DSCI552-Assignment2

Dan Shen

Viterbi School of Engineering, University of Southern California,
Los Angeles, California 90089, USA

(Dated: February 18, 2021)

Abstract

Machine learning is simply making healthcare smarter. This powerful subset of artificial intelligence may be familiar to many in use cases such as speech recognition used by voice assistants, and in creating personalized online shopping experiences through its ability to learn associations. However, machine learning has demonstrated truly life-impacting potential in healthcare – particularly in the area of medical diagnosis. This report shows application of Logistic regression in deciding whether a certain treatment is recommended for the patient or not, including data exploration, data processing, performance evaluation and feature importance testing.

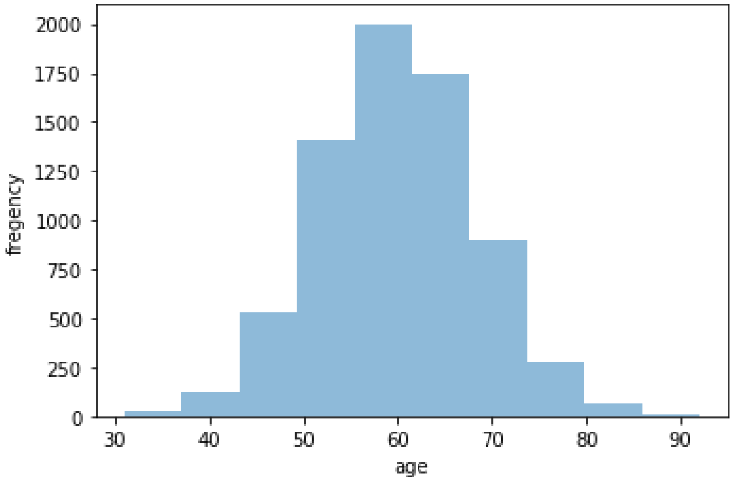
I. INTRODUCTION

The data set is about medical records of patients with 10000 rows and 10 features.

Firstly, I did Data exploration(visualization) with matplotlib and Seaborn. Secondly, I did Data processing, including handling null values, removing outliers and doing one-hot encoding. Thirdly, since it is a classification problem. I choose to use Logistic regression model, and evaluate the performance using roc-auc score. Lastly, after I evaluated the importance of features, I found that the additional 5 features are not that important.

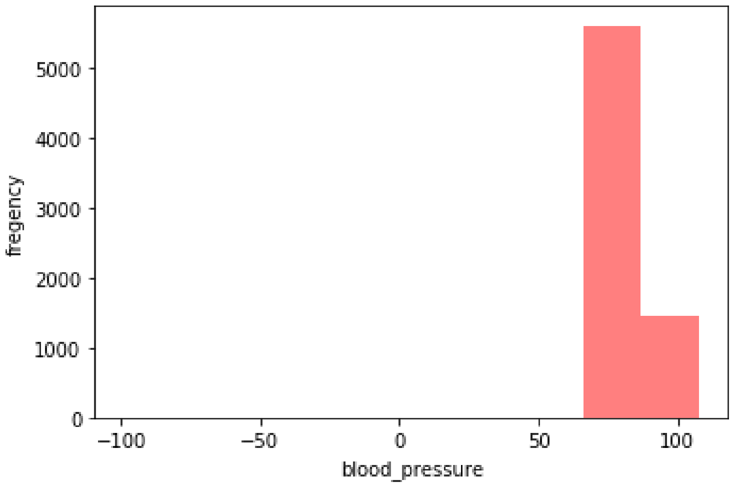
II. DATA EXPLORATION

Before delving more into the features, let us first have a look at the target variable 'age'.

