Laycee Glass

NBM3 Task 2

Predictive Modeling

11/27/24

Western Governors University

## NBM3 Performance Assessment, Task 1

### Student Information

Student Name: Laycee Glass

ID Number: 001896026

Date: 11/27/2024

### Part I: Research Question

### A1. Question

The research question I chose to answer was the following: “What factors contribute most to readmission of patients?”

### A2. Goals of Analysis

Similarly to the question from task 1, the question is broad in nature. But answering it could help hospitals identify contributing factors to readmissions. Being able to identify these could, in turn, help with cost reduction and patient satisfaction.

### Part II: Method Justification

### B1. Assumptions

When working with logistic regression, there are a few assumptions for the data we are using. These assumptions are the following (*Assumptions of logistic regression*, 2024):

1. Binary/categorical dependent variable: Logistic regression requires the dependent variable to have only 2 outcomes (in our case, “Yes/1” and “No/0”).
2. Independence: Observations should be independent of one another.
3. Linearity of logit: Log-odds of dependent should have a linear relationship with the independent variables.
4. Sufficient sample size: Logistic regression requires sufficiently large sample size to produce reliable estimates. General guideline is at least 10 events per predictor variable.
5. Multicollinearity: Independent variables should not be highly correlated with each other.

### B2. Benefits of Python

For this assessment, I used Python. Python is my preferred programming language as it has comprehensive libraries and tools that make performing analysis tasks efficient. These libraries are the following:

* NumPy – offers support for mathematical functions and computations
* Pandas – makes data manipulation easy
* SciPy – has modules for advanced statistical analysis
* Statsmodels – provides functions for estimating and analyzing models
* Matplotlib, seaborn – data visualization
* Scikit-learn – machine learning library with specific tools for regression, as well as a consistent API for model building and evaluation

In addition to this, Python has a consistent and easy syntax to read and follow.

### B3. Justification

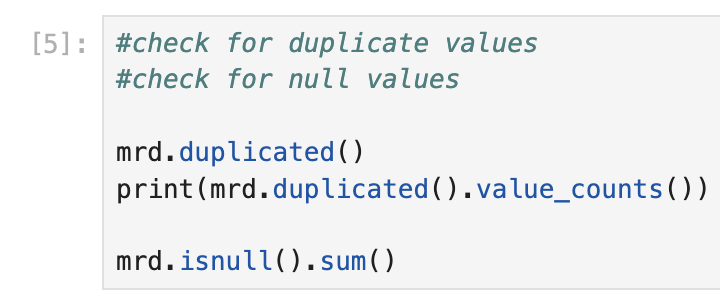
Logistic regression was appropriate to use for this task because the dependent variable, ReAdmis, was binary in nature. Its values were either “Yes/1” or “No/0.” Linear regression does not allow for this and, therefore, would not have been ideal for this task.

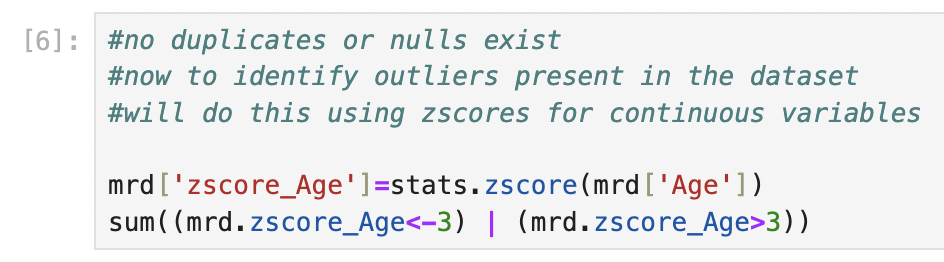
### Part III: Data Preparation

### C1. Data Cleaning Process

In order to perform the logistic regression, some data cleaning needed to take place. The goals of data cleaning were similar to that of task 1. This was to identify and treat null values, duplicates, and identify and retain outliers as they were ‘expected.’ Finally, columns that were not necessary to the model were dropped before exporting to a cleaned CSV.

