**Performance Assessment**

NBM3 — NBM3 TASK 2: LOGISTIC REGRESSION MODELING

Bader Ale

Department of Information Technology, Western Governors University

D208 Exploratory Data Analysis

August 2023

Table of Contents

[Part I: Research Question 3](#_Toc141911136)

[Part II: Method Justification 3](#_Toc141911137)

[Part III: Data Preparation 4](#_Toc141911138)

[Univariate Visualizations 5](#_Toc141911139)

[Bivariate Visualizations 7](#_Toc141911140)

[Part IV: Model Comparison and Analysis 12](#_Toc141911141)

[Initial Regression Model 12](#_Toc141911142)

[Reduced Regression Model 15](#_Toc141911143)

[Model Comparison 16](#_Toc141911144)

[Discussion 18](#_Toc141911145)

[Recommendations 20](#_Toc141911146)

# Part I: Research Question

Our research question for this task was: What patient factors caused high cholesterol levels. The goal is to determine what are the medical factors that could lead to high cholesterol levels the most.

# Part II: Method Justification

In order to perform logistic regression, several assumptions were made. Firstly, the target variable, in this case Hyperlipidemia has to be binary meaning it can either be Yes or No. Secondly, the explanatory variables have to be independent of each other meaning the variables have nothing to do with each other and are separate observations. Thirdly, there should be no extreme multicollinearity among the variables in order for the model to run correctly – this means that no two explanatory variables should be related to each other. Moreover, a reasonably large enough sample size to run the model is required for a well-functioning predictive model. Lastly, there should exist a linear relationship between the explanatory variables of the model and the logit of the response/target variable. (Zach, The 6 Assumptions of Logistic Regression (With Examples), 2020).

Python was chosen over R for two reasons. The first reason is the existing familiarity with Python; the only thing to learn was which libraries needed to be imported and the parameters. Even though R was written specifically for statistical analysis and Python as more of a general programming language, many libraries have been created for Python to perform like R. The second reason is its flexibility and ease of reading/structuring. Certain Python packages were used in order to facilitate statistical calculations and visualizations.

Logistic regression is best suited to answer the research question because is gives an answer that is not numeric but more of a Yes/No (or a 1/0) (Kanade, 2022). Also, it is a classification method that is best suited for situation where you need to determine if the value you are trying to predict fits into a category – in this case, the classification would be Hyperlipidemia\_Yes or Hyperlipidemia\_No. This is also why this is the best method for this specific research question.

# Part III: Data Preparation

In order to begin running the regression model, the data had to be cleaned and pre-processed. This reduces the model’s error and allows for seamless splitting of the data for both x-and y- values. The data preparation included replacing missing values, if any, Furthermore, checking for any duplicated values that could interfere with our model. Lastly, outliers were dealt with accordingly, ensuring that the values used for the model were accurate and within an acceptable range.

Initially, explanatory variables used in the logistic regression model were the same as those used in the linear regression model namely *Area, Age, Income, Marital, Gender, VitD\_levels, Doc\_visits, Initial\_admin,Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, Asthma, Services, Initial\_days –* in this case, the target variable is *Hyperlipidemia\_Yes*.

