D208 - Predictive Modeling

#### Performance Assessment - Task 2: Logistic Regression for Predictive Modeling

#### Medical Readmission Data Set (Clean)

Sam Spencer

Western Governors University

Table of Contents

[Part I 2](#_Toc93887135)

[A1: Research Question 2](#_Toc93887136)

[A2: Objectives and Goals 2](#_Toc93887137)

[Part II 2](#_Toc93887138)

[B1: Summary of Assumptions 2](#_Toc93887139)

[B2: Tool Benefits 2](#_Toc93887140)

[B3: Appropriate Technique 2](#_Toc93887141)

[Part III 2](#_Toc93887142)

[C1: Data Goals 2](#_Toc93887143)

[C2: Summary Statistics 2](#_Toc93887144)

[C3: Steps to Prepare the Data 2](#_Toc93887145)

[C4: Visualizations 2](#_Toc93887146)

[C5: Prepared Data Set 2](#_Toc93887147)

[Part IV 2](#_Toc93887148)

[D1: Initial Model 2](#_Toc93887149)

[D2: Justification of Model Reduction 2](#_Toc93887150)

[D3: Reduced Logistic Regression Model 2](#_Toc93887151)

[Part V 2](#_Toc93887152)

[E1: Model Comparison 2](#_Toc93887153)

[E2: Output and Calculations 2](#_Toc93887154)

[E3: Code 2](#_Toc93887156)

[Part VI 2](#_Toc93887157)

[F1: Results 2](#_Toc93887158)

[F2: Recommendations 2](#_Toc93887159)

[I: Sources 2](#_Toc93887160)

## Part I

### A1: Research Question

Obesity is a large problem with our health care system, it cause many long-term health problems if it isn’t treated. My research will be based around this, can we predict who will be overweight, based on the predictors in our given data set

### A2: Objectives and Goals

My objective and goals for this analysis is for us to assist the chain of hospitals understand what causes obesity within their communities with the given data set. By discovering this, it will allow our chain of hospitals focus more care on patients who need it the most while also still taking care of overweight patients.

Part II

### B1: Summary of Assumptions

Logistic regression has 6 assumptions:

1. The response must be binary – An example of this would be Yes/No M/F, or for our research
2. The observations must be independent. “Logistic regression assumes that the observations in the dataset are independent of each other. That is, the observations should not come from repeated measurements of the same individual or be related to each other in any way.” (Bobbitt, 2020)
3. There can be no multicollinearity with explanatory variables. “Multicollinearity occurs when two or more explanatory variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. If the degree of correlation is high enough between variables, it can cause problems when fitting and interpreting the model. “ (Bobbitt, 2020)
4. If there are outliers, there can’t be a large gap with outliers and other observations.
5. The relationship must be linear with the explanatory variables and the response variable.
6. The sample size must be significant enough to encompass all variables.

### B2: Tool Benefits

The tools that will be utilized in this analysis will be:

* Python
* Anaconda Navigator
* Jupyter notebook

The use of Python seems like an obvious choice, according to InterviewBit, “With more than 137,000 libraries, Python can be used to create applications and models in a variety of fields, for instance, machine learning, data science, data visualization, image and data manipulation, and many more”( (Top 10 python libraries 2022)) some of the libraries that will be used for our analysis are: Pandas, SciKitLearn, Seaborn, MatPlotLib, & NumPy. Anaconda Navigator is a GUI that is very beneficial for our use, It allows launching of applications, such as Jupyter notebook and several others without having to use command line commands. Jupyter Notebook will house all our code, equations, and visualizations.

### B3: Appropriate Technique

Logistic regression is defined as “A statistical method used to predict the outcome of a dependent variable based on previous observations. It's a type of regression analysis and is a commonly used algorithm for solving binary classification problems.” (Joby, 2021) since the outcome of our analysis will be a yes or no, that makes logistic regression perfect for analysis.

## Part III

### C1: Data Goals

The goals for our data are relatively minor changes. First we want to get an understanding of the data, this step will involve us including all variables, numerical and categorical.

The next step is to start cleaning our data and subset it into the variables that are beneficial to our analysis. We then encode the categorical values into 0/1 binary values.

Lastly, as long as our cleaning and wrangling is successfully, we can then start building our logistical model.

