Factors that affect treatment in a Hospital using Machine Learning techniques

Saurabh Jain

Department of Computer Science, University of Southern California,

Los Angeles, California 90089, USA

(Dated: February 25, 2021)

Abstract

Machine Learning can be explained as a method that interpret results from exploring a data set, proposing and fitting a predictive model and evaluating the model's performance. This assignment focuses on the data set that describes the probability of treatment of patients of the hospital based on the factors such as age, blood pressure, family history, some tests and some other genes. The maximum number of patients are in the age group 55-70. Majority of the patients have their blood pressure in the range 85-105. There is good correlation between age and the values of TestA. As the age increases of a patient, the value of TestA increases. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Running Logistic regression for the hospital's data set gives a 74% accuracy, which is an excellent accuracy to predict the future patients' treatment requirements. One of the most interesting factors to be considered is that the most important factors are age, gender, blood pressure and family history.

I. INTRODUCTION

The main task of this assignment is to train a classification model and measure how good that model is. Recommendations are being provided on which factor is the most important to detect if the patient requires a treatment or not. Accuracy, precision, F1-score, AUC score were calculated along with computing regularization, logistical regression and cross validation. Feature engineering and Feature importance was also determined using various methods. Before any computation could be carried out, data was keenly explored, processed and cleaned. Several different methods were used for logistical regression and the one with the best overall efficiency was selected to compute the false positives and the false negatives.

The data consisted of 12 different columns with treatment being the binary outcome variable. The major factors involved - age, blood pressure, and the family history.

II. DATA EXPLORATION

The data consisted of the following factors that were used to compute the logistic regression:

1. treatment: dependent variable - binary
2. age: The minimum and the maximum age is 29 and 93, respectively. The average age is approximately 60. Majority of the patients are in the 55-70 age group, see Fig. 1.
3. blood_pressure: The average blood pressure value in the data set is approximately 101. The minimum is -999, which is an outlier, and such outliers have been cleaned. The maximum blood pressure is 157, which is quite unhealthy. The maximum number of people lie in the 85-105 range, which is quite healthy, see Fig. 2.
4. family_history: boolean value (False: 4824; True: 69)
5. TestA
6. TestB
7. geneC
8. geneD

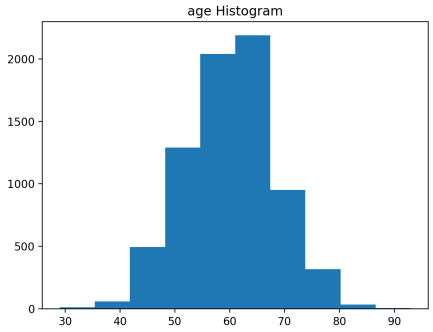

FIG. 1. Age histogram

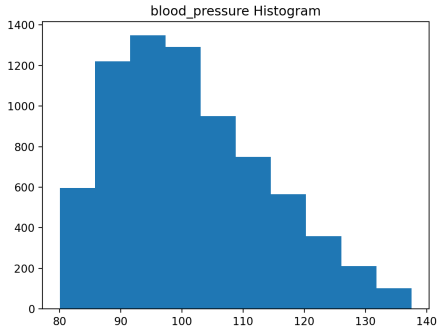

FIG. 2. Blood Pressure histogram

9. geneE
10. geneF
11. gender: (Female: 4656; Non-Female: 2844)
12. blood_test: (negative: 6863, positive: 637)

III. DATA PREPROCESSING

Pandas were used to create dataframes. Since we can note that the minimum value for blood pressure was -999, inter-quartile method was used to remove outliers. The family history column contained 2607 NaN values which is 35% of the total values, and removing those values would not be a great idea. Therefore, True and False values were added to the missing rows in a 1:9 ratio because of the original ratio of True:False. Feature Selection was done using the SelectKBest module available via sklearn. The factors that proved to be the most important were: age, blood pressure, family history, and gender. VIFs were also computed to calculate the multi-collinearity of the variables. The following were the VIFs for the numerical variables:

- treatment 2.322547
- age 19.591299
- blood_pressure 8.763100
- TestA 3.268509

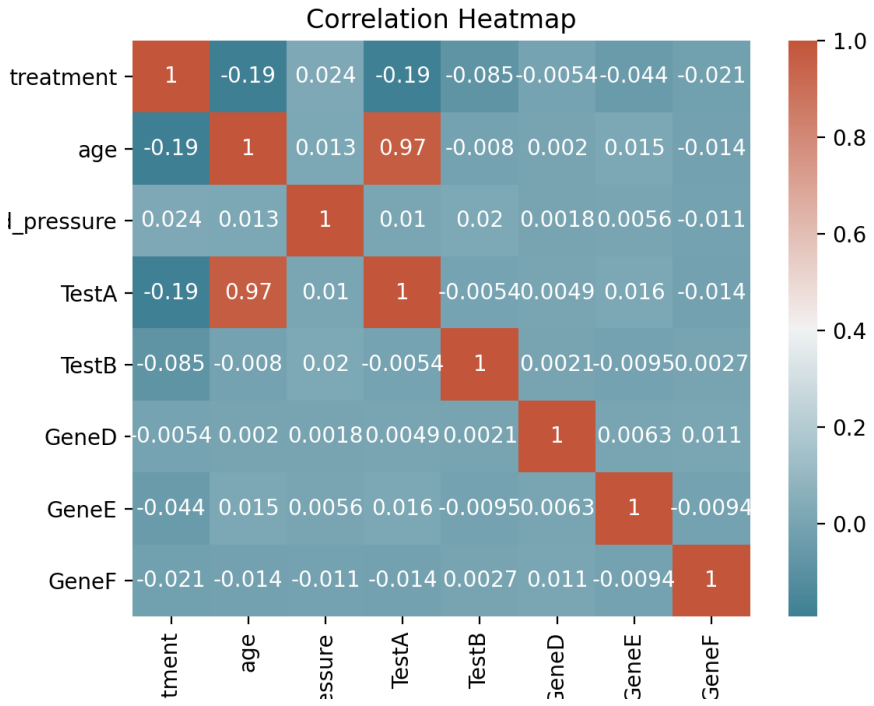

FIG. 3. Heatmap of correlations

- TestB 1.008384
- GeneD 2.214492
- GeneE 2.225922
- GeneF 2.226406

With the age VIF being high, it would not make great sense to neglect the age factor, especially considering age's significance on health, in general. The next being blood pressure, which was treated with the inter-quartile method to remove outliers.

Heatmap of different correlations has been displayed in Fig. 3

One hot encoding is an excellent tool to convert categorical variables to numerical variables. Therefore, one of pandas libraries' methods, pd.get_dummies was used to convert the same.

The data was also scaled using the skearn's StandardScaler() module.

A scatter matrix has also been displayed in Fig. 4 to show the different correlations between variables. We can visualize that the correlation between TestA and age is close to 1.