A screenshot of a graph

Description automatically generated

## Univariate Visualizations

A graph with several blue rectangular bars

Description automatically generated A graph of blue bars

Description automatically generatedA graph of a number of blue bars

Description automatically generated A graph of a number of people

Description automatically generated

A graph of a number of people

Description automatically generated A graph of a number of levels

Description automatically generated

A graph with numbers and lines

Description automatically generated A graph of a number of blue squares

Description automatically generated

A graph of a high complicatization

Description automatically generated A blue rectangular graph with white text

Description automatically generated

A graph with blue squares

Description automatically generated A blue rectangular graph with white bars

Description automatically generated

A graph with blue squares

Description automatically generated A graph with blue squares

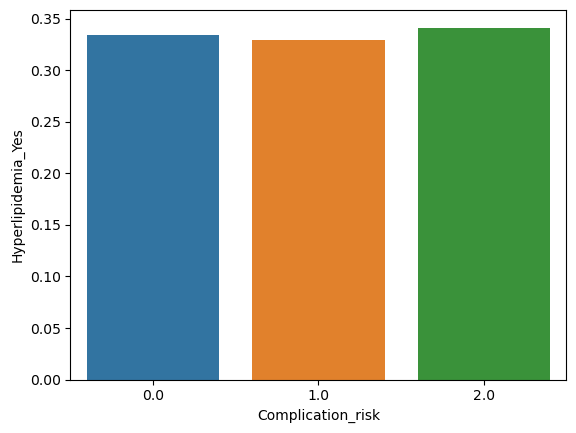
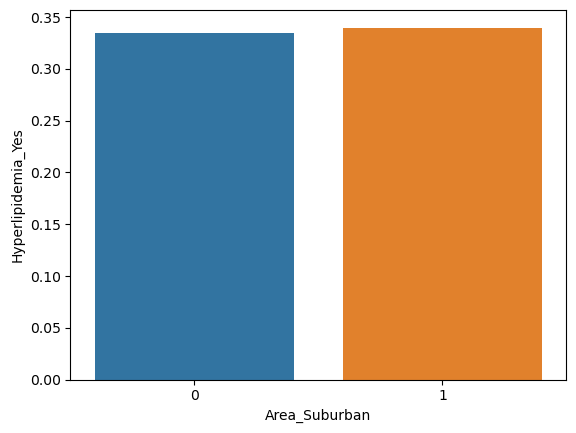
Description automatically generated

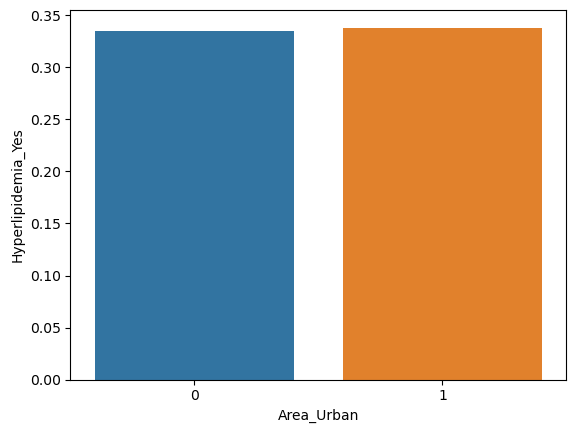
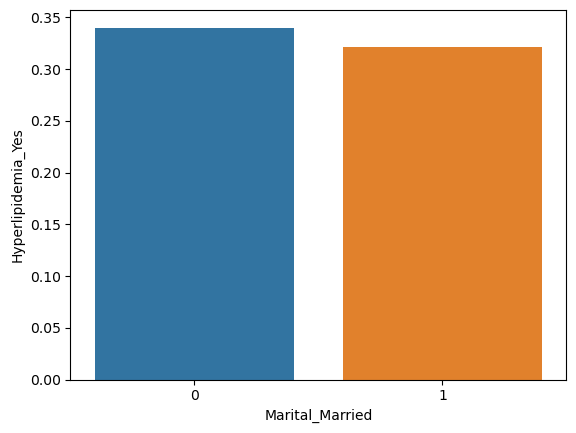
A graph of a bar graph

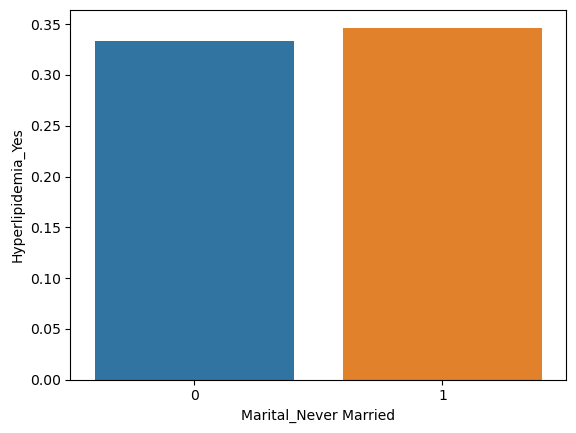
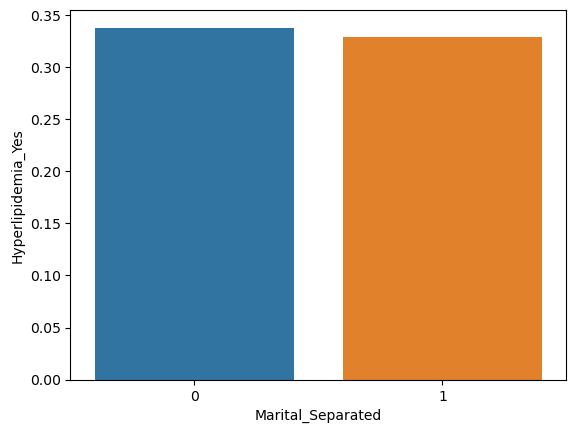
Description automatically generated A graph of a number of blue squares

