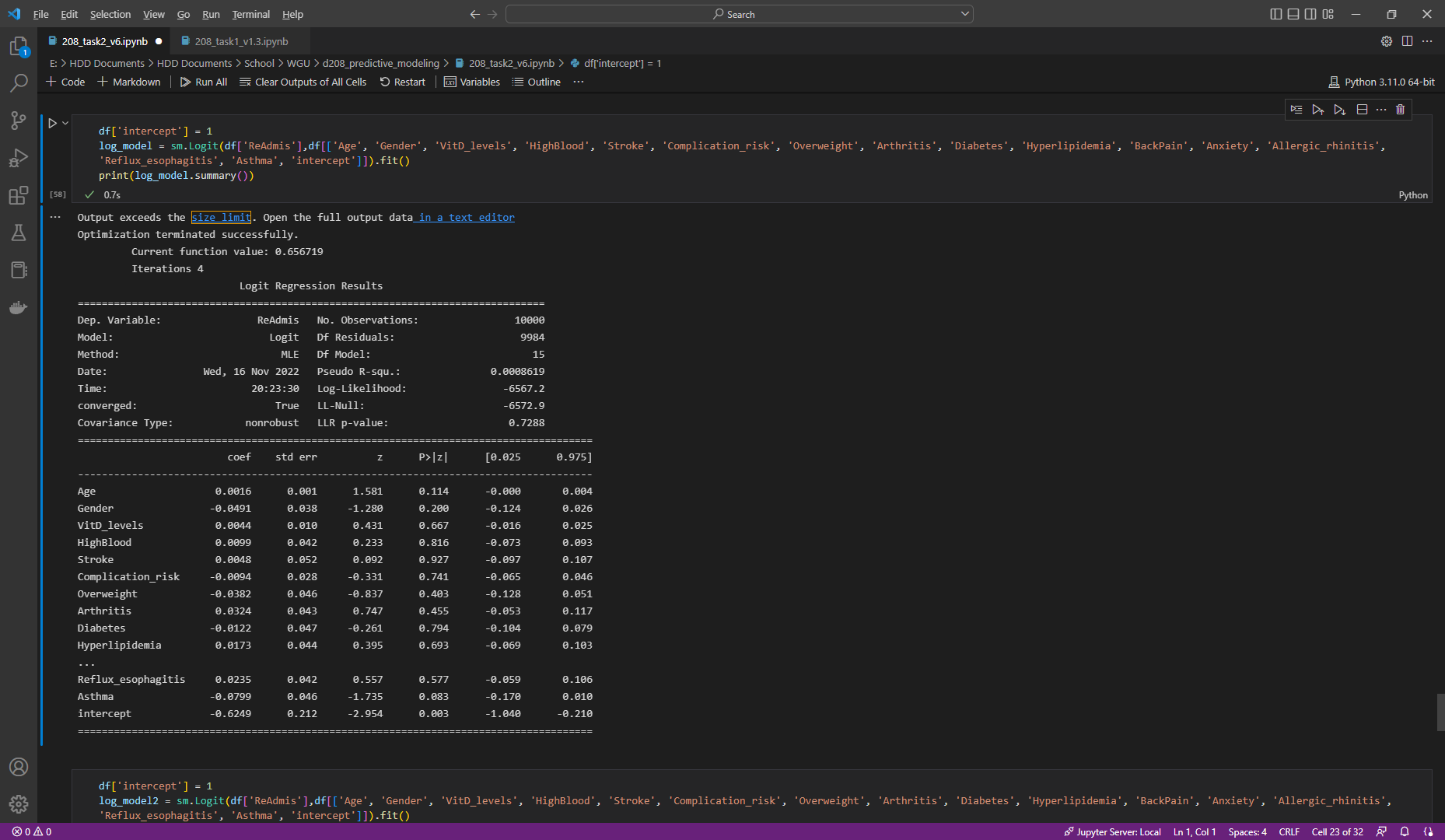
A picture containing text

Description automatically generated

Data provided.

