**D208 Performance Assessment**

**LOGISTIC REGRESSION FOR MEDICAL DATA**

Fahim A. Akbar Student ID 001434895 Masters Data Analytics (January 1, 2021) Program Mentor: Lea Yoakem (877) 435-7948x6422 [fakbar3@wgu.edu](mailto:fakbar3@wgu.edu)

**Part I: Research Question**

**A.  Purpose of using Logistic Regression for Data Analysis**

**1.  Question**

The proposed research question is if the probability of a patient’s readmission is affected by factors such as age, gender, or present medical conditions.

**2.  Goal**

The objective of this data analysis is to predict a patient’s chance of readmission. To do so, we will test to see which independent variables in the dataset are good indicator and fit. Once we decide on which variables to use, we will be doing a logistic regression analysis and predict the probability of patient readmissions. We will also provide a logistic equation along with the statistics for the identified variables.

**Part II: Method Justification**

**B.  Logistic Regression Methods**

**1.  Logistic regression model assumptions**

According to Statology, the following assumptions must be met for the logistic regression model to be a good fit for the present data: (Zach 2020)

* The dependent variable is binary
* Observations are independent. Residual plots against time can be used to check if there is a random pattern in the order of observations. If there isn’t, then the observations do not satisfy this assumption
* No multicollinearity should be present among the variables.
* No extreme outliers should be present.
* Explains the relationship between one dependent binary variable and one or more independent ordinal and continuous variables
* A linear relationship should be present between the independent variables and logit of the dependent variable. Logit is defined as Logit(p) = log(p / (1-p)), where p is the probability of a positive outcome.
* The dataset offers a large sample size. Our data set has over 10,000 entries and is therefore sufficient.

**2.  Benefits of using Python**

In this assessment, Python will be used for the logistic regression model of the medical data. Python provides several packages that allow for statistical analyses, such as Pandas, Scipy and Statsmodels. Python enables easy-to-understand visualizations of variables and the observations within them. It allows you to analyze data easily into separate groups without modifying the original dataset. Python also has commands specific to logistic regression, including LogisticRegression, model.fit, and model.coef\_. These commands will help with performing the regression analyses and visualizing them.

The following packages will be imported and used for this analysis:

* Pandas – Standard import for data science projects. Provides methods to read and visualize data and statistical tools to parse and score data.
* Numpy – Standard import for data science projects. Provides methods to read and visualize data and statistical tools to parse and score data.
* Matplotlib – Standard import for visualizations. This package provides more robust tools to visualize reports and data points.
* Seaborn – Provides descriptive and visually intuitive graphs, plots, and matrices
* Scikit-learn/sklearn – Provides methods and arguments for splitting, training, testing, and fitting data. This package also has arguments for predicting and classifying data as well as applying metrics for models

**B3. Why logistic regression is an appropriate technique to analyze the research question**

For this assessment we are not trying to predict specific values. Instead, we are trying to predict the probability of the event on interest, readmission, occurring. To analyze the probability of patient readmission, Logistic regression is an appropriate technique. This is because the patient readmission is a binary predictor to a categorical variable (Readmission = Yes or No.) The patient observations are the independent variables of the data set that will be tested.

**Part III: Data Preparation**

**C.  Summarize the data preparation process for logistic regression by doing the following:**

**C1. Relevant data preprocessing goal and data manipulations**

To prepare the data for our logistic regression analysis, we must make sure that we fulfill the logistic regression assumptions:

1. data cleaning: we begin by removing missing values, duplicates, and outliers. We will check that the variables in the loaded dataset are in the correct class, if variables were read incorrectly, class will be corrected. Looking for missing values like in d206
2. We will test linearity of continuous variables by generating histograms, and transform non-linear variables using log(), exp() or other functions that ensure variable linearity and fulfil that logistic regression assumption. feature selection: remove features that are dependent on another (logically and mathematically). Geographical data such as latitude and longitude will be removed since they do not pertain to our research question. (inset other features that are removed and why they were removed)
3. Onehot encoding categorical variables: we convert categorical variable to binary variable at this step to prepare our categorical variables for logistic regression. In order to address this data goal, we will use get\_dummies() function.
4. scaling and normalization: to make features comparable, we will run scaling and normalization. In order to achieve this goal, we can use StandardScaler() function to standardize and normalize the data.
5. splitting the data: we split our dataset in 80% for model training and 20% for model testing. In order to split the dataset, we will use train\_test\_split() function.
6. balancing: we used supervised sampling to established a balanced training and testing datasets, where both dependent variable levels truly represent the distribution of the dependent variable in the original dataset. In order to ensure balanced training and testing dataset we will check the frequencies of the dependent variable levels in the original dataset and the resulting dataset and ensure balanced sampling.
7. multicollinearity: we check for multicollinearity between the independent variables, and we will remove variables that violates the logistic regression assumption of predictors independency. In order to address this data goal, we will use variance\_inflation\_factor() function to calculate VIF score for each variable.

If any duplicate data entries, rows, or columns are found, they will be removed. We will also be removing some of the categorical variables that have a high number of unique levels which cannot be condensed.