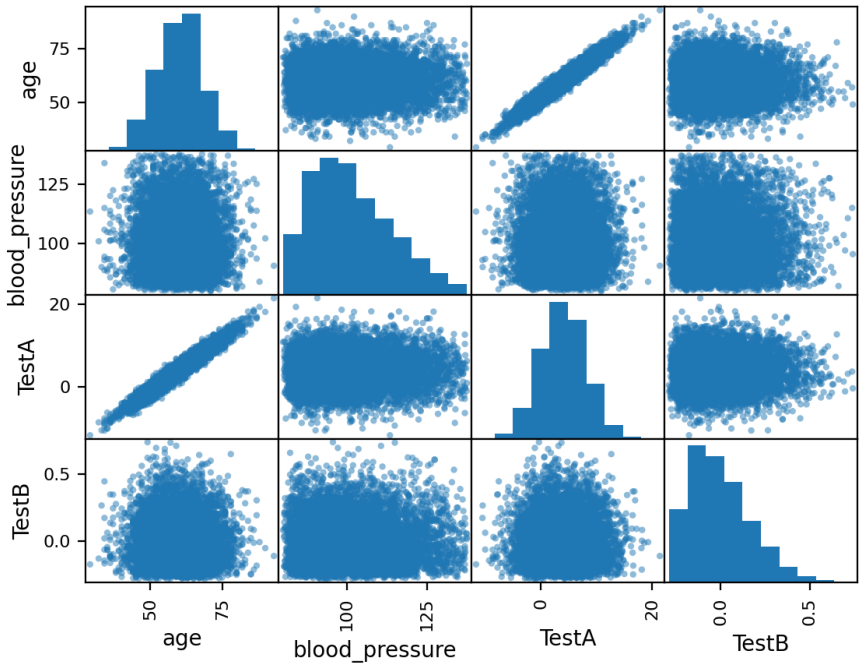

FIG. 4. Scatter Matrix

IV. MODEL SELECTION

The principal task of this assignment is to calculate the accuracy of Logistic Regression and therefore various Logistic Regression techniques were used to calculate the best models. The model with the best accuracy was sklearn's Logistic Regression() method along with LogisticRegressionCV, i.e. cross-validation. The other methods used were Pipeline module, liblinear solver, and the RepeatedKFold module. The sklearn LogisticRegression() method was selected because of its accuracy being better than all other models.

The feature importance of the variables was also calculated using .coef_. False positives and False negatives were also calculated using this method. ROC/AUC score was also calculated.

V. MODEL EVALUATION

The accuracy of the model is 73.6%. The test set is used to calculate the predicted value which is eventually used to calculate the score of the model. A threshold value of 0.5 is used to calculate the probability in the model.

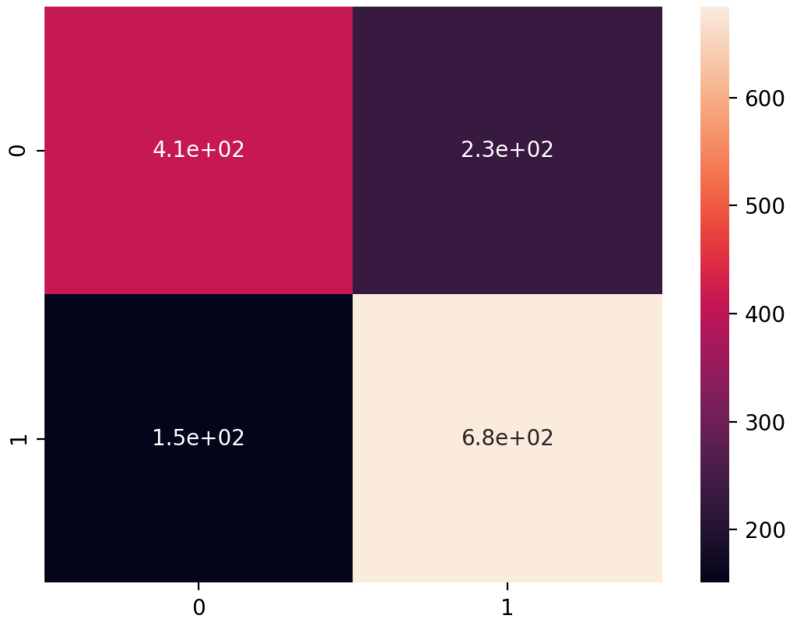

FIG. 5. Heat-map of false positives and false negatives.

Using this the false positives and false negatives are computed and a heatmap of these values is generated as seen in Fig. 5

False Positive = 235 values

False negative = 156 values.

The Cross validation model was computed and following were the accuracy values:

Training accuracy: 74.75%

Validation accuracy: 73.69%

Test accuracy: 73.56%

VI. FEATURE IMPORTANCE

Feature importance was computed using .coef_ method and following were the scores of the respective features:

1. age, Score: -0.60887

2. blood_pressure, Score: 0.29687

3. family_history, Score: 0.05631
4. TestA, Score: 0.05250
5. TestB, Score: -0.21623
6. GeneD, Score: -0.00099
7. GeneE, Score: -0.13586
8. GeneF, Score: -0.08486
9. gender_female, Score: 0.18392
10. gender_non_female, Score: -0.94622
11. blood_test_negative, Score: 0.01021
12. blood_test_positive: 11, Score: -0.07264
13. geneC_active, Score: 0.01316
14. geneC_not_active, Score: -0.00000

This can be visualized in the fig 6

This suggests that age and gender_female are two of the most important factors and the next two being blood_pressure and family_history.