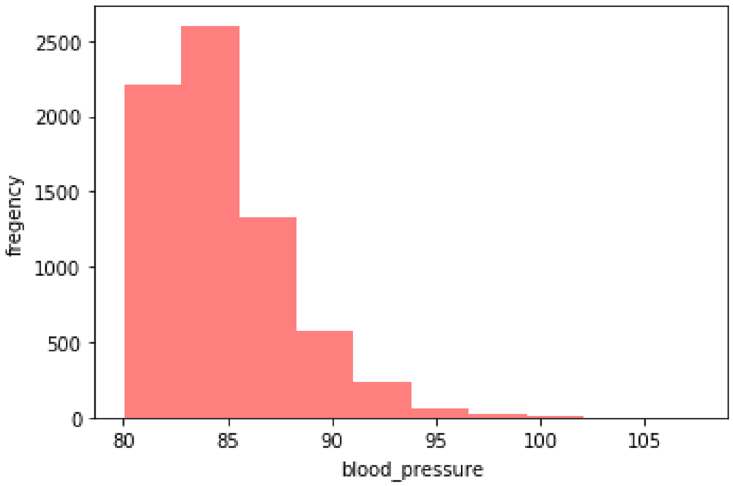


As shown in figure, the age is distributed as normal distribution. The mean value is around 58.

Now let us look at the blood distribution:

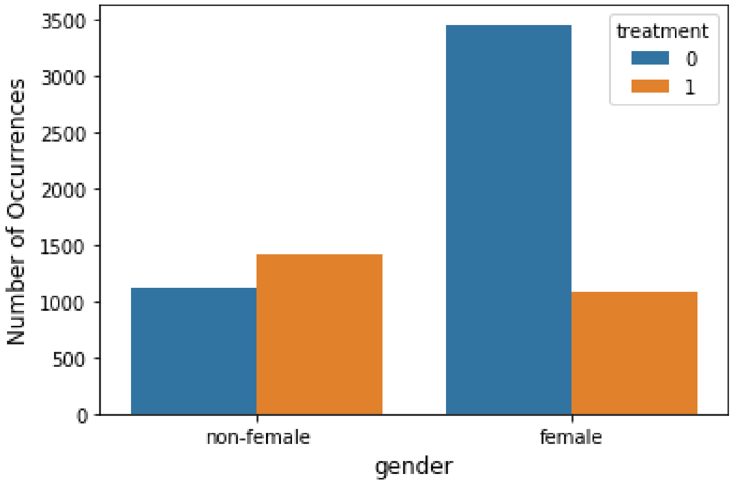


It seems that there are some ouliers.



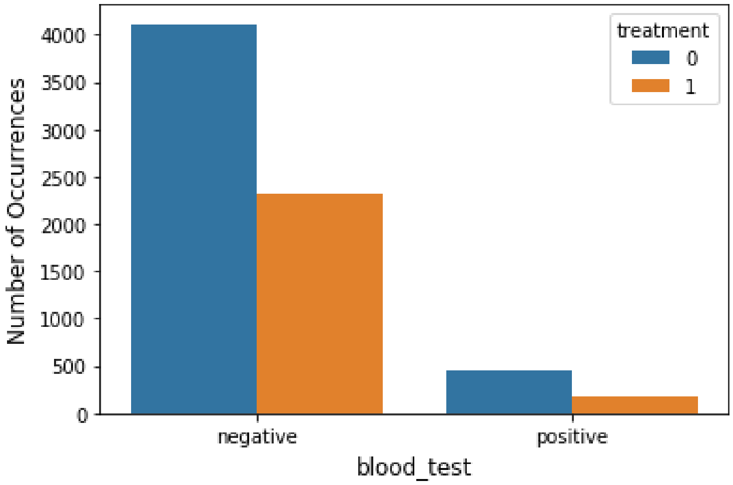
What the figure is like after removing the ouliers.

Now let us look at the categorical variable gender.