The approach

>HERE IS HOW I WANT TO PREPARE MY DATA

>We need the variables as numeric, so we create

>k-1 rule

>u only use 1 fewer than the number of

>if a category has 3 u create 2 dummy variables

>

>Complication risk, high medium low make two dummy variables (high and low)

>create a dummy for high and one for low

>If comp high is 0 and comp low is 0 than medium

>PYTHON CONVERTING STRING FEATURES TO NUMBERS

[**https://www.geeksforgeeks.org/how-to-convert-categorical-string-data-into-numeric-in-python/**](https://www.geeksforgeeks.org/how-to-convert-categorical-string-data-into-numeric-in-python/)

>be careful about ordinal encoding (i.e binary, female, or male as 1,2,3)

>d206 webinar number 3 encoding methods, use the DUMMY variable encoding, not dummy

>

>backup the data,

>make dummy variables

>What’s my statistical reason to exclude a variable

>

>use a command line to clean the data, evaluate the data structure to better input

>name

>examine potential mispell

>find outliers using histograms

>imputing value for missing values such as some mean or remove outliers if beyond 2 std dev of the mean

>most relevant to the ReAdmis

>in cleaning the data it may be discovered that the following variables will be important

>likewise the following relevance for the categoricals, recode them

Ex. Initial\_Admin; “\_\_\_\_”

>

Discrete, ordinal, or numerical

**2. Summary statistics (List the TYPE of variable,**

**1- list all variables and their datatype**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 CaseOrder 10000 non-null int64

1 Customer\_id 10000 non-null object

2 Interaction 10000 non-null object

3 UID 10000 non-null object

4 City 10000 non-null object

5 State 10000 non-null object

6 County 10000 non-null object

7 Zip 10000 non-null int64

8 Lat 10000 non-null float64

9 Lng 10000 non-null float64

10 Population 10000 non-null int64

11 Area 10000 non-null object

12 TimeZone 10000 non-null object

13 Job 10000 non-null object

14 Children 10000 non-null int64

15 Age 10000 non-null int64

16 Income 10000 non-null float64

17 Marital 10000 non-null object

18 Gender 10000 non-null object

19 ReAdmis 10000 non-null object

20 VitD\_levels 10000 non-null float64

21 Doc\_visits 10000 non-null int64

22 Full\_meals\_eaten 10000 non-null int64

23 vitD\_supp 10000 non-null int64

24 Soft\_drink 10000 non-null object

25 Initial\_admin 10000 non-null object

26 HighBlood 10000 non-null object

27 Stroke 10000 non-null object

28 Complication\_risk 10000 non-null object

29 Overweight 10000 non-null object

30 Arthritis 10000 non-null object

31 Diabetes 10000 non-null object

32 Hyperlipidemia 10000 non-null object

33 BackPain 10000 non-null object

34 Anxiety 10000 non-null object

35 Allergic\_rhinitis 10000 non-null object

36 Reflux\_esophagitis 10000 non-null object

37 Asthma 10000 non-null object

38 Services 10000 non-null object

39 Initial\_days 10000 non-null float64

40 TotalCharge 10000 non-null float64

41 Additional\_charges 10000 non-null float64

42 Timely\_admis 10000 non-null int64

43 Timely\_treat 10000 non-null int64

44 Timely\_visits 10000 non-null int64

45 Reliability 10000 non-null int64

46 Options 10000 non-null int64

47 Hrs\_treat 10000 non-null int64

48 Courteous 10000 non-null int64

49 Active\_listen 10000 non-null int64

**Separate the output variable (independent variable): talk about frequencies and if there was any missing values that you removed.**

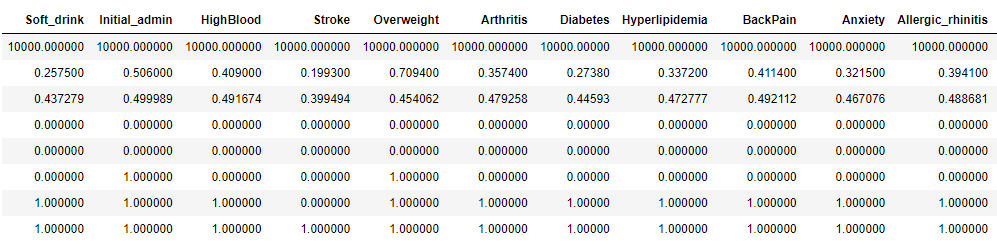
**2- summarize continuous variables table bellow is perfect, just limit to only continuous variables**

**3- summarize categorical variables: frequencies, levels, missing observations,**

P-values and coefficients for the independent variables will be identified to figure out which independent variables impact the probability of the dependent variable. For this analysis, the target dependent variable is ReAdmis. This variable represents whether a patient has been readmitted to the hospital. We believe that performing logistic regression will help predict which patients have a higher chance of being readmitted. This analysis could help the hospital plan to reduce readmissions, saving time and resources for the hospital and the patients. The predictor variables in this study are the patient observations, such as Children, Income, and Initial\_Admin. The tables below show us the standard deviations of each variable as well as the dispersion in the interquartile ranges. From here, we see that Children, Full\_meals\_eaten, Income, and vitD\_supp are not normally distributed variables, while Age, VitD\_Levels, Doc\_visits, and Initial\_days are normally distributed.

The summary statistics of all the variables in the cleaned data sheet are showcased below.





All of the categorical variables were converted to numerical variables so that we can conduct the logistic analysis and compare statistical data. The variable visualizations are shown in section 4. A summary of the statistics for these variables show that Age, BackPain, Reflux\_esophagitis, and Gender are normally distributed. Marital, Soft\_drink, Initial\_admin, Stroke, and Diabetes are not normally distribute and separated by large margins.

>Do a value count on the variable, how many were yes, how many were no

>Range, mean, median, mode, think descriptive statistics.

>”The mean for this variable is \_” when there’s an output, do that for each variable.

