**Dr Keiona Middleton**

11-06-2022

D208

Performance assesement

Part 1 – Page 3

Part 2– Page 3

Part 3 – Page 3

Part 4 – Page 9

Part 5 – Page 11

Appendix H – Page 13

Appendix I – Page 13

**Part I: Research Question**

1.  We would like to prevent patients from being readmitted to the hospital. It is not good for either patients (waste of time and money) or us (one less bed for a new patient) to have patients readmitted. It would be ideal for us to make sure patient’s chances of getting readmitted are as low as possible at time of discharge.

2.  We would like to see if we can determine any factors that increase a patient’s risk for readmission. We can identify any patients that have the identified factors and make sure they can get to a better level of health. Hopefully, this would reduce the occurrence of readmission.

**Part II: Method Justification**

1.  The six assumptions of logistic regression are: binary response variable, independent observations, no multicollinearity, no extreme outliers, linear relationship between explanatory variables & logit of the response, and that the sample size is large enough.

2.  I chose Python simply out of comfort; R would be an equally good tool. Both Python and R could handle logistic regression, therefore both would be able to do this analysis. Python is capable of all the checks we want to do before logistic regression – Checking for nulls, making sure all data is in numerical format, and removing any outliers. All these pre checks make sure we have the most accurate regression. Of course, we can use Python to do the actual regression.

3.  Logistic Regression should be a good model as we are looking to predict a binary event (yes or no to patient getting readmitted).

**Part III: Data Preparation**

1. I would want to get rid of any unnecessary columns first. We were supplied with 50 data points. There’s no point in keeping any columns that aren’t going to be used in our analysis. I would make sure to get rid of any nulls or outliers as both would alter my analysis results. I would also have to change my categorical variables to numerical as we can’t do regression on categorical variables.

2.  Our target variable Readmission is currently a categorical variable. Age and doc visits are numerical. High Blood, Stroke, Overweight, and Complication risk are all categorical. Readmission, High Blood, Stroke, Overweight are data points for all patients. They are all answered with either yes or no. Complication Risk is answered is high, medium, low. About 40 percent of patients are medium, 30 percent are high, 20 percent are low. We have patients between 18 and 89 with an even distribution between age groups. We have 10000 observations. We have 7 data points. The mode for Re Admission is 0 or No. The mode for Complication Risk is 1 or Medium. The mean for Doc Visits looks to be around 5. The mode for High Blook is 0 or no. The mode for Stroke is 0 or no. The mode for overweight is 1 or yes.

3. We would drop columns with df.drop. We would check for nulls with df.isnull. We didn’t have any nulls values. We would get rid of any nulls with techniques like dropping rows or imputing values. We would change out categorical values to numerical values with replace. We would get rid of any outliers by checking z score and dropping any rows with a score higher than 3 or lower than -3.

4.  The univariate visualizations are below :

Chart, bar chart

Description automatically generated

Chart

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

The bivariate distributions are below

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated

Chart

Description automatically generated

5 The clean data is attached as medical\_clean\_task2dropped. A snippet of the top 10 rows is also below

Table

Description automatically generated

**Part IV: Model Comparison and Analysis**

1. The model with all predictors is below

A picture containing table

Description automatically generated

2. So we see that our pseudo r – squared is quite low. It is .0002577. Usually, 0.2 to 0.4 is a good range. I’m thinking we should use backwards stepwise elimination to remove some variables in the hopes that it will increase our pseudo r -squared. Typically, we want to have p-values under .05. So looking at the p values of our variables (P . |z|), we would remove Doctor Visits, Stroke, Complication Risk and High Blood. That leave us with Overweight and Age. Typically, they wouldn’t be considered either as their p-value is above .05 but they are the lowest.

3.  Graphical user interface, text

Description automatically generated

1.My initial variable selection was based on common sense / a hunch. It makes sense that older patients, high risk patients, patients that already had to see the doctor a lot, and overweight / patients with a stroke or high blood would all have a higher risk of being readmitted due their preexisting conditions. Our reduced model has our two most significant variables, removing all other less significant variables by means of backwards stepwise elimination. We can see Psuedo R-Squared has dropped from .0002577 to .0002455, so we explain less variability with our reduced model.

1. Here is the confusion matrix

Graphical user interface, text, application, email

Description automatically generated

1. Confusion matrix code is above the actual plots

**Part V: Data Summary and Implications**

1.Risk of readmission could be computed by using the following equation : ln (p / 1- p) = -.0631 + .0016 (Age) - .0386 (Overweight). This tells us a one unit increase in Age, increased log odds of admission by .0016. This also tells us being overweight decreases leg odds of readmission by .0386. Our model has very little statistical or practical significance. We actually see that our model explains less variability, as pseudo r squared dropped from .0002577 to .0002455 I think having a data table with mostly categorical variables is the main drawback of this data set.

2 I think there are two options here. The first might be try different variables (or even all data available to us). I think we’re on the right course but perhaps the initial variable choices were wrong. The second would be gather more data. For example, overweight could be anything between the patient only being a lb overweight vs 200 lbs overweight. Our data doesn’t allow us to make that distinction while we know there’s a big difference between those two levels of overweight.

H.  None Used

I.  None used