The following is the code I used to achieve these goals:





A close-up of a computer screen

Description automatically generated

### C2. Summary Statistics

The dependent variable (x) for the question posed was ReAdmis, whereas the independent variables (y) were the following: Age, VitD\_levels, Doc\_visits, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, and Initial\_days.

The following is the code to get the summary statistics for these variables, along with their outputs:

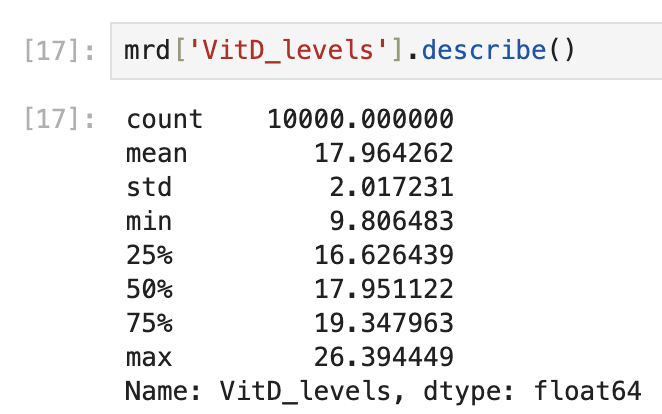
A screenshot of a computer code

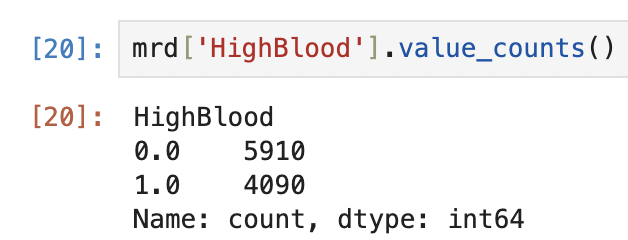
Description automatically generated

A screenshot of a computer code

Description automatically generatedA screen shot of a computer

Description automatically generated



A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

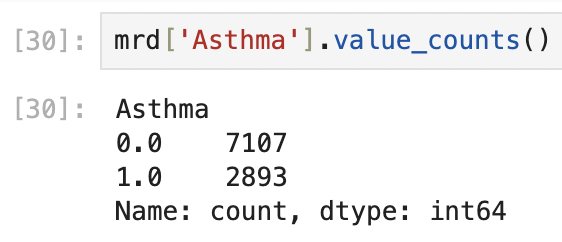
A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

ReAdmis: 36.69% of patients were readmitted.

Age: The minimum was 18 and the maximum was 89. The average age was 53.5.

VitD\_levels: The minimum was 9.81 and the maximum was 26.39. The average Vitamin D level was 17.96.

Doc\_visits: The minimum was 1 and the maximum was 9. The average number of doctor visits was 5.01.

Initial\_days: The minimum was 1 and the maximum was 71.98. The average number of days initially spent in the hospital was 34.46.

HighBlood: 40.9% of patients had high blood pressure.

Stroke: 19.93% of patients had a stroke.

Overweight: 70.94% of patients were overweight.

Arthritis: 35.74% of patients had arthritis.

Diabetes: 27.38% of patients had diabetes.

Hyperlipidemia: 33.72% of patients had hyperlipidemia.

BackPain: 41.14% of patients had back pain.

Anxiety: 32.15% of patients had anxiety.

Allergic\_rhinitis: 39.41% of patients had allergic rhinitis.

Reflux\_esophagitis: 41.35% of patients had reflux esophagitis.

Asthma: 28.93% of patients had asthma.

All variables had a count of 10000.