>The dataset consists of 50 original columsn and 10,000 records. For the purpose of analysis, certain demographics were removed. Any categorical containing more than 3 levels

>binomial gender,

>measures of \_ thru histograms

>The clean shows no more outliers

> “df of categorical features .describe(loop through all) ”

**3. Data Preparation Steps**

The first step in preparing the data is to make sure that there are no missing data entries in any of the columns. Next, we want to make sure there aren’t any duplicated data entries. We’ll also make sure that there aren’t any duplicated columns or rows to further prevent dealing with repeated entries. For the logistics analysis, several columns in the dataset were deemed irrelevant and were subsequently dropped from the dataset (i.e latitude, longitude). The “yes/no” entries for the categorical variables will be converted to 1 and 0, respectively. The marriage variable will also be condensed to fewer unique answers, “Married” and “Non-Married.” Gender will be condensed to “female” and “non-female” to reduce the levels. “Initial\_admin” will be condensed to "Emergency Admission/non-Emergency Admission" to reduce the levels. Services and complication risks will be dropped due to the high number of levels and upon further inspection, these levels could not be condensed to a binary set. Once all the necessary modifications to the data set are made, we will proceed with creating the initial logistics regression model.

**Code used for preparing data:**

# import packages that will be used for the logistics regression analysis

import pylab

import seaborn as sb

sb.set(style="white")

sb.set(style="whitegrid", color\_codes=True)

import sklearn

from sklearn.metrics import confusion\_matrix

from sklearn import preprocessing

from sklearn.decomposition import PCA

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve

from sklearn.metrics import classification\_report

from sklearn import metrics

import matplotlib.pyplot as plt

plt.rc("font", size=14)

import numpy as np

import scipy.stats as stats

import statsmodels.api as sm

import statsmodels.formula.api as smf

from IPython.core.display import HTML

from IPython.display import display

import pandas as pd

from pandas import Series, DataFrame

from sklearn.metrics import classification\_report, confusion\_matrix

from imblearn.over\_sampling import SMOTE

# import data set that will be used for the logistics regression analysis

pd.set\_option('display.max\_columns', None)

df = pd.read\_csv (r'C:\Users\fahim\Documents\0\_WGUDocuments\d208\1medical\_clean.csv')

# rename the item columns accordingly

df.rename(columns={'Item1':'Timely\_admis','Item2':'Timely\_treat',

'Item3':'Timely\_visits','Item4':'Reliability',

'Item5':'Options','Item6':'Hrs\_treat',

'Item7':'Courteous','Item8':'Active\_listen'},inplace=True)

df.head()

df.info()

# drop all the demographic columns we don't need for this logistics regression analysis

df.drop(['City','State','County','Area','Zip','Lat','Lng','Population','TimeZone','Additional\_charges','TotalCharge','Interaction','UID','Customer\_id','Job','CaseOrder'],axis = 1,inplace=True)

# verify that all the columns were dropped before proceeding

df.info()

#check if there is any duplicate data entries present in columns

df[df.duplicated()]

# check if there are any duplicated columns in the data set - if there are none then the output should be False

df.columns.duplicated().any()

# check if there are any duplicated rows in the data set - if there are none then the output should be False

df.duplicated().any()

# convert categorical yes/no values to numeric 1/0 values

df = df.replace(to\_replace = ['Yes','No'],value = [1,0])

df

# convert the non-married Marital status values to "Married/Not Married", then convert "Married/Not Married" to "1/0"

#this will make the Marital variable easier to work with during regression analysis

df['Marital'] = df['Marital'].replace(['Divorced','Widowed','Separated','Never Married'],'Not Married')

df['Marital'] = df['Marital'].replace(['Married','Not Married'],[1,0])

df

# Showcase the unique values for the Services variable

df['Gender'].unique()

#convert the non-Female gender values to "Female/non-female", then convert "Female/non-female" to "1/0"

df['Gender'] = df['Gender'].replace(['Male','Nonbinary'],'non-female')

df['Gender'] = df['Gender'].replace(['Female','non-female'],[1,0])

df

# Showcase the unique values for the Services variable

df['Services'].unique()

# Drop the services variable since these values cannot be condensed

df.drop(['Services'],axis = 1,inplace=True)

# Showcase the unique values for the Complication\_risk variable

df['Complication\_risk'].unique()

# Drop the services variable since these values cannot be condensed

df.drop(['Complication\_risk'],axis = 1,inplace=True)

# Showcase the unique values for the Initial\_admin variable

df['Initial\_admin'].unique()

# convert the non-emergency admission status values to "Emergency Admission/non-Emergency Admission", then convert "Emergency Admission/non-Emergency Admission" to "1/0"

#this will make the Marital variable easier to work with during regression analysis

df['Initial\_admin'] = df['Initial\_admin'].replace(['Elective Admission','Observation Admission'],'non-Emergency Admission')

df['Initial\_admin'] = df['Initial\_admin'].replace(['Emergency Admission','non-Emergency Admission'],[1,0])

df

# describe the dataframe and showcase summary statistics of the variables

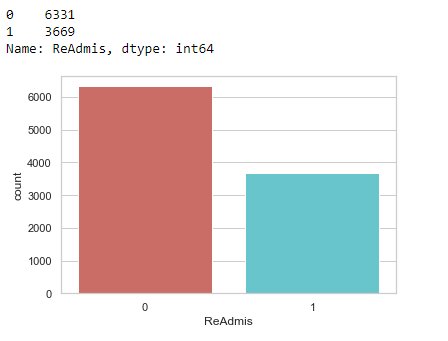
df.describe()