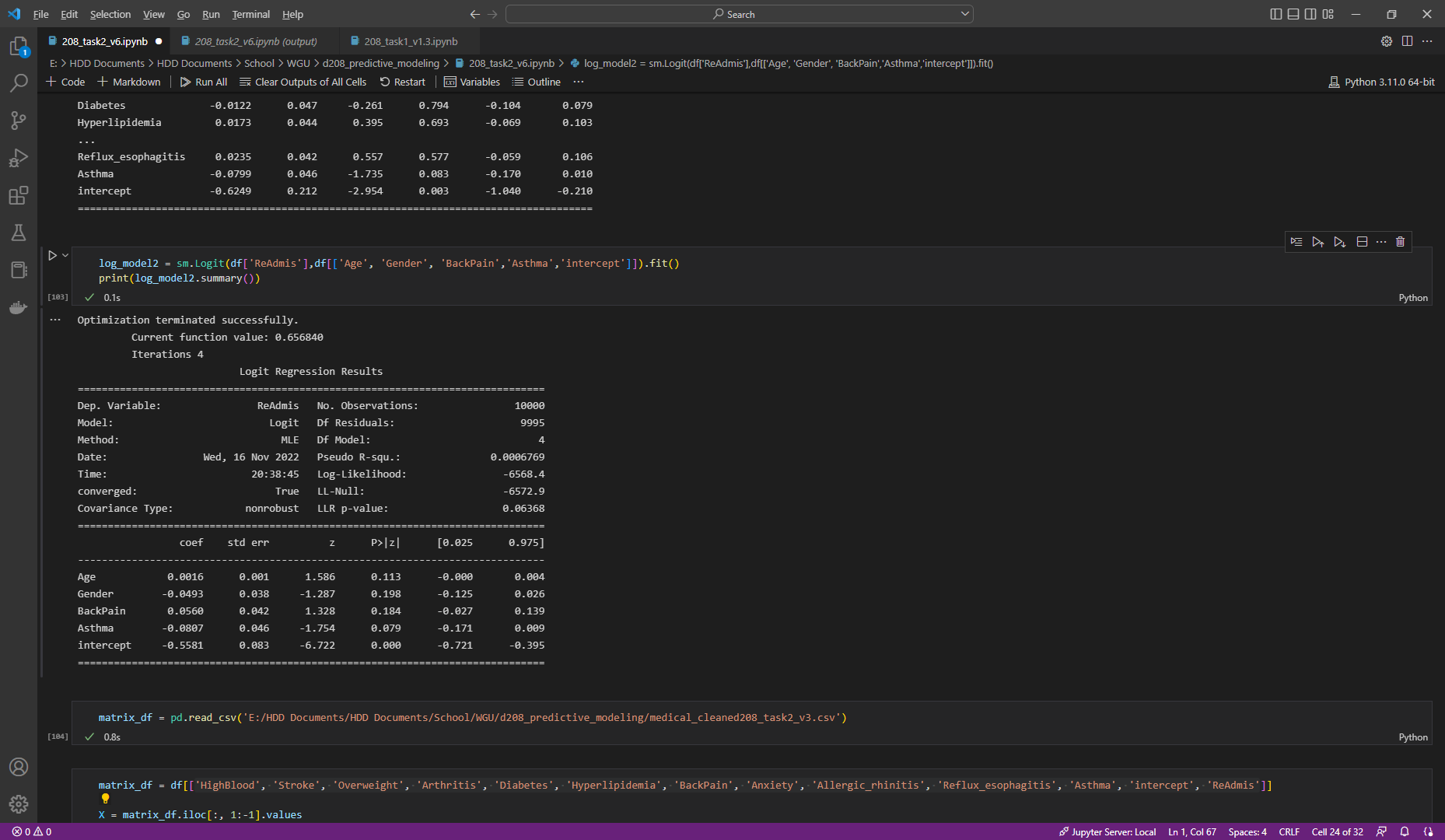
## **D. MODEL COMP AND ANALYSIS**

1. Comp initial and reduced multiple regression model:
   1. I used Logit from statsmodel to run an initial regression on all of the identified variables from the cleaned-up data set. As I was noting in the previous analyses, the data here is not at all balanced—it only has a Pseudo R-squared value of 0.0009.

 A screenshot of a computer

Description automatically generated with medium confidence

* 1. After looking at the P-values for the available data, I wasn’t left with enough variables to run a sound reduced model. I pulled variables with less than 0.2 values and was left with Age, Gender, BackPain, Asthma, and the intercept. Running a model with these values returned even less confidence in the model, leading me to conclude that the data was too skewed to evaluate.



## **E. Data Set Analysis**

1. Explaining data analysis process
   1. I created a confusion matrix using the original variables because the Pseudo R-squared value was slightly higher than the original model.
   2. I opted to use the LogisticRegression package because it was covered in the source materials. I also considered using SMOTE, but the set up was outside of the scope for this assignment.

Graphical user interface, text

Description automatically generated

* 1. Unfortunately, the matrix only confirms the lack of balance in the model. **This brings me to the conclusion that there are dependencies in the data that cannot be explained by just the comorbidities listed in my research question.**

Graphical user interface, application

Description automatically generated

* 1. Despite this finding, I wanted to put together one more matrix to compare. I printed out the classification report for this model to take one more look at the F-score. The confusion matrix already highlighted the disproportionate data, but due to the Precision and F-scores that printed, (0% for ReAmis = 1) I can assert that this model does not predict a logical regression.