### C3. Univariate and Bivariate Visualizations

The following is the code, and subsequent charts, used to create the univariate visualizations for the continuous variables and categorical variables, respectively:

A screenshot of a computer code

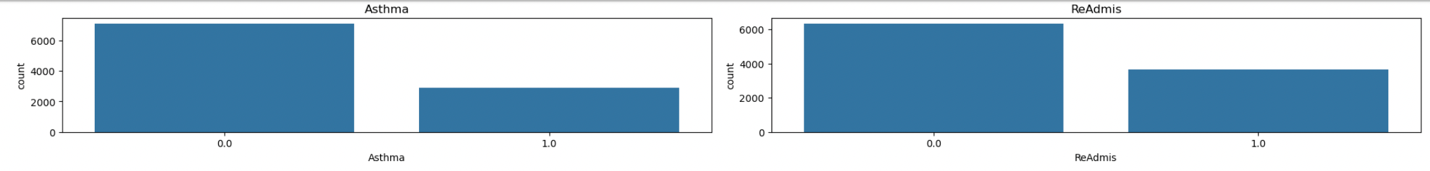
Description automatically generated

A computer code with green and purple text

Description automatically generatedA graph of a diagram

Description automatically generated with medium confidenceA graph of different sizes and shapes

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

The following is the code, and subsequent charts, used to create the bivariate visualizations for the continuous variables and categorical variables, respectively:

A screenshot of a computer code

Description automatically generated

### A screenshot of a computer code Description automatically generatedA comparison of a graph Description automatically generated with medium confidenceA screenshot of a graph Description automatically generated

### A group of blue and orange bars Description automatically generatedA group of blue and orange bars Description automatically generatedC4. Data Transformation Goals

For the purpose of this task, the only transformation that took place was re-expression of categorical variables using the .map function. The variables ReAdmis, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, and Asthma were re-expressed by mapping “Yes” values to 1 and “No” values to 0.

The following is the code used to perform this transformation:

A screenshot of a computer code

Description automatically generated

### C5. CSV File

The cleaned dataset was submitted alongside this document as part of the task submission. The data was exported as a CSV file using the following code:

A close up of a sign

Description automatically generated

### Part IV: Model Comparison and Analysis

### D1. Initial Logistic Regression Model

The initial regression model was constructed using the variables named in section C2. It was created using the .assign(const=1) function. This was done to add the intercept term directly to the DataFrame. The initial model had 15 variables that was eventually reduced to focus on the most meaningful variables to the dependent variable.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThe following is the code used to generate this model and its output:

### D2. Feature Selection

As shown in the output of the code above, there were multicollinearity issues that might have present. For this reason, variance inflation factor (VIF) was used to eliminate variables. There were 2 variables that were eliminated through this process, as their VIF was greater than 5 and indicated high multicollinearity. Once that was complete, backwards elimination was used to reduce the model further based on p-values. Variables whose p-values were greater than 0.05 were eliminated as they were not statistically significant. This resulted in 5 more variables being removed from the model.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThe following code was used for the VIF process as stated above (GeeksforGeeks, 2024):

A white background with red text

Description automatically generatedThe code was then repeated after each variable removal. The following was the code used for the backwards elimination process:

A screenshot of a computer

Description automatically generatedThis was repeated again, removing each variable greater than 0.05 until the following model was left:

### A screenshot of a computer Description automatically generatedD3. Reduced Logistic Regression Model

The final, reduced model is shown above. The variables that remained, seemingly most significant to the dependent variable, were Initial\_days, HighBlood, Stroke, Arthritis, Diabetes, Hyperlipidemia, Anxiety, and Asthma.

### E1. Model Comparison

The initial regression model that was created had many variables, most of which ended up not being important to the model itself. Just under half of the initial variables were eliminated using both VIF and backwards elimination.

When comparing these two models, many of their metrics appear the same or very similar. To demonstrate a clear difference, the AIC values of both were calculated. The initial (full) model had an AIC value 6.3007 points higher than the reduced model. Ideally, a difference of 10 between the models would indicate a strong difference. Substantial differences occur between 2 and 6. Anything less than 2 would not be significant. Therefore, the reduced model is better, as ours falls into the substantial category (Kumar, 2023).