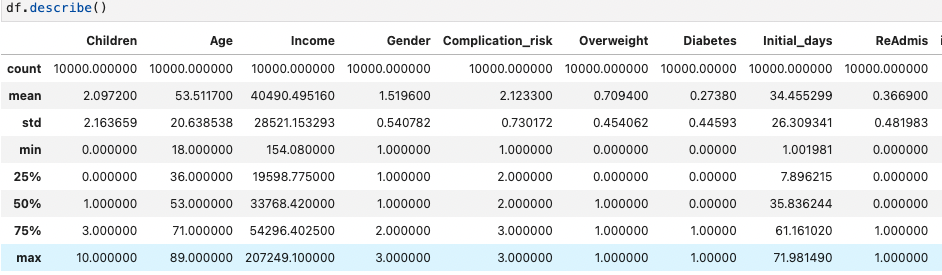
### C2: Summary Statistics

The target variable is Overweight, and the predictors are illnesses associated with obesity, which in turn increases chances for hospital admissions and readmission, and demographics of the patients. Predictors: 'Income', 'Children', 'Age', 'Full\_meals\_eaten', 'Initial\_days', 'Overweight', 'Diabetes', 'HighBlood', 'Gender', 'Asthma', 'Soft\_drink', 'Doc\_visits', 'Anxiety'.

Using the describe function of pandas, we can get an understanding of the data.

* The age range is from 18-89, but the average age is 53 years old.
* The number of children ranges from 0-10, but the average amount of children is 2
* Income ranges from $154- $207,250 with an average of approximately $40,000
* Total Charges per day ranges from $2100-$9200 with the average amount charged per day being a little over $5,000

The output of df.describe():



### C3: Steps to Prepare the Data

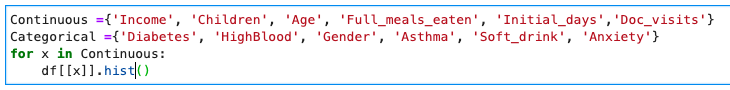
The following steps are the preparations steps and code:

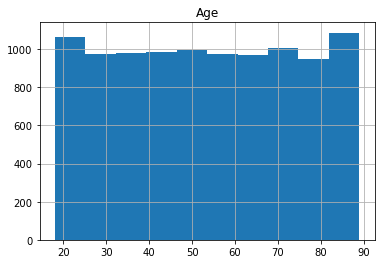
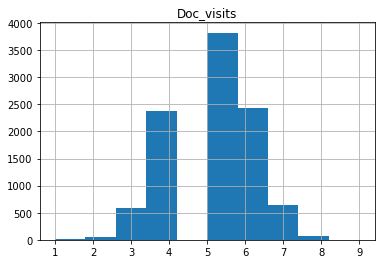
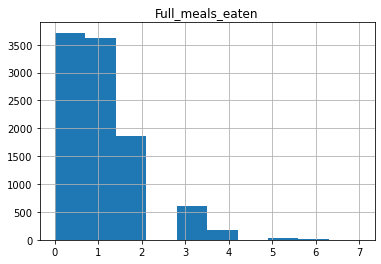
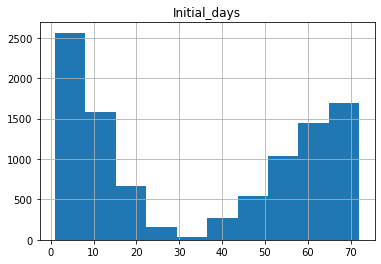
* Import necessary packages and libraries (these were mentioned in B2)
* Review the data and get an understanding of the data
* Convert categorical variables into numeric.
* Check for nulls and dupe and remove if necessary
* Remove variables that aren’t necessary for our analysis
* Generate bivariate and univariate visualizations (C4)
* Provide the cleaned data.

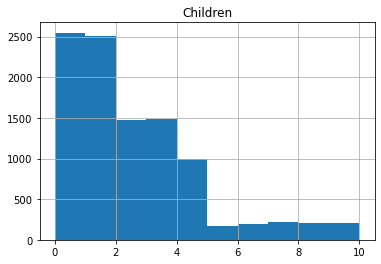
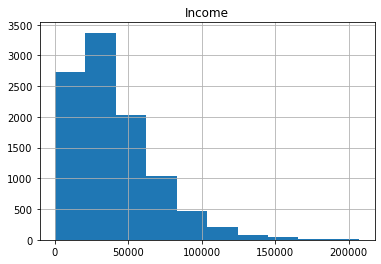
### C4: Visualizations

#### Univariate

Below you will see the visualizations for the continuous variables used in the analysis and the code that produced it.