# now that all the modifications have been made, export the prepared dataset

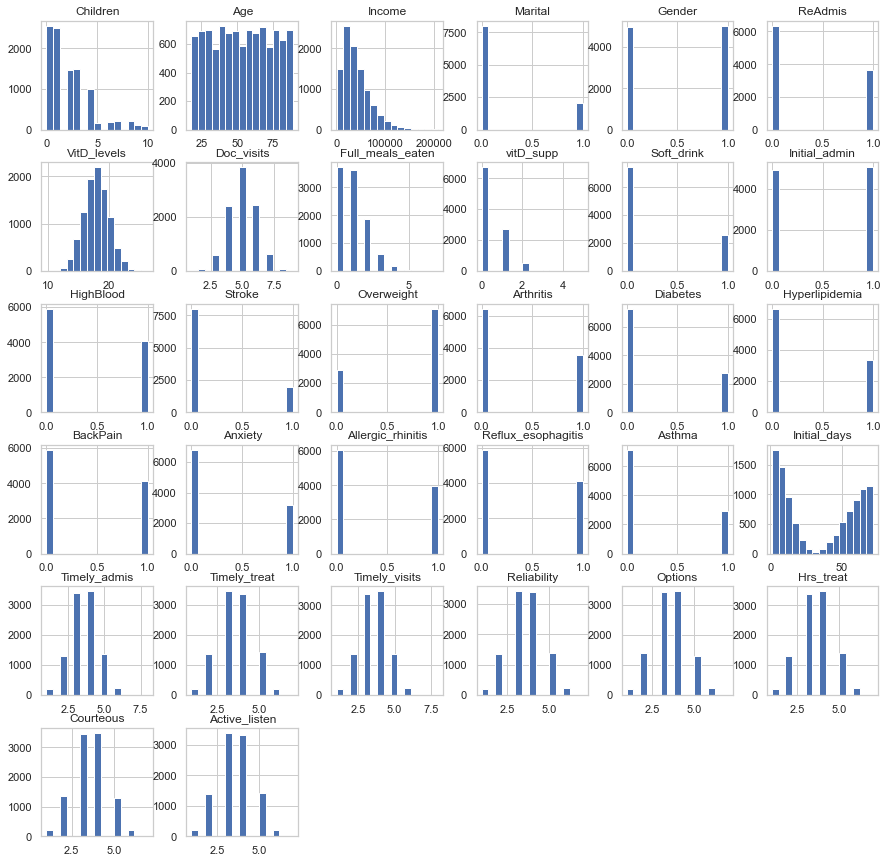
df.to\_csv(r'C:\Users\fahim\Documents\0\_WGUDocuments\d208\2medical\_clean-PREPAREDTASK2\_12-24-2022.csv')

**4. Univariate and bivariate visualizations of the distributions of variables**

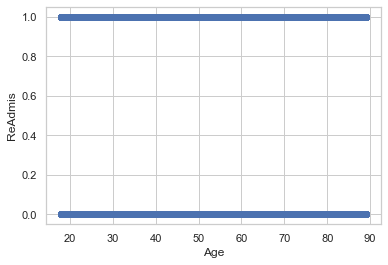
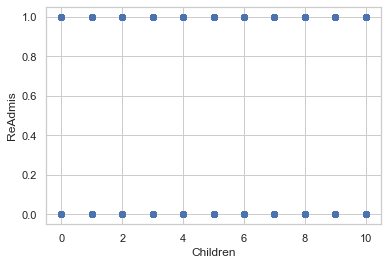
The cleaned data set was used to prepare visualizations of the variables we chose for the logistic regression. First we observed the target variable.

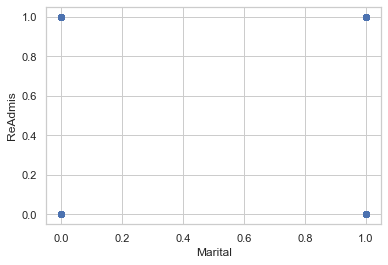
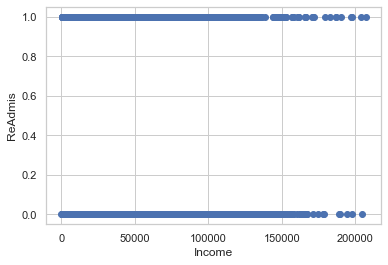


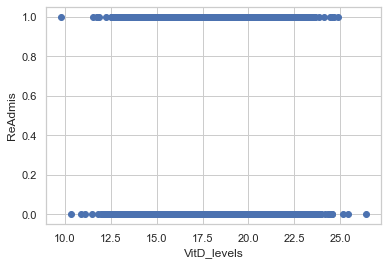
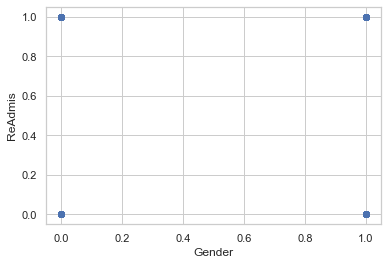
Here we see that the majority of patients were not readmitted. Next, we review the independent variables, using histograms.