VII. INTERPRETATION

The accuracy of the model is 73.6% with the Cross validation model test accuracy being 73.56%. The ROC AUC score is computed to be 78.3%, F1-score to be 77.7% and the AUC score to be 78.4%.

One of the graphs that is a great representation of False positive rate and the True positive rate is shown in Fig. 7

Another graph which illustrates the Precision vs. Recall values is displayed in Fig. 8.

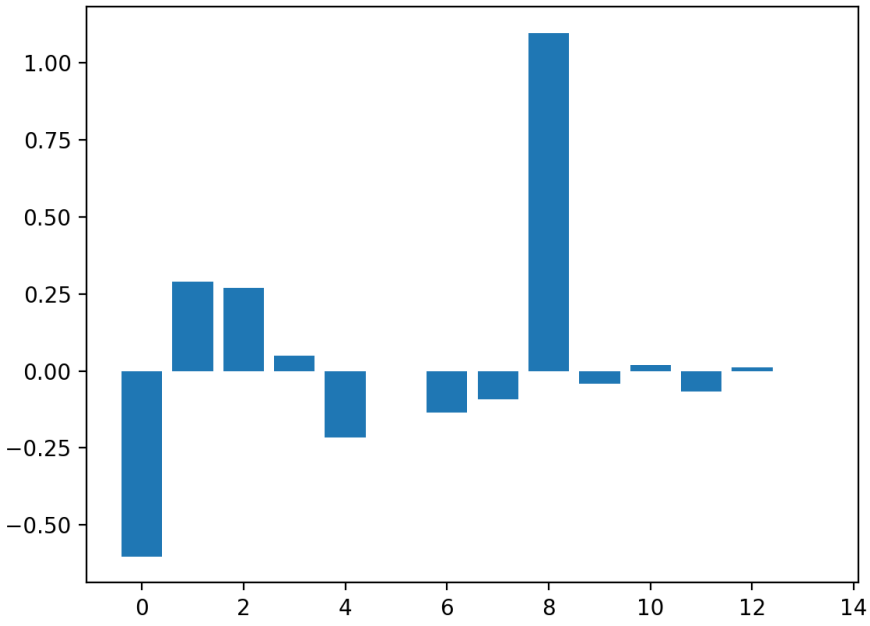

FIG. 6. Importance graph

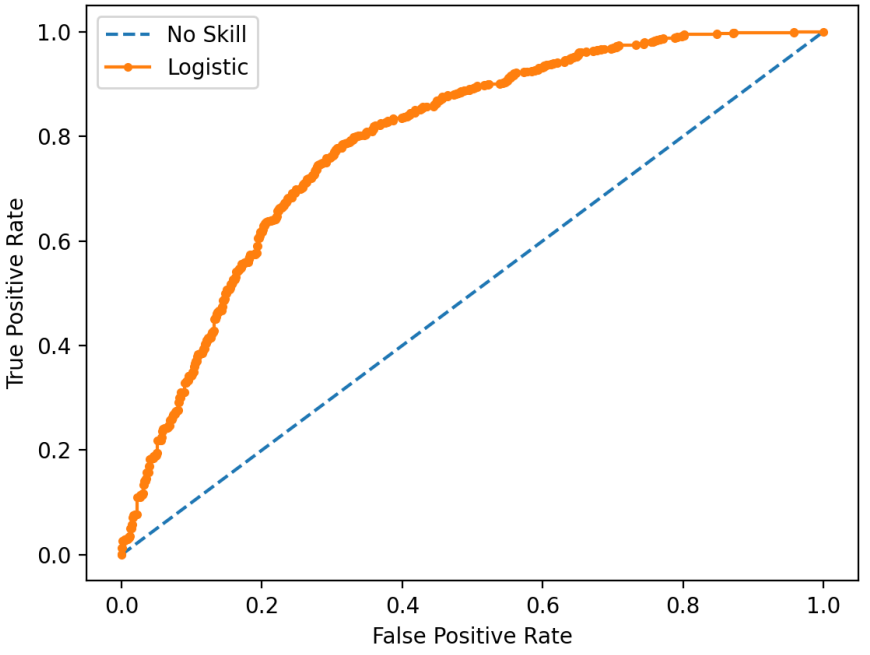

FIG. 7. Importance graph

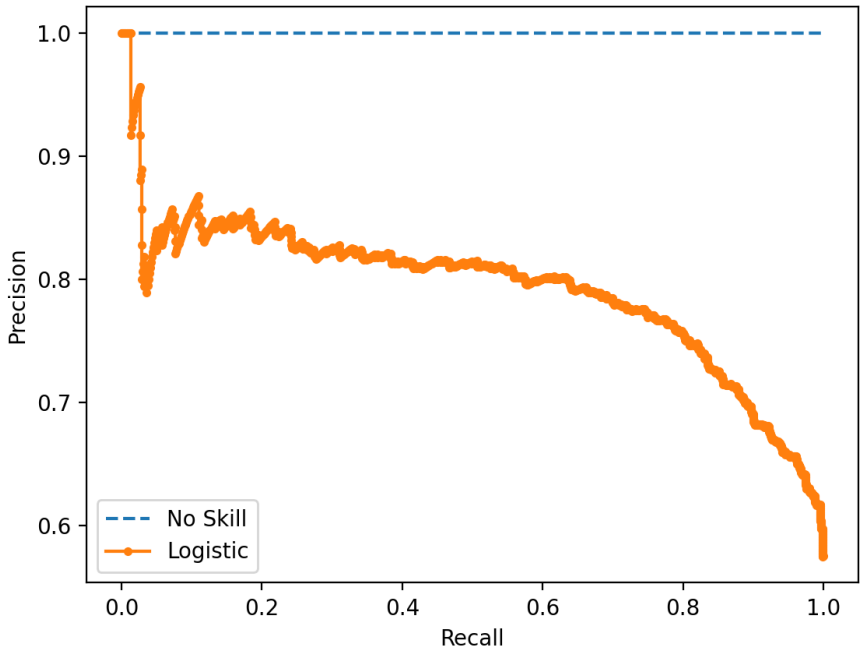

FIG. 8. Precision vs. Recall

VIII. CONCLUSIONS

It can be concluded that the the accuracy of the model is 73.6% with the Cross validation model test accuracy being 73.56%. The ROC AUC score is computed to be 78.3%, F1-score to be 77.7% and the AUC score to be 78.4%.

The most important factors can be considered as age and gender_female and the next two being blood_pressure and family_history.

There were 235 false positive values and 156 false negative values.

Age and TestA are directly correlated where if the age of the patient increases, the value of TestA increases.

DATA AVAILABILITY

Data is available at blackboard names "ps2_public.csv"

CODE AVAILABILITY

Code is available at `https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps2-sjain681`

- feature.py: Feature selection
- hw2.py: Most-effective Logistic regression models with plots
- lmisc_models.py: Miscellaneous with not so efficent models
- visualizations.py: Visualizations

ACKNOWLEDGMENTS

I would like to thank professor Kristina Lerman, professor Keith Burghardt, the TAs and the graders of the course DSCI 552 for all their efforts towards this assignment.

Appendix A: References

1. Jeff Macaluso. "Testing Logistic Regression Assumptions in Python." Jeff Macaluso, 27 May 2018, jeffmacaluso.github.io/post/LogisticRegressionAssumptions/.
2. "Sklearn.logistic_model.LogisitcRegression." Scikit, scikit-learn.org/stable/modules/generated/sklearn.logistic_model.LogisticRegression.html.