Chart, bar chart

Description automatically generated with medium confidenceLogo

Description automatically generated with medium confidenceLogo

Description automatically generatedLogo

Description automatically generatedLogo

Description automatically generatedA picture containing bar chart

Description automatically generatedLogo

Description automatically generated

Bivariate visualizations

Text

Description automatically generated with medium confidenceLogo

Description automatically generated with medium confidenceGraphical user interface

Description automatically generated with medium confidenceA picture containing graphical user interface

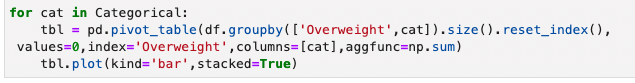
Description automatically generatedLogo

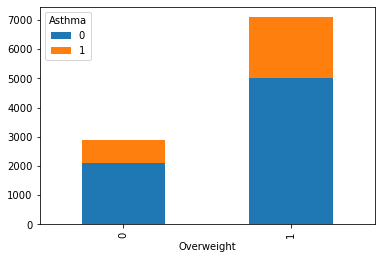
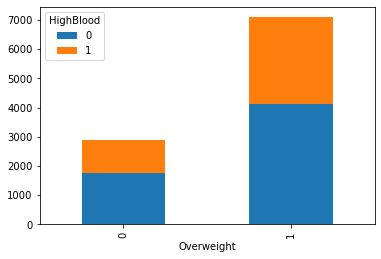
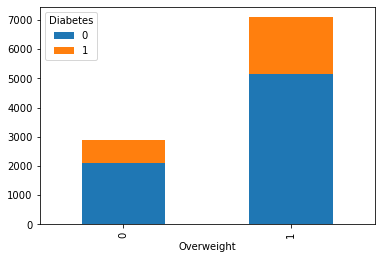
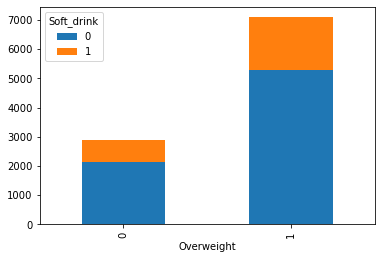
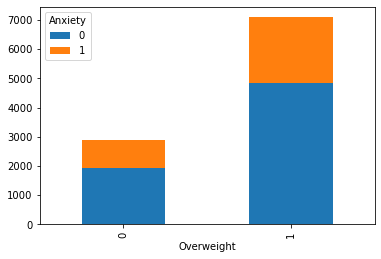
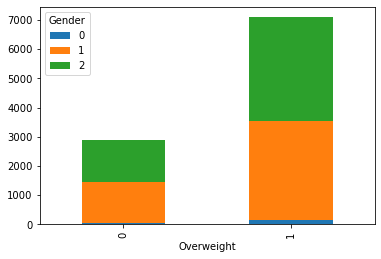
Description automatically generatedA picture containing logo

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

Categorical Bivariate visualizations





### C5: Prepared Data Set

Prepared data set will be attached along with reduced.

## Part IV

### D1: Initial Model

Below is our initial model with all of our variables used.

A picture containing chart

Description automatically generated

Params:

Table

Description automatically generated with medium confidence

Logistical Regression below:

Table

Description automatically generated

### D2: Justification of Model Reduction

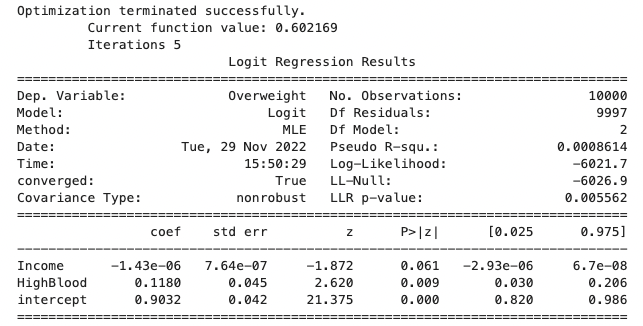
If you look at the output of our model, most of the values have a p-values over .05. The general rule of p-values is .05 or lower is considered statistically significant, and the only values that are at or below our threshold are: Income and HighBlood.