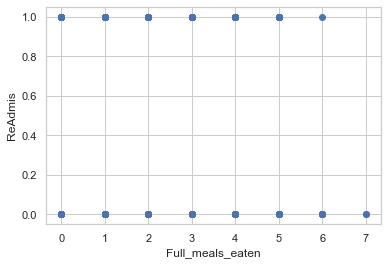
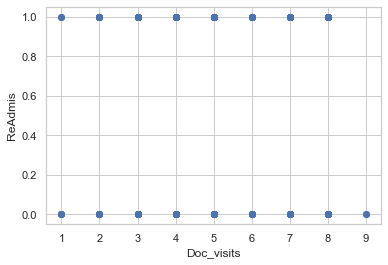


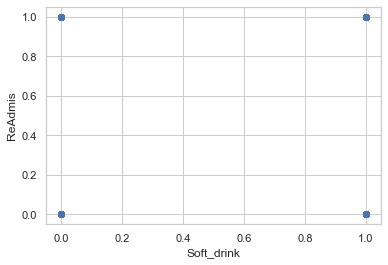
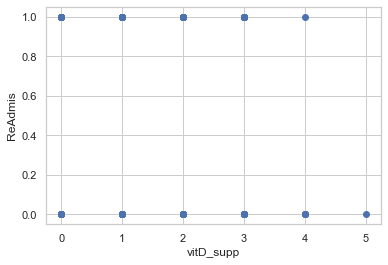
For bivariate visualizations, the target ReAdmis is compared with the potential predictors in the scatterplots shown below:

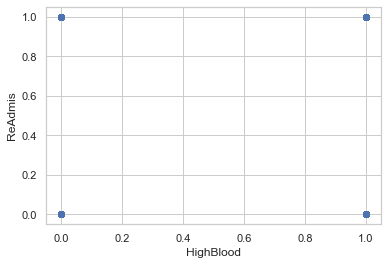
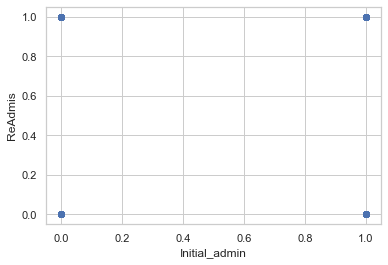


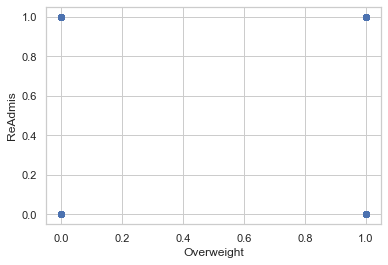
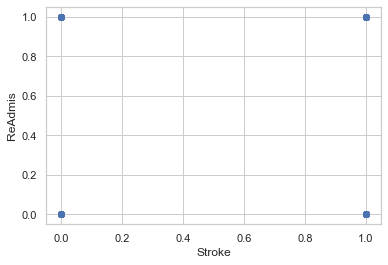


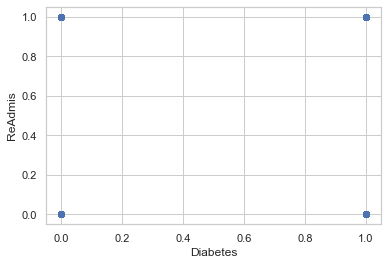
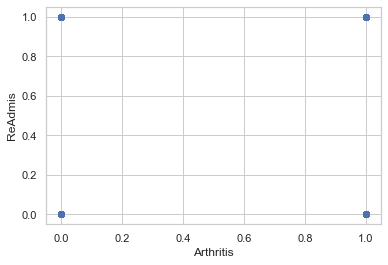


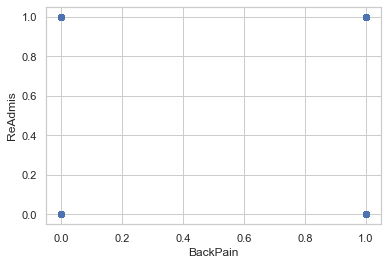
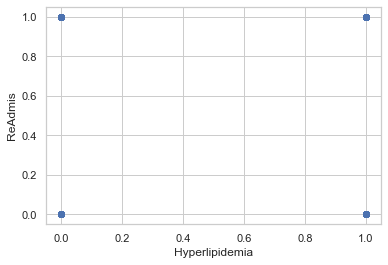


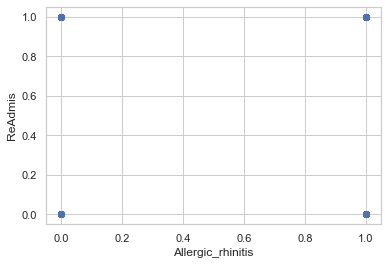
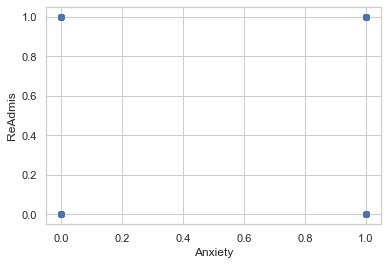


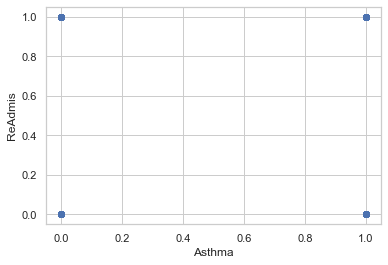
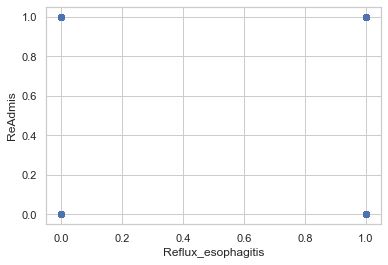


