DATA MINING APPLICATIONS

Northeastern University

ALY6040, WINTER 2020

MODULE 1 PROJECT ASSIGNMENT

FINDING PATTERNS IN DATA & EDA

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: 02/23/2020

Introduction

In this assignment, we have worked on the "Healthcare Dataset Stroke Data" dataset obtained from Kaggle website. We have performed initial exploratory analysis using statistical methods in order to obtain insights from the data as visual representations. The analysis has helped us to understand the factors responsible for patients to suffer a heart stroke.

Analysis

We have obtained the dataset for our project from Kaggle website which belongs to owners Mckinsey & Company and Analytics Vidya. Since, healthcare is a very vast and complex field, we were interested to explore and perform analysis on a healthcare dataset. Moreover, Kaggle provides datasets that are capable of implementing machine learning algorithms and analysis. The dataset aims at predicting the causes and chances of patients to suffer a stroke. Some attributes on which the prediction of stroke depends are Age, Gender, BMI, Smoking-Status and Work-Type. The healthcare dataset on stroke data will help us solve issues in the healthcare sector as we would be performing analysis on the causes and factors affecting a patient to suffer a heart stroke.

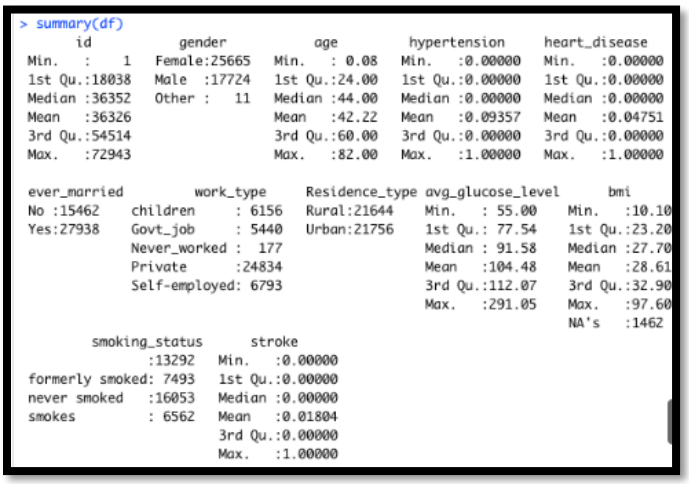
- It has been observed that, around 795,000 people every year in the US suffer heart stroke. Also, the risk of stroke to occur for a smoker is double than that of non-smokers.
- Moreover, patients with high blood pressure have higher chances of suffering a stroke. This dataset will help us in analyzing and predicting the factors that are responsible in suffering a stroke that could help medical industry in improving their treatments for heart patients.
- Hence, the analysis will help us to perform data cleaning, model building and carry out predictions.

Using the healthcare dataset, we are trying to solve the social problem which will enable us to predict the factor levels that are major causes of people suffering a heart stroke.

- This predictive approach will help medical industries to improvise on their facilities in preventing a heart stroke in patients.
- The model building will involve the analysis on factors like, Gender, Age, Hypertension, Glucose level and BMI which will help medical industries in understanding the causes of a heart stroke so that they can formulate prevention strategies at different medical institutions.

We have performed explanatory analysis using R Studio.

In order to perform data cleaning, we have analyzed the summary for the data. We have analysed the Healthcare dataset on the basis of certain factors like, BMI, Age, Gender, Smoking Status and Hypertension.



```
> summary(df)
       id          gender          age         hypertension     heart_disease
 Min.   :    1   Female:25665   Min.   : 0.08   Min.   :0.00000   Min.   :0.00000
 1st Qu.:18038   Male  :17724   1st Qu.:24.00   1st Qu.:0.00000   1st Qu.:0.00000
 Median :36352   Other :   11   Median :44.00   Median :0.00000   Median :0.00000
 Mean   :36326                  Mean   :42.22   Mean   :0.09357   Mean   :0.04751
 3rd Qu.:54514                  3rd Qu.:60.00   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :72943                  Max.   :82.00   Max.   :1.00000   Max.   :1.00000

 ever_married        work_type     Residence_type avg_glucose_level      bmi
 No :15462    children     : 6156   Rural:21644    Min.   : 55.00    Min.   :10.10
 Yes:27938    Govt_job     : 5440   Urban:21756    1st Qu.: 77.54    1st Qu.:23.20
              Never_worked :  177                  Median : 91.58    Median :27.70
              Private      :24834                  Mean   :104.48    Mean   :28.61
              Self-employed: 6793                  3rd Qu.:112.07    3rd Qu.:32.90
                                                   Max.   :291.05    Max.   :97.60
                                                                     NA's   :1462

       smoking_status      stroke
              :13292   Min.   :0.00000
 formerly smoked: 7493   1st Qu.:0.00000
 never smoked   :16053   Median :0.00000
 smokes         : 6562   Mean   :0.01804
                         3rd Qu.:0.00000
                         Max.   :1.00000
```