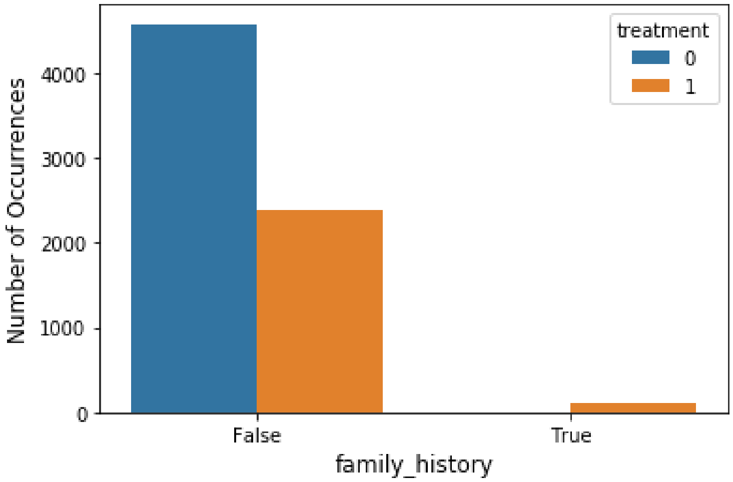
From the figure we may guess that female patients tend not to be suitable for this treatment

Now let us look at the distribution of blood test.



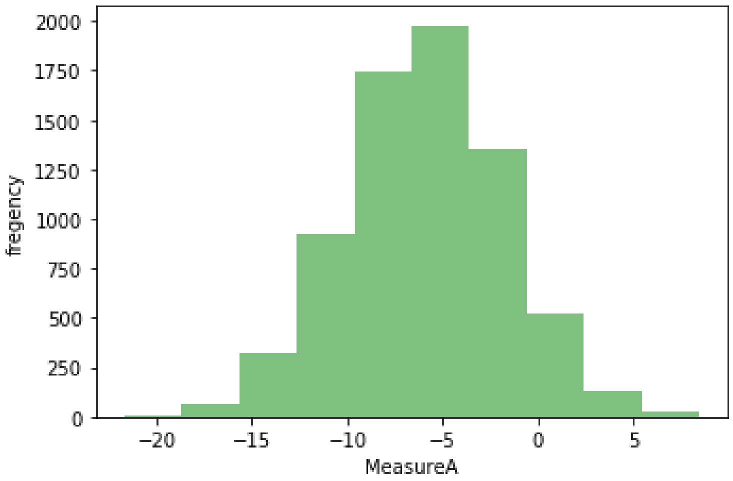
Blood test may not be a critical factor, as shown in the figure

Now let us look at the influence of family history.



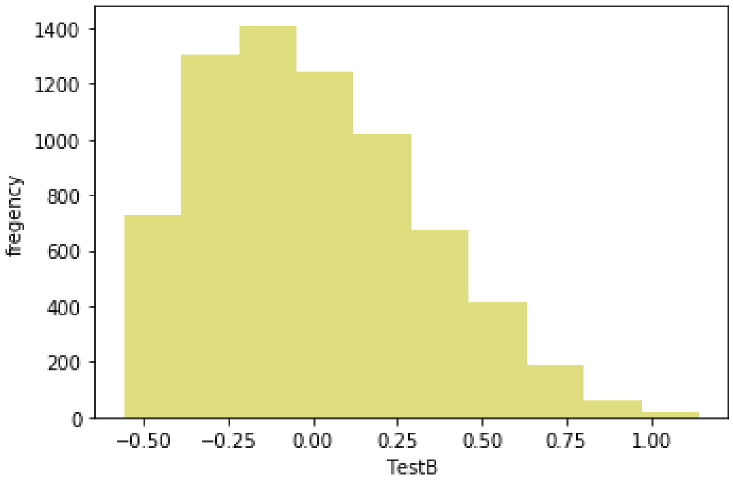
It seems that family history matters a lot.

Now let us look at the Measure A distribution.



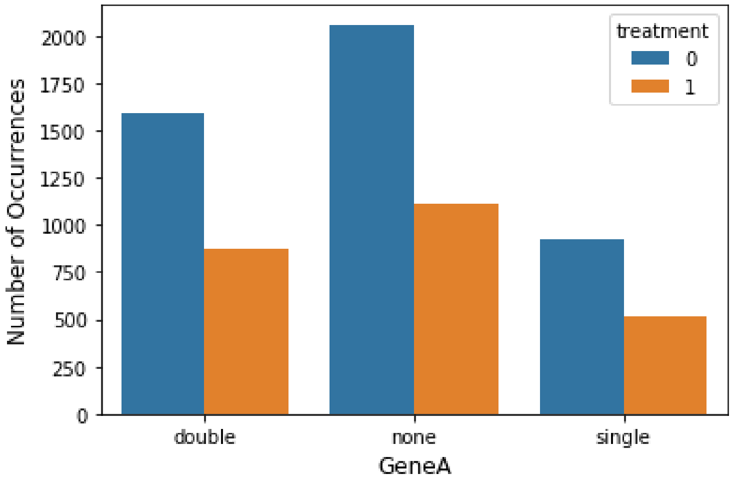
We do not exactly know what Measure A is, the mean value is around -5

