DATA MINING APPLICATIONS  
  
  
ALY6040, WINTER 2020  
MODULE 2 PROJECT ASSIGNMENT

Data Munging and Data Wrangling

SUBMITTED BY: KEYUR SHAH, ANUPREETA MISHRA, SHIVANI ADSAR,

NIKHIL SAKINAL, SHRADDHA GOPALKRISHNAN, JAYETA BISWAS

NUID: 001089242, 001050752, 001399374, 001069354, 001376444, 001068448

SUBMITTED TO: PROF. JUSTIN GROSZ

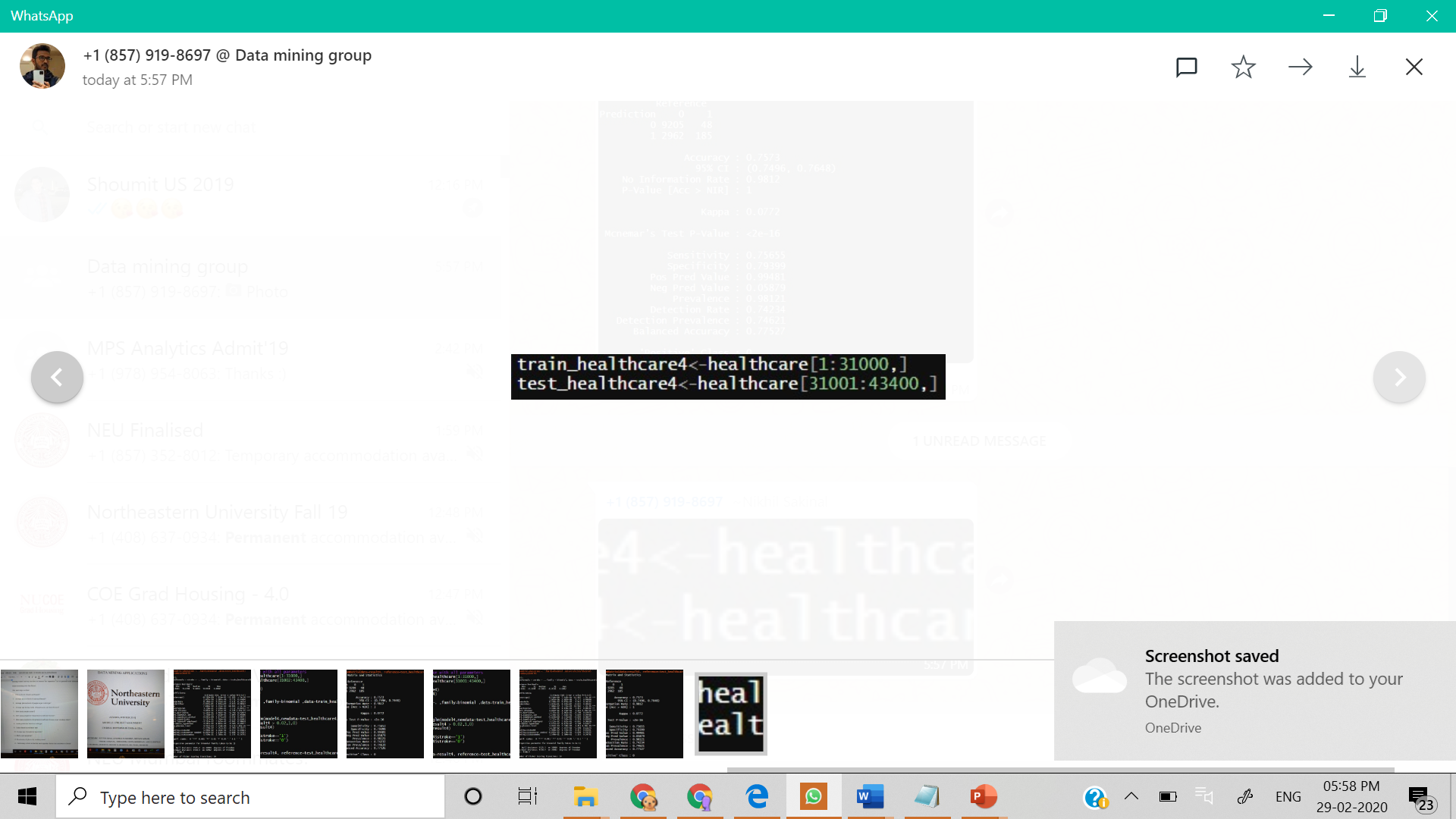
DATE: 03/01/2020

**Introduction**

In this assignment, we have worked on the “Healthcare Dataset Stroke Data” dataset obtained from Kaggle website. The data has been cleaned so we are now running early correlations on the data. We use logistic regression, which is basically a classification algorithm (1). In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). The analysis has helped us to understand the factors responsible for patients to suffer a heart stroke.