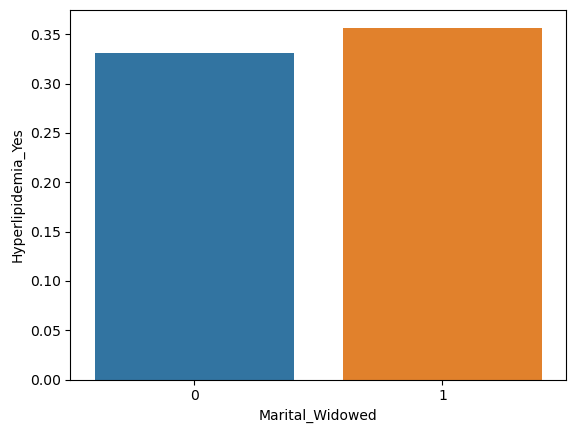
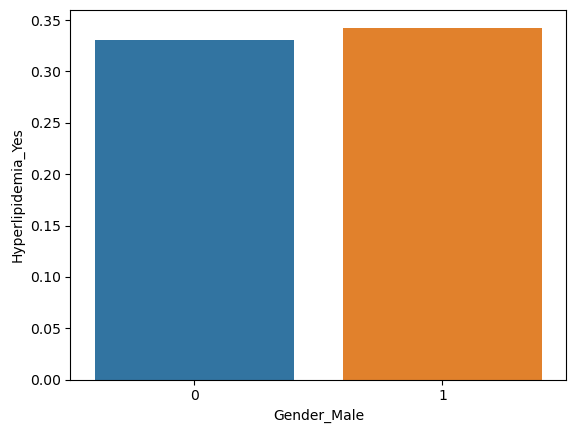
Description automatically generated

## Bivariate Visualizations

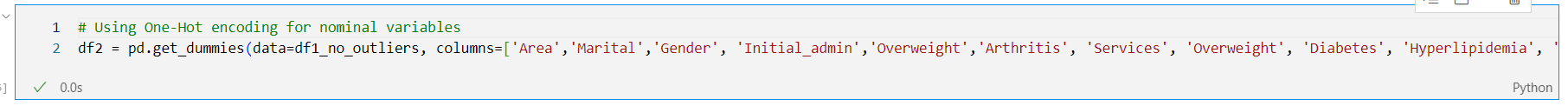
 

The goal for transforming the data is to prepare it in a way that is easy for the model to use and that minimizes error. This included re-expressing categorical variables accordingly so the code could run efficiently. The variables needed to be re-expressed were *Area, Marital, Gender, Initial\_admin, Overweight, Arthritis, Services, Overweight, Diabetes, Hyperlipidemia, and Asthma.* This was accomplished by using the Pandas functionality *pd.get\_dummies –* the code is shown below. No missing values or duplicates were noted in the dataset.



A copy of the final code is attached.

# Part IV: Model Comparison and Analysis

## Initial Regression Model

An initial logistic regression model was run using SKLearn library. Before running, the data was split into testing and training as shown below. The dependent variable3 was specified as Hyperlipidemia\_Yes and the rest of the variables were included in the X-data set.



Figure 2: X and Y variable assignments

The following shows the target variable classes, coefficients, and intercept of the initial model:

A screenshot of a computer

Description automatically generated

The model was scored using both the training set and the testing set. A big difference in scores would mean that the model is not performing well and would not give good real-world predictions.