A screenshot of a phone

Description automatically generated The following is the code used to generate the AIC values for both models:

### E2. Confusion Matrix

A screenshot of a computer program

Description automatically generated A confusion matrix weas created for the model. The following is the code and subsequent matrix generated by such:

Here, the results showed a very high degree of accuracy. There were 1880 correct true negatives and 1057 correct true positives, with 31 false positives and 32 false negatives. The accuracy score is 98%, as found by taking the true positives and true negatives and dividing by all the scores (Kundu, n.d.). So, in our case,

### E3. Code

The code for each portion of this task has been shown previously within this document, as well as attached in the .ipynb file.

### Part V: Data Summary and Implications

### F1a. Regression Equation

= -62.2382+ (1.1462 \* Initial\_days) + (0.7634 \* HighBlood) + (1.3657 \* Stroke) - (0.8798 \* Arthritis) + (0.4170 \* Diabetes) + (0.3946 \* Hyperlipidemia) - (0.7199 \* Anxiety) – (1.0142 \* Asthma)

### F1b. Coefficient Interpretation

According to the equation listed above, the intercept is -62.2382. According to the odds ratio, the probability of readmission is near 0 (*e*-62.23820)*.* This assumes all other factors remain constant.

Using the same method and still assuming all other factors remain constant, for every additional hospital stay, the odds of readmission increase by 215%.

Patients with high blood pressure are 2.15 times more likely to be readmitted.

Patients who had a stroke are 3.92 times more likely to be readmitted.

Patients with arthritis have 59% lower odds of being readmitted.

Patients with diabetes are 1.52 times more likely to be readmitted.

Patients with hyperlipidemia are 1.48 times more likely to be readmitted.

Patients with anxiety have 51% lower odds of being readmitted.

Patients with asthma have 64% lower odds of being readmitted.

From this, it appears that Initial\_days and Stroke have the largest positive effect, meaning these patients tend to have higher readmission odds. In contrast, Asthma and Arthritis have the largest negative effect, meaning these patients tend to have lower readmission odds (*Home)*.

### F1c/d. Reduced Model Significance and Disadvantages

Our model would be considered statistically significant based on its LLR p-value of 0.000 (less than the threshold of 0.05). Regarding practical significance, this model definitely lacks insight. The small sample size is one factor that could influence the model. Adding more samples of people of all ages will help with the model’s generalizability.

Another limitation for this data is the limitations of health conditions listed. Some conditions, such as anxiety, may interact with other variables (some even unmodeled), which could oversimplify its effect on the model.

Or, like in the case of the variable Initial\_days, there may be influence of other factors that we cannot quantify or is not available. Initial\_days could be impacted by hospital policy or severity of the patient’s condition.

There are also limitations in the way the model is interpreted. The identified predictors may not translate to enforceable action. For example, patients who had a stroke were almost 4 times more likely to be readmitted. While knowing this is helpful in some manner, there may be no specific ways to reduce the risk associated with it.

### F2. Recommendations

As the model lacks practical significance, my recommendation would be to gather even more data. More data would lead to better generalization between healthcare settings. As noted, there were no children in our dataset, which is not representative of the population.

Another recommendation would be further modeling in the way of random forests, or something similar in nature, in order to capture a better view of complex relationships with the data. Logistic regression may not show these relationships solely, but in addition to another model, may help give a better overview of the predictor variables on the dependent variable.

### Part VI: Demonstration

### G. Panopto Recording

A Panopto recording has been attached with this paper.

### H. Code Sources

GeeksforGeeks. (2024, October 14). *Detecting multicollinearity with VIF - python*. https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

### I. Sources

*Assumptions of logistic regression*. Statistics Solutions. (2024, June 11). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

*Home*. OARC Stats. (n.d.). https://stats.oarc.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/

Kumar, A. (2023, November 30). *AIC in logistic regression: Formula, example*. Analytics Yogi. https://vitalflux.com/aic-in-logistic-regression-formula-example/#:~:text=When%20comparing%20models%2C%20the%20difference,a%20strong%20difference%20between%20models

Kundu, R. (n.d.). *Confusion matrix: How to use it & interpret results [examples]*. V7. https://www.v7labs.com/blog/confusion-matrix-guide