CAPSTONE GROUP - 6

Final Project

ALY 6020 Predictive Analytics

SRI SRAVYA TAMMINEEDI,

KEYUR RASHMIKANT SHAH

Northeastern University

Date: 06/25/2020

Note:

Sri Sravya Tammineedi, Keyur Rashmikant Shah College of Professional Studies, Northeastern University, Boston, MA 02115.

This report was created as a part of the project to gain hands-on experience in classification using the Decision Trees and rules on R Programming.

Instructor Name: Richard He

Sri Sravya Tammineed, Keyur Rashmikant Shah is now a student at Department of Analytics, Northeastern University

Contact: tammineedi.s@husky.neu.edu, shah.keyur@husky.neu.edu

NUID: 001086975, 001089242

INTRODUCTION

The dataset "Healthcare Dataset Stroke Data" is taken from the Kaggle site http://www.strokecenter.org/[2]. The data has been cleaned so we are presently running early correlations on the data.

Logistic regression [4]: It is fundamentally a classification algorithm. In regression analysis, logistic regression is assessing the boundaries of a logistic model (a type of binary regression). The linear regression model is a straight association between the dependent variable with no change and the independent variable. Generalized Linear Model (GLM) is a direct blend of autonomous factors that address the needy variable. The traditional sort of GLM is a fundamental linear regression. Essential linear regression works splendidly when the dependent variable is commonly distributed.

Random forests [1]: It is also known as random decision forests are an outfit learning technique for classification, regression, and different undertakings that work by developing a large number of choice trees at training time and yielding the class that is the method of the classes (classification) or mean prediction regression of the individual trees.

Decision Trees [3]: Decision trees will by and large be the methodology for choice for predictive modelling since they are modestly direct and are also particularly reasonable. The basic goal of a choice tree is to section a populace of information into littler segments. A regression tree is used to predict continuous quantitative information.

The analysis has helped us to comprehend the elements answerable for patients to endure a heart stroke with the accuracy of the models. The analysis of this assignment is done using R studio in R language.

ANALYSIS

The main aim of this analysis to create a model that can predict what kind of patients are more prone to get a stroke in order to help medical industry in improving heart treatments.

1) **Collecting the Data:** Data set has 12 columns and 43, 400 rows**.**

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	30669	Male	3.00	0	0	No	children	Rural	95.12	18.0		0
2	30468	Male	58.00	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
3	16523	Female	8.00	0	0	No	Private	Urban	110.89	17.6		0
4	56543	Female	70.00	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
5	46136	Male	14.00	0	0	No	Never_worked	Rural	161.28	19.1		0
6	32257	Female	47.00	0	0	Yes	Private	Urban	210.95	50.1		0
7	52800	Female	52.00	0	0	Yes	Private	Urban	77.59	17.7	formerly smoked	0
8	41413	Female	75.00	0	1	Yes	Self-employed	Rural	243.53	27.0	never smoked	0
9	15266	Female	32.00	0	0	Yes	Private	Rural	77.67	32.3	smokes	0
10	28674	Female	74.00	1	0	Yes	Self-employed	Urban	205.84	54.6	never smoked	0
11	10460	Female	79.00	0	0	Yes	Govt_job	Urban	77.08	35.0		0
12	64908	Male	79.00	0	1	Yes	Private	Urban	57.08	22.0	formerly smoked	0
13	63884	Female	37.00	0	0	Yes	Private	Rural	162.96	39.4	never smoked	0
14	37893	Female	37.00	0	0	Yes	Private	Rural	73.50	26.1	formerly smoked	0
15	67855	Female	40.00	0	0	Yes	Private	Rural	95.04	42.4	never smoked	0
16	25774	Male	35.00	0	0	No	Private	Rural	85.37	33.0	never smoked	0
17	19584	Female	20.00	0	0	No	Private	Urban	84.62	19.7	smokes	0
18	24447	Female	42.00	0	0	Yes	Private	Rural	82.67	22.5	never smoked	0
19	49589	Female	44.00	0	0	Yes	Govt_job	Urban	57.33	24.6	smokes	0
20	17986	Female	79.00	0	1	Yes	Self-employed	Urban	67.84	25.2	smokes	0