A screenshot of a computer code

Description automatically generated

The equation includes all 17 explanatory variables in the initial regression model and their respective coefficients. Moreover, the constant was also included. Our next task is to reduce the number of features and run a reduced regression model.

## Reduced Regression Model

In order to reduce the explanatory variables, Recursive Feature Elimination (RFE) was used. For this, we used the SKlearn library and the *RandomForestRegressor* as the estimator parameter. RFE was chosen due to its reduced complexity and selection of estimator algorithm – RFE works by first running the all the explanatory variables first and assigning an coefficient or “value of importance” to each one. In this specific case by specifying the parameters *step=1* , one by one the features were dropped from least importance and the top 3 (explicitly given in *n\_features\_to\_select=3*) (Brownlee, 2020). A screenshot of the code is shown below.

A screenshot of a computer program

Description automatically generated

Here, we specified the algorithm to remove one feature at a time and to reach a maximum of three total variables to use in the reduced model.

A screenshot of a computer program

Description automatically generated

The RFE mechanism chose *Complication\_risk, Initial\_days, and Initial\_admin\_Emergency Admission.* We will now use this as input into our reduced model.

A screenshot of a computer error

Description automatically generated

As before, the new parameters were passed to the model object and the model was executed. Below is a screenshot of the model intercept, coefficients, and the target variable.

A screenshot of a computer code

Description automatically generated

Using the above information, a logistic regression model equation was then computed:

*ln(y/(1-y)) = -0.66695895 + 0.02214502(Complication Risk) - 0.03500518(Initial\_Days) + 0.00724142(Initial\_admin\_Emergency Admission)*

The coefficients can help us determine how each explanatory variable might affect the chance of the target variable – the sign determines how it will affect it (whether positively or negatively) and the value can tell us how it effects the odds of the target variable. For example, the coefficient for *Complication\_risk* is positive meaning it increases the odds of the target variable being 1 for *Hyperlipidemia\_Yes*. In order to see by how much, we can use the equation:

For our example of *Complication\_Risk,*  the above equation would 1.0223 – this means that the depending on the complication risk (0 for Low, 1 for Medium and 2 for High) the combined effect of the risk level AND the coefficient would have 1.0223 times the odds for the patient having Hyperlipidemia (Zach, How to Interpret Logistic Regression Coefficients (With Example), n.d.)

Just like before, the reduced model was scored on the training and test data, and they were both very similar.

A screenshot of a computer code

Description automatically generated

A confusion matrix was also executed – we can see that the reduced model correctly predicted 67.08% of the time with the test data while the remaining 32.92% would have given a true-negative (not correctly predicted).

A screenshot of a graph

Description automatically generated

## Model Comparison / Discussion

Both models performed fairly well – the reduced model, even with the removal of 14 variables by the RFE mechanism, only dropped by 2-3 percentage points which means the model can more or less predict with 67.08% precision that a person will have high cholesterol even with only 3 variables.

## Recommendations

Based on the initial and reduced logistic models, the model can predict with around 67% accuracy is a person would have high cholesterol based off of only 3 parameters. In the real world, there are a plethora of other factors to consider in order to check for hyperlipidemia – this model only takes into account those already hospitalized. Essentially, it can be determined from three variables if the person would have hyperlipidemia without actually testing for it.

There are numerous other factors involved and the data itself is not spread to a more general global population, so the limitations are quite a few. In general, it was shown that the reduced logistic model had similar results with only three of the initial 17 predictive variables.

Works Cited

Brownlee, J. (2020, August). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/rfe-feature-selection-in-python/

Kanade, V. (2022, April). *What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices*. Retrieved from Spiceworks: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#\_002

Zach. (2020). *The 6 Assumptions of Logistic Regression (With Examples)*. Retrieved from Statology: https://www.statology.org/assumptions-of-logistic-regression/

Zach. (n.d.). *How to Interpret Logistic Regression Coefficients (With Example)*. Retrieved from Statology: https://www.statology.org/interpret-logistic-regression-coefficients/