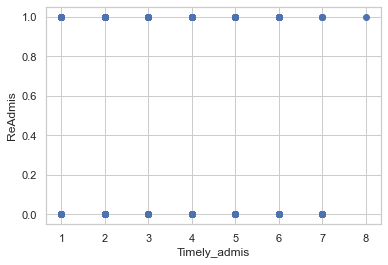
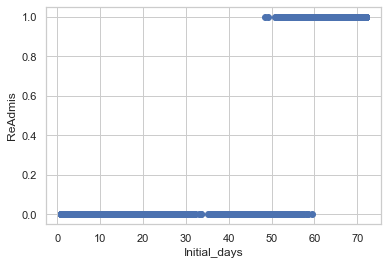


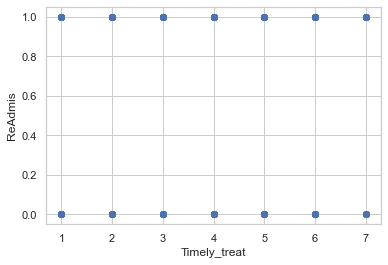


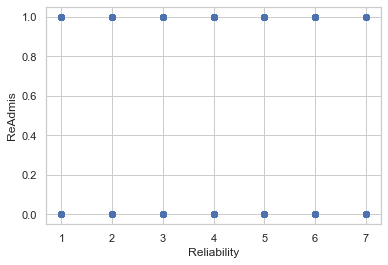
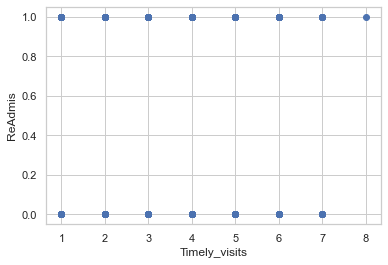


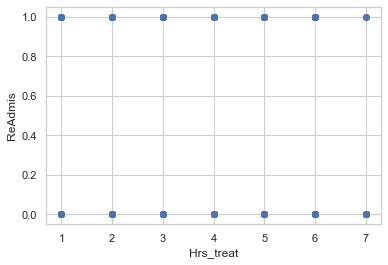
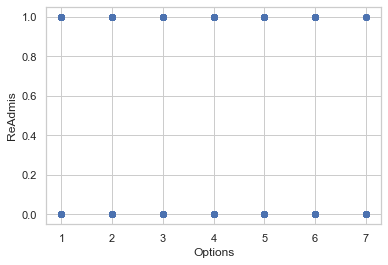


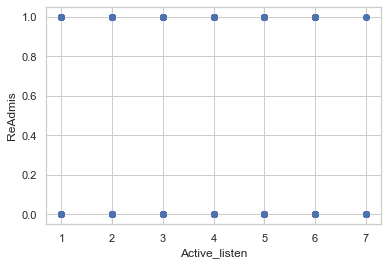
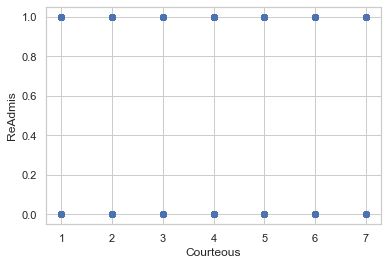












>the first model is SUPPOSED to be bad, so u can show us how to make it a better (not perfect) model

**5.  Provide a copy of the prepared data set.**

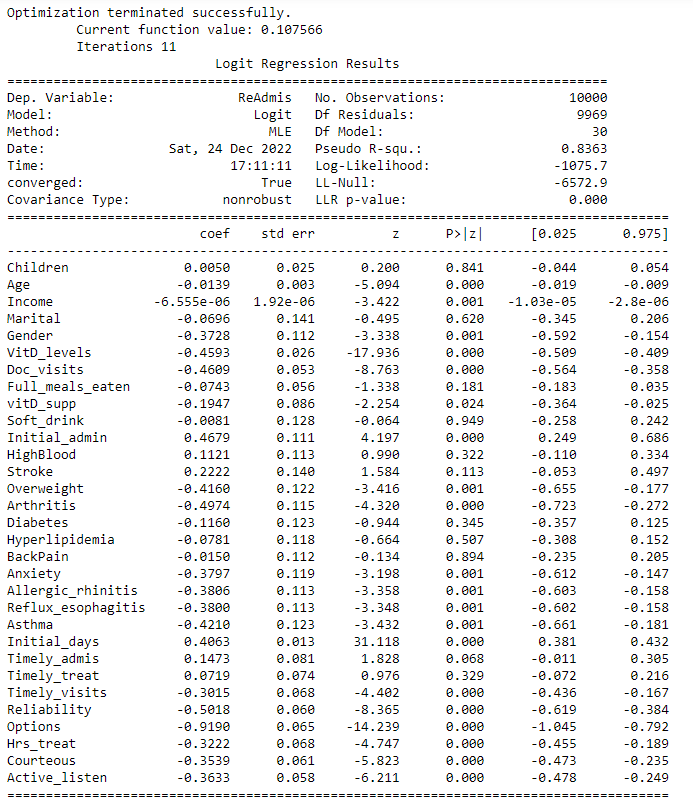
A copy of the cleaned data set, titled “2medical\_clean-PREPAREDTASK2\_12-24-2022.csv” is provided in the task submission.