Graphical user interface, text

Description automatically generated

## **F. Data Summary and Implications**

1. Discuss the results of analysis:
   1. The regression equation for my reduced model is:
      1. -3.203(0)+470.4 = -4.046.
   2. The confusion matrix has a true positives value of 1300 and a false negatives value of 740. This means that the probability of receiving both a true positive and a false negative equate to an accuracy of ~63%, which is well below the confidence level of 99% that we would like to see.
   3. accuracy = (TP + TN) / (TP + FN + FP + TN)
   4. The technique I used for variable selection is explained in my step-by-step explanations. While searching for answers to my research question, I quickly hit a wall when it became obvious through the correlation matrix that I wasn’t going to find connection between patient reviews and readmission likelihood. I continued to explore this avenue against the Initial\_days variable. The residual plot for the model is ultimately the same as the original, since there was no discernable connection between the two tested variables.
   5. **The most statistically significant variable was Asthma, based on a negative coef less than – 0.8 and a p-value of 0.079.**
   6. **Age and BackPain had a slight, but not statistically significant impact on readmission.**
2. Output and calculations are included in the code section of this communication.
3. Code is included in the code section of this communication.

## **F. DATA/Implications Summary**

1. Evaluation results:
   1. This analysis was ultimately unbalanced, which impacts its statistical and practical significance. Asthma was found to be a predictor that addresses the probability of a patient returning to the hospital chain. Based on this evaluation, the medical center can closely treat patients with this comorbidity.
   2. This means that in all likelihood, the data was insufficient to answer the research question as it was presented. As such, the evaluation as a whole will likely need to be reassessed for additional data.
   3. Comorbidities, by definition, can overlap with invisible factors to continue to skew data. The primary takeaway from this evaluation should be that additional data points need to be audited in order to better grasp the extent of impact.
   4. Despite these concerns, Asthma was still found to have significant variable.
2. Recommendations:
   1. This medical center should spend resources to house data in a particular way that better designates comorbidity information.
   2. In order to maintain the health of this analysis, the medical center should launch campaigns to gather as much comorbidity data as possible. I would also recommend keeping track of the variation KPIs in order to segment response efforts.
   3. This analysis should be run quarterly in order to ensure the value of this assessment.

## **G. Panopto**

[Logistic Regression Modeling – NBM2 | D208](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f9d578fd-90d3-4f2b-bf6f-af50004d5f17)

## **H. Reference web sources**

*A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm*. (2010, May 18). God, Your Book Is Great !! https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/

## **G. Acknowledge sources**

Brownlee, Jason. “4 Automatic Outlier Detection Algorithms in Python - MachineLearningMastery.com.” *MachineLearningMastery.com*, 7 July 2020, machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python.

jssuriyakumar. “How to Show Values on Seaborn Barplot? - GeeksforGeeks.” *GeeksforGeeks*, 2 Jan. 2022, www.geeksforgeeks.org/how-to-show-values-on-seaborn-barplot.

“Problems Importing Imblearn Python Package on Ipython Notebook.” *Stack Overflow*, 12 Oct. 2016, stackoverflow.com/questions/40008015/problems-importing-imblearn-python-package-on-ipython-notebook.

“plt.figure(figsize=(10,20)) Is Not Working.” *Stack Overflow*, 29 June 2019, stackoverflow.com/questions/56815400/plt-figurefigsize-10-20-is-not-working.

“UndefinedMetricWarning: F-score Is Ill-defined and Being Set to 0.0 in Labels With No Predicted Samples.” *Stack Overflow*, 1 Apr. 2017, stackoverflow.com/questions/43162506/undefinedmetricwarning-f-score-is-ill-defined-and-being-set-to-0-0-in-labels-wi.

Codeacademy Team. “Seaborn Styling, Part 2: Color | Codecademy.” *Codecademy*, 2020, www.codecademy.com/article/seaborn-design-ii.

Zach, and View all posts by Zach. “Pandas: How to Drop Rows That Contain a Specific Value.” *Statology*, 14 May 2021, [www.statology.org/pandas-drop-rows-with-value](http://www.statology.org/pandas-drop-rows-with-value).

“Getting ValueError: The Indices for Endog and Exog Are Not Aligned.” *Stack Overflow*, 10 May 2016, stackoverflow.com/questions/37144913/getting-valueerror-the-indices-for-endog-and-exog-are-not-aligned.

akhilvasabhaktula03. “How to Create a Stacked Bar Plot in Seaborn? - GeeksforGeeks.” *GeeksforGeeks*, 2 Jan. 2022, www.geeksforgeeks.org/how-to-create-a-stacked-bar-plot-in-seaborn.