Fig 1: Loading the Data

2) **Data Preparation:** Bmi (Body Mass Index) has 1,462 NA (~3%), Hence mean of the bmi was used to replace the NA values. Smoking status has 30% NA values which are replaced with a relationship between work type and smoking status. 90% of the children did not mention their smoking status and out of the children who mentioned their smoking status never smoked. So, for better analysis, we filled the smoking status of the children who did not mention their smoking status to never smoked.

bmi	smoking_status	stroke
18.00000	never smoked	0
39.20000	never smoked	0
17.60000	never smoked	0
35.90000	formerly smoked	0
19.10000	never smoked	0
50.10000	never smoked	0
17.70000	formerly smoked	0
27.00000	never smoked	0
32.30000	smokes	0
54.60000	never smoked	0
35.00000	never smoked	0
22.00000	formerly smoked	0
39.40000	never smoked	0
26.10000	formerly smoked	0

Fig 2: Replacing the Null Values

```
> summary(healthcare)
       id            gender               age         hypertension heart_disease ever_married  work_type        
 Min.   :    1   Length:43400       Min.   : 0.08   0:39339      0:41338       No :15462    Length:43400      
 1st Qu.:18039   Class :character   1st Qu.:24.00   1: 4061      1: 2062       Yes:27938    Class :character  
 Median :36352   Mode  :character   Median :44.00                                           Mode  :character  
 Mean   :36326                      Mean   :42.22                                                             
 3rd Qu.:54514                      3rd Qu.:60.00                                                             
 Max.   :72943                      Max.   :82.00                                                             
 Residence_type avg_glucose_level      bmi        smoking_status        stroke   
 Rural:21644    Min.   : 55.00    Min.   :10.10   Length:43400       0:42617  
 Urban:21756    1st Qu.: 77.54    1st Qu.:23.40   Class :character   1:  783  
                Median : 91.58    Median :28.10   Mode  :character            
                Mean   :104.48    Mean   :28.61                               
                3rd Qu.:112.07    3rd Qu.:32.60                               
                Max.   :291.05    Max.   :97.60                               
```

Fig 3: Summary of the data with Null values replaced

3) **Exploratory data analysis (EDA):** The bar plot shows the comparison of hypertension as per the smoking status of people. Hypertension of patients who never smoked is highest and patients who smoke is lowest.

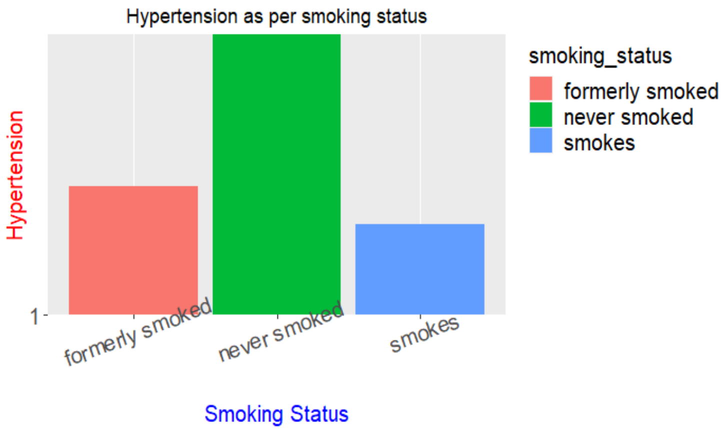

Fig 4: Bar plot of Hypertension

Patients who never smoke have higher chances of heart diseases count and those who smoke have less chance of heart diseases. This is because, patients who never smoked, assumes that only people who smokes will get effected and will not be careful but they started showing some symptoms of heart diseases slowly.