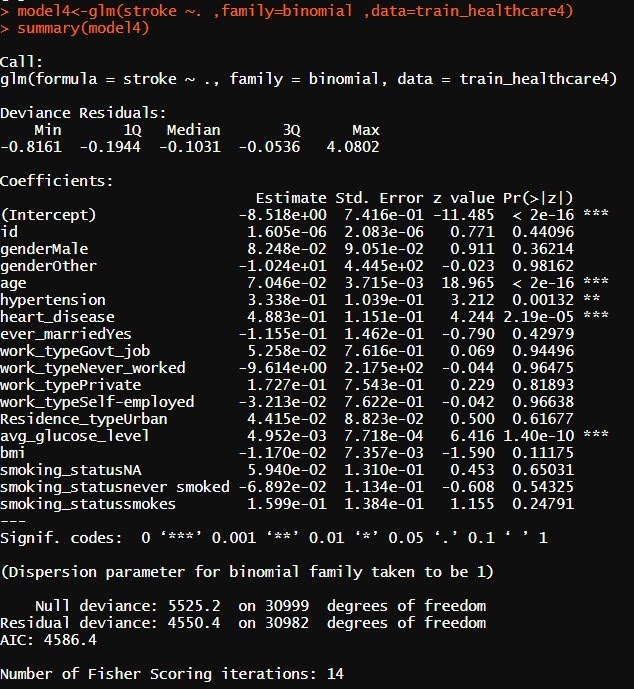
**Analysis**

Here, we follow according to the pre-model steps for the analysis as seen below:

1. We clean the data
2. We divide the data into two different datasets in the ratio 70(Training):30(Test) as seen in Fig 1.
   1. Training: Data Set used to build the Model
   2. Test: Simulation of live data

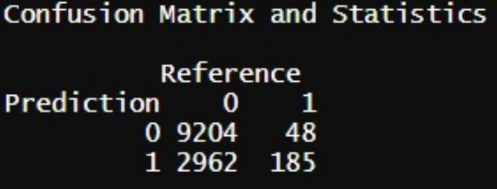
*Fig 1. Division of data*

1. Logistic Regression is the regression we used as we are looking for answering questions with binary output. In our dataset we are having stroke column which has 0(No stroke) and 1(Stroke)`. It is used to build a model based on binary dependent variable.
   1. For the dataset we have used logistic regression using R using packages tidyverse and caret. The reason we have used logistic regression is to obtain a binary classification to find whether a patient will have a stroke or not as seen in Fig 2.



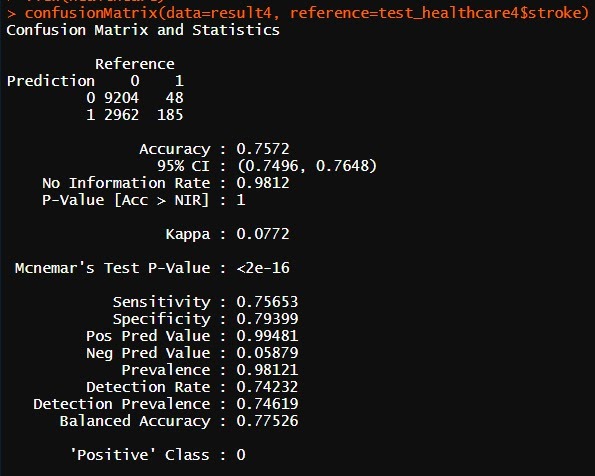
*Fig 2. Logistic Regression on Stroke Data*

* 1. From the above table we can see that age, heart disease, and glucose level are highly statistically significant factors which affect the stroke levels of patients. However, even though hypertension is significant but relatively less dominant than the other factors mentioned above. The remaining factors do not significantly affect stroke as their z-score values are very small.
  2. y (stroke) =  id + gender + age + hypertension + heart\_disease + ever\_married + work\_type + Residence\_type + avg\_glucose\_level + bmi + smoking\_status

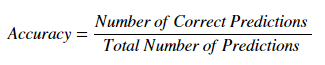
1. We have used Confusion Matrix to understand the accuracy of the model by observing the True Positives and False Positives. It helps us in summarizing the correct and incorrect predictions on our classifier. It portrays the cardinality of variables which are broken down by each class. Following is the matrix we have obtained using our logistic model as seen in Fig 3.

*Fig 4. Confusion Matrix Interpretation*

*Fig 3. Confusion Matrix Output*

* 1. We can infer from the confusion matrix that:

*Fig 5. Accuracy of the Model*

1. True Positive- Actual values were predicted correctly for 9204 patients
2. False Negative- 48 patients who were not supposed to have stroke were predicted incorrectly
3. True Negative- 2962 patients were correctly predicted to not have a stroke
4. False Positive- It is observed that 185 patients were not supposed to have a stroke but the model predicts they have a stroke
   1. As we can see the accuracy of the model is 75.72% which summarizes the performance of the classification algorithm, the formula used is:
5. Next, we intend on performing analysis using algorithms like, the K-Means Clustering, Decision Trees and Random Forest to enhance our predictions on the people who suffered stroke. We would check the homoscedasticity of different columns with a stroke. Performing LASSO to find the most important factors that contribute to Stroke.

**Conclusion**

* Age, heart disease, and glucose level are highly statistically significant factors which affect the stroke levels of patients.
* The accuracy of the model is 75.72% which summarizes the performance of the classification algorithm.
* Clearly, there are a lot of things left to analyze using algorithms like, the K-Means Clustering, Decision Trees and Random Forest

**References**

1. Eckerson, W. (n.d.). Secrets of Analytical Leaders. Retrieved from https://learning.oreilly.com/library/view/secrets-of-analytical/9781935504344/chap13.xhtml
2. The Internet Stroke Center. (n.d.). Retrieved from <http://www.strokecenter.org/patients/about-stroke/stroke-statistics/>