Observations: It can be interpreted that *bmi* has 1462 missing values which is ~3% of the dataset, the best way to get a better analysis was to take the average of the *bmi* and replace all the missing values with the average *bmi.*

- According to *National Aphasia Association* "smoking status does have an effect on stroke" hence we cannot have many missing values.
- In future analysis we figured that around 90% of the children opted did not answer their smoking status and out of the children who did answer 90% of them answered they never smoked.

- So, to perform accurate analysis, we have filled the smoking statues of the children who did not opt to answer with “never smoked” .

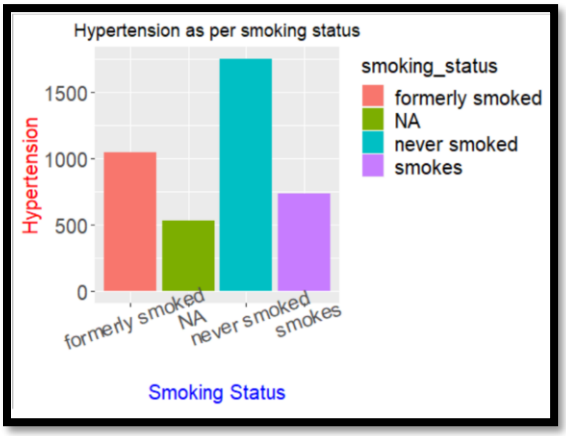
```
> Missing_Data_Smoking = round(100*sum(healthstroke_data_training$smoking_s
tatus ==="")/nrow(healthstroke_data_training),2)
> Missing_Data_Smoking
[1] 30.63
> Missing_Data_BMI = round(100*sum(healthstroke_data_training$bmi== "")/nro
w(healthstroke_data_training$bmi),2)
> Missing_Data_BMI
```

We can also see that 13,292 are blank under smoking status which is ~30% of the dataset.

```
> healthstroke_data_training=healthstroke_data_training %>%
+   mutate(bmi= ifelse(is.na(bmi),mean_bmi,bmi))
> View(healthstroke_data_training$bmi)
>
> #seperate model for with Smoking data and without smoking data and joinin
g the Stroke data
```

We have replaced the null values in BMI column with the mean of those values, since they contributed to 3% of the dataset.

```
> mean_bmi = mean(healthstroke_data_training$bmi,na.rm = T)
> mean_bmi
[1] 28.60504
```



Observation: The histogram shows the comparisons of hypertension as per the smoking status of people. It can be seen that the hypertension of people who have “Never Smoked” is the highest and people who “Smoke” is the lowest.

Fig.1: Histogram for Hypertension vs. Smoking Status

Observation: The histogram shows the comparisons of Heart Disease Count as per the Smoking Status. It can be seen that, people who formerly smoked have higher chances of heart disease count and those who smoke have a lower chance of heart disease.

Fig.2: Histogram for Heart Disease vs. Smoking Status

In order to perform accurate analysis on the patients data, we would perform regressions and build different models. Moreover, 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 homoscedasticity of different columns with stroke. Performing LASSO to find the most important factors that contribute to Stroke.

Conclusion

- The hypertension of people who have “Never Smoked” is the highest and people who “Smoke” is the lowest.
- People who formerly smoked have higher chances of heart disease count and those who smoke have a lower chance of heart disease.
- Clearly, there are a lot of things left to analyze using algorithms like, the K-Means Clustering, Decision Trees and Random Forest

References

1. The Internet Stroke Center. (n.d.). Retrieved from http://www.strokecenter.org/patients/about-stroke/stroke-statistics/
2. Bulman, A. G. (n.d.). Probability and Counting Rules. In ELEMENTARY STATISTICS: A STEP BY STEP APPROACH, TENTH EDITION (10th ed., p. A-549). New York
3. Siegle, D. (2015, June 14). ANOVA, Regression, and Chi-Square. Retrieved from https://researchbasics.education.uconn.edu/anova_regression_and_chi-square/#.
4. Evans, James. R. (n.d.). University of Cincinnati. In S TATISTICS, DATA ANALYSIS, AND DECISION MODELING, FIFTH EDITION (5th ed., p. A-229). International.