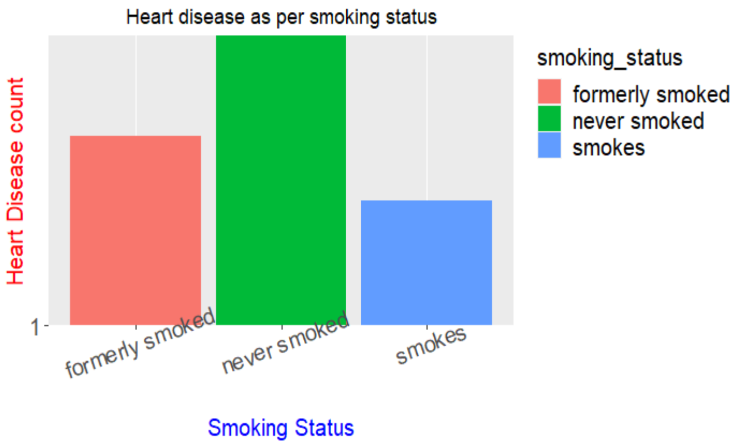

Fig 5: Bar plot of Heart Disease

The bar plot shows the relation between the age group and the smoking status for the prediction of the stoke. We can observe that the formerly smoked are the highest number with average age group of 50 years old and the least is the never smoked people with the average age group of below 40 years.