### D3: Reduced Logistic Regression Model

Below is the reduced model using the variables income an HighBlood

A picture containing text

Description automatically generated



## Part V

### E1: Model Comparison

Between the 2 models, there is a slight improvement in terms of fit. If you look at the R score, there is an increase of .0006927 between the two models. while the model benefitted from the removal of the predictors, there is room for more improvement.

I used the K fold cross validation to test the accuracy of the model, the model is 68% accurate reduced.

### E2: Output and Calculations

Graphical user interface, text, application, email

Description automatically generated

### A screenshot of a cell phone Description automatically generated with medium confidence

### E3: Output and Calculations

Code is attached with submission in a Jupyter Notebook. Along with a screenshot below

Graphical user interface, text, application

Description automatically generated

## Part VI

### F1: Results

The regression equation for Overweight is Income = -1.43e-06 + (0.1180\* HighBlood)

The coefficients tell us that income is trending negatively to being overweight, but HighBlood is the opposite, high blood pressure appears to have a positive relationship with our target variable, overweight increase with individuals who are overweight.

The primary limitation of the research and analysis that we conducted is that there is no way for us to include everyone in analysis with 10,000 observations there are several demographics that aren’t represented, and business decisions could be made and negatively impact the demographics that aren’t represented.

### F2: Recommendations

My recommend course of action from our analysis is gathering a larger, much more diverse sample size to conduct research on. Obesity is a problem that is prevalent in nearly every demographic and ethnicity, though, it effect some more than others. This analysis should be used as a supplemental analysis and further analysis would need to be conducted to get an accurate understanding of what causes patients initial visit resulting in hospital admission.

According to (Chandra, 2021) “Data from population-based studies showed nearly linear relationship between body mass index (BMI) and blood pressure.6,7 Findings from the Framingham Heart Study showed that 78% of essential (primary) hypertension in men and 65% in women can be attributed to obesity”

Some individuals are just prone to obesity due to genetics or lack of nutrient rich food or disability that immobilizes them, but for all individuals who are overweight, it would be beneficial to have regular health screenings and having blood pressure checked during these screenings. It is also recommended that overweight individuals check their blood pressure regularly at home as well. High blood pressure is typically a comorbidity for many people and the other illnesses they have in addition to high blood pressure are associated with obesity therefore it is critical for patients to have their health checked regularly.

If key leadership and the community act on the findings of this analysis, the correlation between obesity and high blood pressure can be reduced. This will keep less people from needing to come to the chain of hospitals which in turn would lower the risk for fines and penalties associated with excessive 30-day readmission.

### G: Provide a Panopto recording

* Panopto is attached with submission

### H: Third party Code

* [Pandas](https://pandas.pydata.org/)
* [NumPy](https://numpy.org/)
* [matplotlib](https://matplotlib.org/)
* [Seaborn](https://seaborn.pydata.org/)
* [Jupyter Notebook](https://jupyter.org/)

# I: Sources

Bobbitt, Z. (2020, October 13). *The 6 assumptions of logistic regression.* Statology. Retrieved November 9, 2022, from https://www.statology.org/assumptions-of-logistic-regression/

Centers for Disease Control and Prevention. (2020). *NAMCS/NHAMCS - Ambulatory Health Care Data Homepage*. Centers for Disease Control and Prevention. Retrieved November 27, 2022, from https://www.cdc.gov/nchs/ahcd/index.htm

Joby, A. (2021, July 29). *What is logistic regression? learn when to use it*. Learn Hub. Retrieved November 26, 2022, from https://learn.g2.com/logistic-regression

*Top 10 python libraries*. InterviewBit. (2022, August 22). Retrieved November 10, 2022, from https://www.interviewbit.com/blog/python-libraries/#:~:text=With%20more%20than%20137%2C000%20libraries,data%20manipulation%2C%20and%20many%20more.

*Weight-loss strategies for prevention and treatment of hypertension*. professional.heart.org. (n.d.). Retrieved November 29, 2022, from https://professional.heart.org/en/science-news/weight-loss-strategies-for-prevention-and-treatment-of-hypertension/Commentary#:~:text=Data%20from%20population%2Dbased%20studies,(BMI)%20and%20blood%20pressure.&text=Findings%20from%20the%20Framingham%20Heart,can%20be%20attributed%20to%20obesity.