**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 12

Appendix H – Page 13

Appendix I – Page 13

**Part I: Research Question**

1.  We would like to determine if we can predict a patient’s total charges based on their prior health history and other data points available to us.

2. It would allow the hospital to predict how many money it would have coming in for a day. It would also allow the hospital to give patients an estimation of the cost of their visit.

**Part II: Method Justification**

1.  There are four assumptions for multiple linear regression. One is that variables are normally distributed. Normally distributed data would be data that is not skewed or have any outliers. Any skewed data or outliers would distort our findings. The second assumption is that there is a linear relationship between the independent variable and the dependent variables. If the relationship is not linear, our analysis wouldn’t show the true relationship between variables. Nonlinear regression would be a better tool for nonlinear relationships. The third assumption is correct data (reliably). If our data was incorrect or was measured incorrectly, we run the chance of seeing relationships that don’t exist. Finally, we assume homoscedasticity. This would mean we want the variance of error at all levels around our prediction. In simpler terms, any difference between the expected result at any point vs the actual result at any point should be around the same.

2.  I chose Python simply out of comfort; R would be an equally good tool. Both Python and R could handle multiple linear regression, therefore both would be able to do this analysis. Python can check for nulls, change data to numerical, and remove outliers. This will allow us to prep the data so we have the most accurate analysis. Python can also do the actual regression, making Python able to do all parts of the regression.

3.  Multiple Regression is appropriate to use as we are looking to determine the relationship between a dependent variable and multiple independent variables. Multiple regression seems to be a better tool than linear regression as we believe there should be a better relationship when we look at multiple variables vs a single variable. For example, we should expect an 85 year old diabetic to have a higher charge compared to a 21 year old diabetic as our younger diabetic would likely need less services. If we were only using linear regression, our relationship would only be able look at age with initial charges or diabetic with initial charges. With multiple regression, we can look at both age and diabetic in terms of initial charges.

**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 ones that I’m not looking at around. I would make sure to get rid of any nulls or outliers as both would alter my analysis. I would also have to change my categorical variables to numerical.

2.  Our target variable Total Charge is a numerical variable. Age and doc visits are numerical. High Blood, Stroke, Overweight, and Complication risk are all categorical. High Blood, Stroke, Overweight are all yes or no question. Age ranges between 18 and 89 and is distributed evenly. Total Charges ranges from around 2000 to 10000. It’s distribution is U- shaped, more data points at the low and high end . Doc Visits is pretty normally distributed (outside of very few patients recording 4 visits), and ranges from 1-9

3. We would drop columns with df.drop. We would check for nulls with df.isnull. We didn’t have any nulls but we have a few ways to get rid of nulls including 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 score higher than 3 or lower than -3.

4.  The univariate visualizations are below :

Chart, histogram

Description automatically generated

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

Graphical user interface

Description automatically generated

5 The clean data is attached as medical\_clean\_task1dropped.csv. A snippet of the first 10 rows is attached below as well

Application, table

Description automatically generated with medium confidence

**Part IV: Model Comparison and Analysis**

1. The model with all predictors is below

Graphical user interface, application

Description automatically generated

2. For a variable to be statistically significant, p-value should be under .05 (P > |t|) on the above chart. As we can see, Re Admission, Complication Risk, and High Blood are all under that. I will use those for the reduced model.

3. Table

Description automatically generated

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

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 three most significant variables, indicated by their lower p value.

1. Here are the residual plots

Graphical user interface, chart

Description automatically generated Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated

1. Residual plots code are above the actual plots

**Part V: Data Summary and Implications**

1. Total charges could be computed by using the following equation – Total Charges = (Re Admission \* 5084.17) + (1858.52 \* Complication Risk) + (1369.25 \* High Blood). This tells us that patients who have been readmitted would be charged 5084 more per day, patients with medium complication risks pay 1858 more per day, patients with high complication risks pay about 3717 more per day, and patients with high blood pay 1369 more per day. Our R-Squared is 42 which means these three variables explain 42% of the variability. I think we’d want a model with a higher r – squared to show to patients so we might be able to use this data for the hospital. A better model would be needed if we were to show this to patients.
2. Further analysis to try to discover more statistically significant variables could be done. That could be done by both looking at variables we already have or discovering new ones

H.  None Used

I.  None used