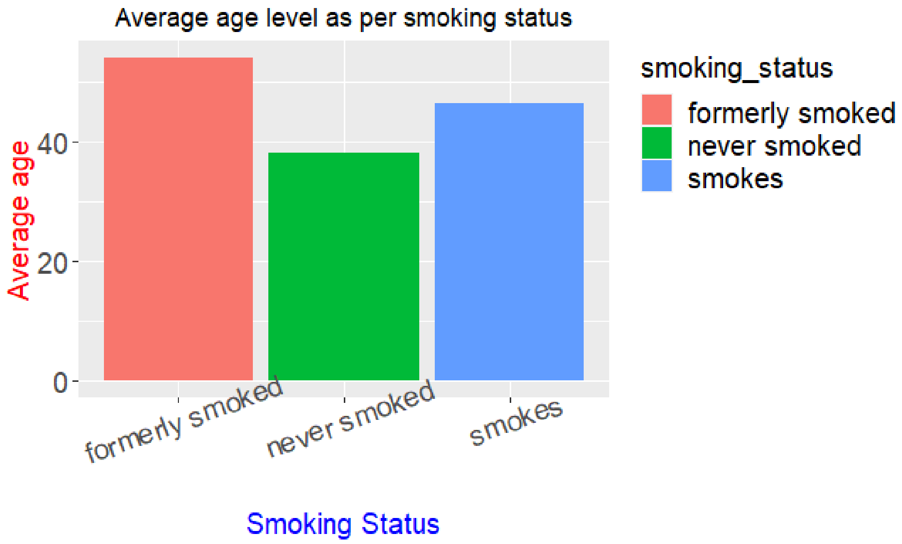

Fig 6: Bar plot of age

4) **Models:** The data is divided into training and test data. 80:20 where 80% is training data and 20% is test data. For the better accuracy we performed the analysis on the training data set and data validation in test data.

```
> train_healthcare<-healthcare[1:39000,]
> test_healthcare<-healthcare[39001:43400,]
> str(train_healthcare)
'data.frame':	39000 obs. of  12 variables:
 $ id               : int  30669 30468 16523 56543 46136 32257 52800 41413 15266 28674 ...
 $ gender           : chr  "Male" "Male" "Female" "Female" ...
 $ age              : num  3 58 8 70 14 47 52 75 32 74 ...
 $ hypertension     : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 2 ...
 $ heart_disease    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ ever_married     : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 2 2 2 2 ...
 $ work_type        : chr  "children" "Private" "Private" "Private" ...
 $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 1 2 2 1 1 2 2 1 1 2 ...
 $ avg_glucose_level: num  95.1 88 110.9 69 161.3 ...
 $ bmi              : num  18 39.2 17.6 35.9 19.1 50.1 17.7 27 32.3 54.6 ...
 $ smoking_status   : chr  "never smoked" "never smoked" "never smoked" "formerly smoked" ...
 $ stroke           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
> str(test_healthcare)
'data.frame':	4400 obs. of  12 variables:
 $ id               : int  28742 60600 19420 50472 15768 67042 22372 33477 595 19047 ...
 $ gender           : chr  "Female" "Male" "Male" "Female" ...
 $ age              : num  78 22 82 32 37 4 17 31 56 52 ...
 $ hypertension     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ heart_disease    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ ever_married     : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 1 1 2 2 2 ...
 $ work_type        : chr  "Private" "Never_worked" "Private" "Govt_job" ...
 $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 1 2 2 2 1 1 2 2 1 1 ...
 $ avg_glucose_level: num  89.1 151 90.9 59.2 85.8 ...
 $ bmi              : num  33.3 25 25.7 24.8 36.3 14.3 28.3 22.2 38.4 22.3 ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
 $ stroke           : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
```

Fig 7: Dividing of the data

1) **Logistic Regression:** We have a binary classification in the outcome patient suffers stroke or not hence we used Logistic regression.

   Model Fitting: Factors that significantly affect stroke are age, hypertension, heart disease, and glucose level and it can be seen through higher z score value.

```
> summary(model)

Call:
glm(formula = stroke ~ ., family = binomial, data = train_healthcare)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8269  -0.1950  -0.1053  -0.0555   4.1364

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -8.717e+00  7.285e-01 -11.966  < 2e-16 ***
id                      8.446e-07  1.860e-06   0.454 0.649746
genderMale              1.005e-01  8.005e-02   1.255 0.209444
genderOther            -1.124e+01  7.332e+02  -0.015 0.987765
age                     6.854e-02  3.276e-03  20.921  < 2e-16 ***
hypertension1           3.186e-01  9.317e-02   3.419 0.000628 ***
heart_disease1          5.977e-01  1.008e-01   5.930 3.03e-09 ***
ever_marriedYes        -1.192e-01  1.302e-01  -0.915 0.359949
work_typeGovt_job       4.072e-01  7.511e-01   0.542 0.587746
work_typeNever_worked  -9.556e+00  1.832e+02  -0.052 0.958398
work_typePrivate        5.052e-01  7.453e-01   0.678 0.497860
work_typeSelf-employed  3.924e-01  7.512e-01   0.522 0.601433
Residence_typeUrban     3.896e-03  7.872e-02   0.049 0.960531
avg_glucose_level       4.257e-03  6.942e-04   6.132 8.69e-10 ***
bmi                    -1.114e-02  6.571e-03  -1.696 0.089954 .
smoking_status          2.826e-02  4.153e-02   0.680 0.496224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6927.5  on 38999  degrees of freedom
Residual deviance: 5724.5  on 38984  degrees of freedom
AIC: 5756.5

Number of Fisher Scoring iterations: 15
```

Fig 7: Summary of Logistic Regression

Model Evaluation: Confusion Matrix allowed us to compare the correct and incorrect number of predictions made by our logistic model.

This classification algorithm gave an accuracy of **76.34%** with 95% confidence interval.

- True Positive-Actual values were predicted effectively for 3282 patients.
- False Negative-17 patients who shouldn't have strokes were predicted erroneously.
- True Negative: 1024 patients were effectively anticipated to not have a stroke
- False Positive-It is observed that 77 patients shouldn't have a stroke, however, the model predicts they have a stroke.

```
> confusionMatrix(data=result, reference=test_healthcare$stroke)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3282   17
         1 1024   77

               Accuracy : 0.7634
                 95% CI : (0.7506, 0.7759)
    No Information Rate : 0.9786
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0932

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.76219
            Specificity : 0.81915
         Pos Pred Value : 0.99485
         Neg Pred Value : 0.06994
             Prevalence : 0.97864
         Detection Rate : 0.74591
   Detection Prevalence : 0.74977
      Balanced Accuracy : 0.79067

       'Positive' Class : 0
```