**Part IV: Model Comparison and Analysis**

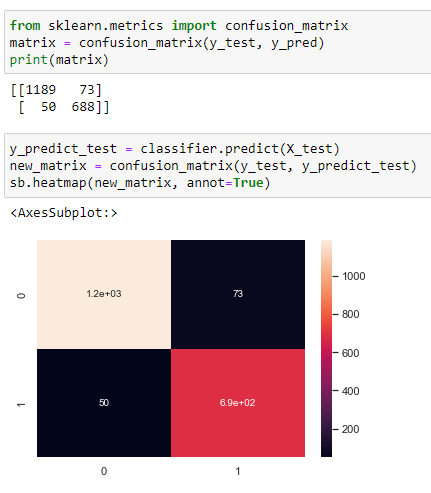
**D.  Comparison an initial and a reduced logistic regression**

**1. Initial logistic regression model**

Logit is used to run initial model. The output provided us with the P-value, coefficients, standard error values, and Z-value of the variables.

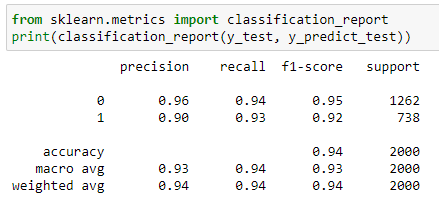
****

For us to better visualize the data, we also created the following confusion matrix:



**2.  Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question**.

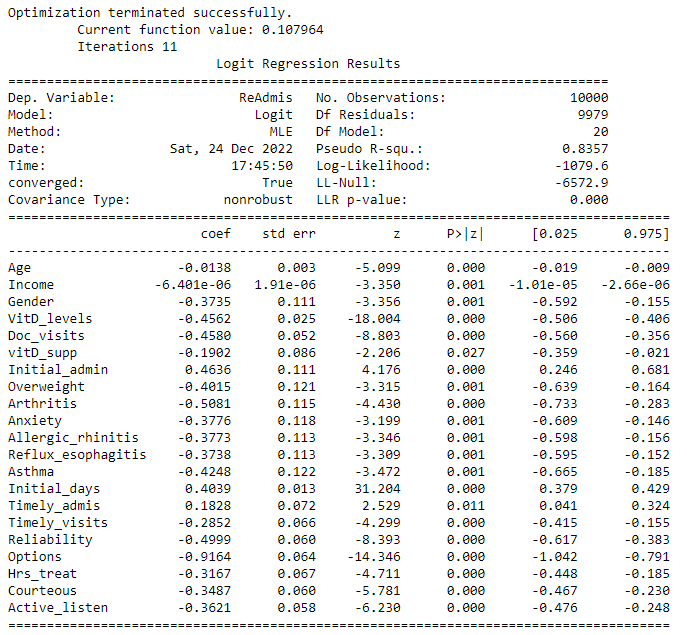
With the confusion matrix created, we generated a classification report to get the precision score of the initial model. This showed that the initial model predictions were 94% correct.



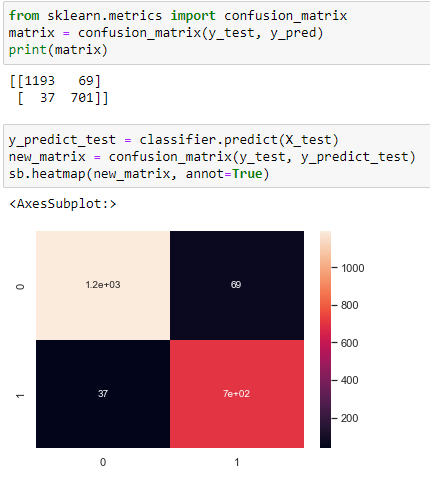
**Statistically based variable selection method:** To create our reduced model, the variables that had a P value below .05 statistical significance level were chosen. Variables with a P value above .05 statistical significance were dropped. The following variables were chosen for the reduced model: 'Age', 'Income', 'Gender', 'VitD\_levels', 'Doc\_visits', 'vitD\_supp','Initial\_admin', 'Overweight','Arthritis', 'Anxiety','Allergic\_rhinitis', 'Reflux\_esophagitis', 'Asthma', 'Initial\_days','Timely\_admis', 'Timely\_visits', 'Reliability','Options', 'Hrs\_treat', 'Courteous', 'Active\_listen'.

**3.  Reduced logistic regression model.**

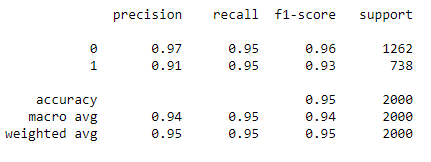
Now that we have our reduced model, we can proceed and run Logit again.



The new model provided us with this new confusion matrix and prediction percentage:

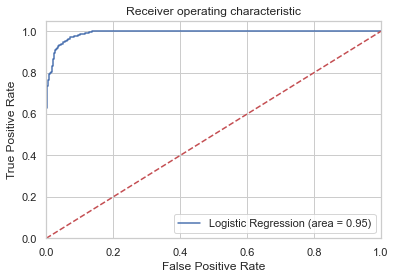


The classification report for the reduced model showed that the reduced model predictions were 95% correct, demonstrating a 1% increase in precision.



The Pseudo R squared value in the initial model was .8363 while the Pseudo R squared valuein the is .8357. This shows that the goodness of fit between the two models is almost the same. This also suggests that the variables we dropped may have statistically significant but only by a factor of .0006.

Using the reduced model and variables, we built a Receiver Operating Characteristic (ROC) graph. This linear slope is that of a random classifier and the blue curve is used to show classifiers true positive. If the blue arc is closer to the top left, then variables used have generated a greater confidence for the model. Because this is what we see with the graph produced, we can confirm our results with confidence.



**Code used to support the implementation of the logistic regression models.**

**E.  Analyze the data set using your reduced logistic regression model by doing the following:**

The full code for the project is provided at the end section **Part VI: Demonstration.** Furthermore, a pdf print of the Jupyter notebook used for running the python scripts is attached in task submission.

