**Performance Assessment: Task 2**

**A1. Research Question**

My research question for this performance assessment is, “What factors are associated with readmission?”

**A2. Goals**

The goal of this analysis is to determine whether there are variables within the provided data set that directly correlate with readmission to the hospital.

**B1. Summary of Assumptions**

There are several assumptions of a logistic regression model. One assumption is linearity of the log odds, meaning there is a linear relationship between the predictor variables and the log odds of the target variable. Another assumption is that the target variable is binary. A third assumption is that there is no multicollinearity, meaning no strong relationship between the predictor variables. Another assumption is that there are no extreme outliers in the data.

**B2. Tool Benefits**

For this performance assessment I chose to use Python. One benefit of using Python is that it is easy to read and build upon. Another benefit of Python is that the regression summary function includes more model evaluation metrics than R which makes it simpler to evaluate the model.

**B3. Appropriate Technique**

Multiple logistic regression is an appropriate technique for analysis because the target variable is a binary categorical variable. A logistic regression model allows me to analyze the relationship between readmission and my chosen predictor variables.

**C1. Data Cleaning Goals**

Upon loading the data, I included ‘keep\_default\_na=False’ to keep values of NONE from getting converted to nulls. I renamed the 8 survey response variables using the .rename() function to better align the names with the data and to make my life easier. The first goal of my data cleaning process was to assess for duplicates, which I accomplished using the .duplicated() function. Once I determined there were no duplicates, I assessed for nulls using the .isnull() function. After confirmation of no null values, I created boxplots of all quantitative values to assess for outliers using the seaborn .boxplot() function. There were 3 variables that contained outliers. To remove the outliers, I replaced the outliers with nulls and imputed the nulls with the median value.

**C2. Summary Statistics**

The dependent variable for my model is readmission. The twelve independent variables for my model are: children, age, income, vitamin D levels, doctor visits, full meals eaten, vitamin D supplements, initial days, total charges, additional charges, timely treatment, and active listening. I performed summary statistics on the data set for the quantitative variables using the .describe() function in order to quickly get the characteristics of each variable, including the full range and average value. The function cannot be used to assess categorical variables, so I also used .value\_counts() function to obtain the summary statistics for the target variable.

A table with numbers and text

Description automatically generated

**A table with numbers and text

Description automatically generated**

**A black text on a white background

Description automatically generated**

**C3. Visualizations**

**A graph with blue bars

Description automatically generated**

A graph of age and age of patients

Description automatically generatedA graph with blue squares

Description automatically generatedA blue graph with white text

Description automatically generatedA chart with a blue line

Description automatically generated with medium confidenceA graph of a doctor visits

Description automatically generatedA graph of a number of meals

Description automatically generated**A graph of vitamin d supplements

Description automatically generated**A graph of a number of days

Description automatically generatedA graph of blue and white lines

Description automatically generated with medium confidenceA graph of a number of patients

Description automatically generated with medium confidenceA graph of a survey response

Description automatically generatedA graph of blue rectangular objects

Description automatically generated with medium confidence**A graph with lines and dots

Description automatically generatedA diagram of a box plot

Description automatically generatedA diagram of a box plot

Description automatically generatedA diagram of a box plot

Description automatically generated**

**A graph with lines and a line in the middle

Description automatically generated with medium confidenceA graph with lines and dots

Description automatically generatedA graph with lines and dots

Description automatically generatedA chart with lines and dots

Description automatically generatedA graph with lines and numbers

Description automatically generated with medium confidenceA diagram of a box plot

Description automatically generatedA diagram of a survey

Description automatically generatedA graph with lines and dots

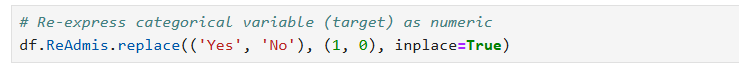
Description automatically generated**

**C4. Data Transformation**

According to the data dictionary provided with the data set, the survey responses were rated “on a scale of 1 to 8 (1 = most important, 8 = least important),” which felt counterintuitive. To make interpreting results of my analysis more straight forward, I chose to reverse the variables so that 1 = least important and 8 = most important using the .replace() function. I completed this prior to my univariate and bivariate analysis. I also re-expressed my categorical variable as numeric. Since the target variable contained ordinal data, I used the .replace() function again.

A screenshot of a computer program

Description automatically generated

****

**C5. Prepared Data Set**

See “D208\_PA2\_MV\_clean” csv file for prepared data set.

**D1. Initial Model**

A screenshot of a computer

Description automatically generated

**D2. Justification of Model Reduction**

After observing the regression results of my initial model, I tested for multicollinearity between the predictor variables by obtaining the variance inflation factor (VIF) for each. I removed the variable with the highest VIF value greater than 10, then reran the VIF values. I continued removing variables one by one until all VIF values were less than 10. From there, I used the backward stepwise elimination method to remove the least significant figures one by one until all the remaining predictor variables had p-values less than 0.05. This method ensured that I was left with a refined final model with statistically significant features.

**D3. Reduced Logistic Regression Model**

A screenshot of a computer

Description automatically generated

**E1. Model Comparison**

The LLR p-value for both my initial and final model is 0.00, implying that both regression models are meaningful. The pseudo r-squared value for my initial and final model are also very similar (0.932 vs. 0.926) which implies statistical significance since it is close to one.

**E2. Output and Calculations**

A screenshot of a computer program

Description automatically generated

**E3. Code**

See “D208 PA2 MV” ipynb file for code.

**F1. Results**

The regression equation for my final model is:

This equation can be interpreted as follows: Keeping all things constant, one unit increase in initial days would increase the log odds of readmissions by ~0.99. Keeping all things constant, one unit increase in additional charges would increase the log odds of readmissions by ~4.1\*10-5.

The p-values for all my variables are below 0.05. The LLR p-values for my final model is also below 0.05. The pseudo r-squared value is also acceptable and close to 1.00. Due to these results, my model qualifies as statistically significant. My model is also practically significant, as hospitals will want to know what factors are related to readmissions to reduce rates of readmission and decrease hospital spend.

**F2. Recommendations**

Since the model is statistically significant as is, I would work to implement this model and use it to predict readmissions based on initial days and additional charges. This model was a good start and showed that correlation does exist in the data. I would also further explore other variables in the data set and see if this model can be expanded to include other variables.

**G. Panopto Demonstration**

Link to video included with submission.

**H. Sources of Third-Party Code**

**I. Sources**