Fig 8: Accuracy of the Model

2) **Random Forest:** We perform this method on the target variable which is "stroke" and the significant variables. This classification algorithm gave an accuracy of **98%** with **2.16%** Error.

```
Call:
 randomForest(formula = stroke ~ age + hypertension + heart_disease +      avg_glucose_level, data = train_healthcare, mtry = 5, importance = TRUE,      do.trace = 100)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 2.16%
```

Fig 9: Out of bag (OOB) Error rate

The true positive of the model is

	Class 0 :No stroke	Class 1:Stroke
Class 0 :No Stroke	True Positive	False Negative
Class 1:Stroke	False Positive	True Negative

Fig 10: Interpretation of Confusion matrix

- True Positive-Actual values were predicted effectively for 4295 patients.
- False Negative-94 patients who shouldn't have strokes were predicted erroneously.
- True Negative: 11 patients were effectively anticipated to not have a stroke.

```
> confusionMatrix(data=health_pred, reference=test_healthcare$stroke)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4295   94
         1   11    0

               Accuracy : 0.9761
                 95% CI : (0.9712, 0.9804)
    No Information Rate : 0.9786
    P-Value [Acc > NIR] : 0.8835

                  Kappa : -0.0045

 Mcnemar's Test P-Value : 1.22e-15

            Sensitivity : 0.9974
            Specificity : 0.0000
         Pos Pred Value : 0.9786
         Neg Pred Value : 0.0000
             Prevalence : 0.9786
         Detection Rate : 0.9761
   Detection Prevalence : 0.9975
      Balanced Accuracy : 0.4987

       'Positive' Class : 0
```

Fig 10: Accuracy of the Model

3) **Decision tree:** The default class variable is the eleventh section in train_healthcare, so we need to remove it from the training data outline anyway gracefully it as the target factor vector for a course of action. The healthmodel object right now contains a C5.0 choice tree. We can see some central information about the tree by creating its name as underneath:

```
Call:
C5.0.default(x = train_healthcare[-11], y = train_healthcare$stroke)

Classification Tree
Number of samples: 24000
Number of predictors: 11

Tree size: 2

Non-standard options: attempt to group attributes
```

Fig 11: Classification of the trees

Confusion matrix of the model gives us the accuracy of 100% with confidence interval of 95%.

- True Positive-Actual values were predicted effectively for 5971 patients.
- False Positive-It is observed that 137 patients shouldn't have a stroke, however, the model predicts they have a stroke.

```
> confusionMatrix(data=health_pred, reference=test_healthcare$stroke)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5971    0
         1    0  137

               Accuracy : 1
                 95% CI : (0.9994, 1)
    No Information Rate : 0.9776
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.9776
         Detection Rate : 0.9776
   Detection Prevalence : 0.9776
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

Fig 12: Accuracy of the Model

CONCLUSION

- Age, hypertension, heart disease, and glucose level are highly statistically significant factors which affect the probability of stroke in patients.
- Logistic Regression gave us the important factors for the analysis to predict the problem and accuracy is 76.34%.
- Random forest leads to more accurate results with 98% accuracy; hence it is a better model than logistic regression.
- Decision tree is a better model for our data with accuracy of 100% than other two models.

Reference:

1. Random forest. (2020, June 22). Retrieved June 25, 2020, from https://en.wikipedia.org/wiki/Random_forest.
2. The Internet Stroke Center. (n.d.). Retrieved June 25, 2020, from http://www.strokecenter.org/.
3. Decision Trees: An Overview. (2019, September 21). Retrieved June 07, 2020, from https://www.aunalytics.com/decision-trees-an-overview/.
4. Learn Generalized Linear Models (GLM) using R. (n.d.). Retrieved June 24, 2020, from https://www.kdnuggets.com/2017/10/learn-generalized-linear-models-glm-r.html.