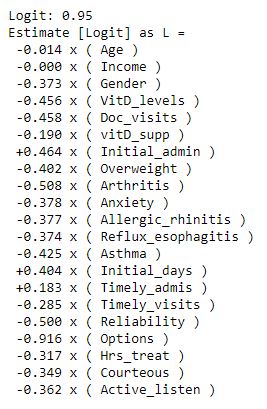
**Part V: Data Summary and Implications**

**F. Summary of findings and assumptions**

**Data analysis results:**

**Regression equation for the reduced model:**

After conducting the analysis, we created the following logistic regression equation for the reduced model:



**Interpretation of coefficients of the statistically significant variables of the model:**

A higher chance of readmission correlates with the positive coefficients, while a lower chance correlates with the negative coefficients. Based on our analysis, the following variables are linked to a patient’s higher chance of readmission: Initial\_admin, Initial\_days, and Timely\_admis. The remaining variables indicate a lower likelihood, but all variables used to create this model can potentially influence a patient’s chance of being readmitted.

**Statistical and Practical significance of the model:**

The final logistic regression model was both statistically and practically significant. We were able to determine which variables had the greatest impact on the likelihood of a patient being readmitted. We were also able to see the how when these variables increase or decrease, the chance of being readmitted also increases or decreases accordingly. The reduced model showed us that there is a strong correlation between a patient being readmittance and their initial administration, initial days spent at the hospital, and their timely admissions. The hospital can use our analysis to further investigate the correlation that a patient’s initial days, initial admission and timely admission has with their likelihood of being readmitted to the hospital.With the case of initial admissions, they could identify which of the readmitted patients were initial emergency admissions and which were not.These were strong predictor variables for a patient’s chance of being readmitted, so now the hospital could plan more efficient protocols involving these kinds of patients. They can also plan protocols based on the other strong predictor variables listed. By doing so, they could decrease the likelihood of patient readmission which will in turn help the hospital save time and resources. This can also benefit the patients because giving them the best treatment ahead of time based on their conditions will help to further reduce their chances of being readmitted. If any patient shows a positive result for any of the predictor variables, the hospital could plan ahead and anticipate their readmission in case preventive care or protocols aren’t enough. A patient who shows a combination of these variables would in turn have a higher chance of being readmitted. Knowing this could help the hospital further determine which patients are at the highest risk and they can plan treatments accordingly.

**Data analysis limitations:**

The first limitation we address is that the predictor variables are quite limited. They do not provide a diagnosis or solution to the higher readmission rates. Rather, they can be used to help predict if a patient has a high chance of readmission. These results are purely predictive and must be treated as such. Furthermore, the analysis is limited by the data set provided, as there could be additional variables that were not accounted for during the surveys. Our goal in this investigation was to see which variables provided could have correlated to a patient’s chance of readmission. Our investigation could be even more insightful if we had variables related to specific diseases and aliments that may have caused patients to return to the hospital, such as cancers. Additional variables related to patients could range from genetic history, recent exposure to certain viruses, additional health complications, and so on. The data provided is what the hospital deemed most important, so we are limited to work within that given framework. Lastly, as with any logistic regression, there is the potential of the model being overfit.

**Recommended course of action**

Improving the treatment process is the best thing to help decrease the likelihood of readmission. By making sure the patient receives the most accurate diagnosis and procedures as soon as possible, they in turn won’t have to spend too much time at the hospital getting treated. Reducing the time for their initial days spent at the hospital could also help to reduce their chances of being readmitted to the hospital. To address concerns of the model being overfit, we could run the regression multiple times or having additional analysts test for different variables could be beneficial to increase overall confidence. Lastly, we recommend a more detailed survey and adding more variables related to the specific reasons why patients were admitted into the hospital as well as what kind of treatment they’ll be receiving. For instance, patients who needed chemotherapy would need to be readmitted to the hospital since it’s a multi-appointment treatment. Knowing the relationship between why the patients are admitted to the hospital to begin with would prove valuable in further investigations and future models.

**Part VI: Demonstration**

**Link to the Panopto Video recording:**

[**https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=aba630d4-fffc-4db8-8660-af7600656d4a#**](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=aba630d4-fffc-4db8-8660-af7600656d4a)

**Sources for third party code:**

**Code for importing packages, preparing data, and performing logistic regression:**

<https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial/notebook>

<https://www.kaggle.com/code/miguelrodriguezolmos/plot-the-logistic-map-with-python-matplotlib/notebook>

<https://www.kaggle.com/code/vipulgandhi/linear-regression/notebook>

<https://www.kaggle.com/code/alexnystrom/hospital-data-logistic-regression>

<https://www.kaggle.com/code/neisha/heart-disease-prediction-using-logistic-regression>

<https://thepythonguru.com/python-string-formatting/>

<https://www.w3resource.com/pandas/dataframe/dataframe-itertuples.php>

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

<https://intellipaat.com/community/68715/converting-statsmodels-summary-object-to-pandas-dataframe>

>Code used for properly selecting variables

**References**

Zach. (2020, October 13). The 6 Assumptions of Logistic Regression (With Examples). Statology. Statology.org/assumptions-of-logistic-regression/

Kohli, S. (2019) Understanding a classification report for your machine learning model, Medium. Medium. Available at: <https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>.

**Full Code used for this Project**

~~Full Code used for D209 Task 2 Submission~~

~~DATA CLEANING AND PREPARATION CODE~~