Now let us look at the distribution of Test B.



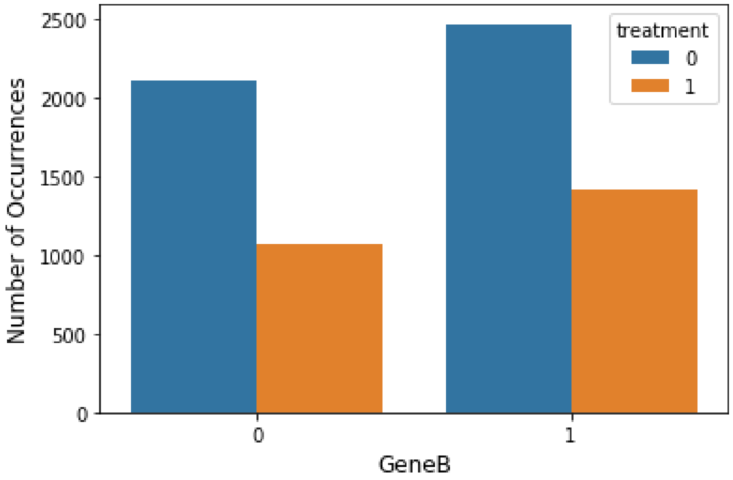
The Test B is distributed as right skewed distribution.

Now let us look at the Gene A.



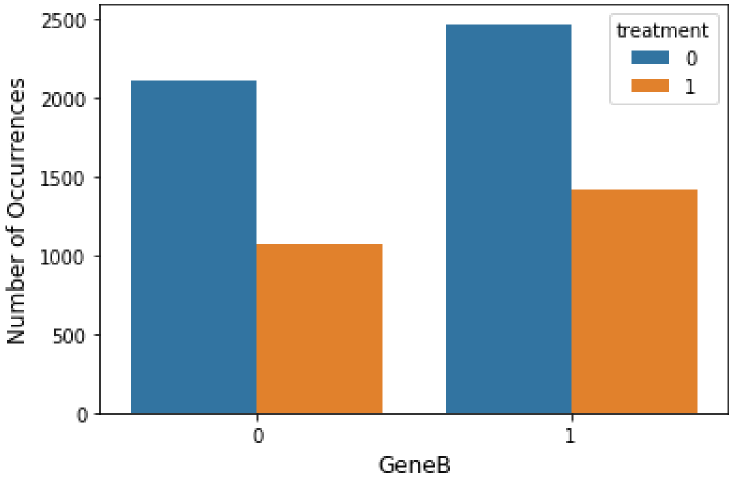
We do not know the meaning of Gene A, it may not be a critical feature.

Now let us look at Gene B.



Gene B may not matters a lot as the figure shows.

Lastly let us look at Gene C.



Same as Gene A and Gene B, Gene C may not be a critical factor.

III. DATA PREPROCESSING

Null value handling. By use the .info() function, I noticed that there are 2932 null value in column 'family$_{h}istory'.Ichoosetoremovethemsincethereisnosuchaperfectwaytofillthenull.$

Outliers removing. I noticed that there are 3 outliers in column blood pressure, which are all negative values. I removed these 3 rows so they will not affect the model fitting.

Feature scaling. For categorical features like gender and blood test, I choose to use one-hot encoding to handle them. Also, I transfered numerical variables like age to categorical by cutting them into several parts, and then one-hot encode them too, since I found that it would improve the performance.

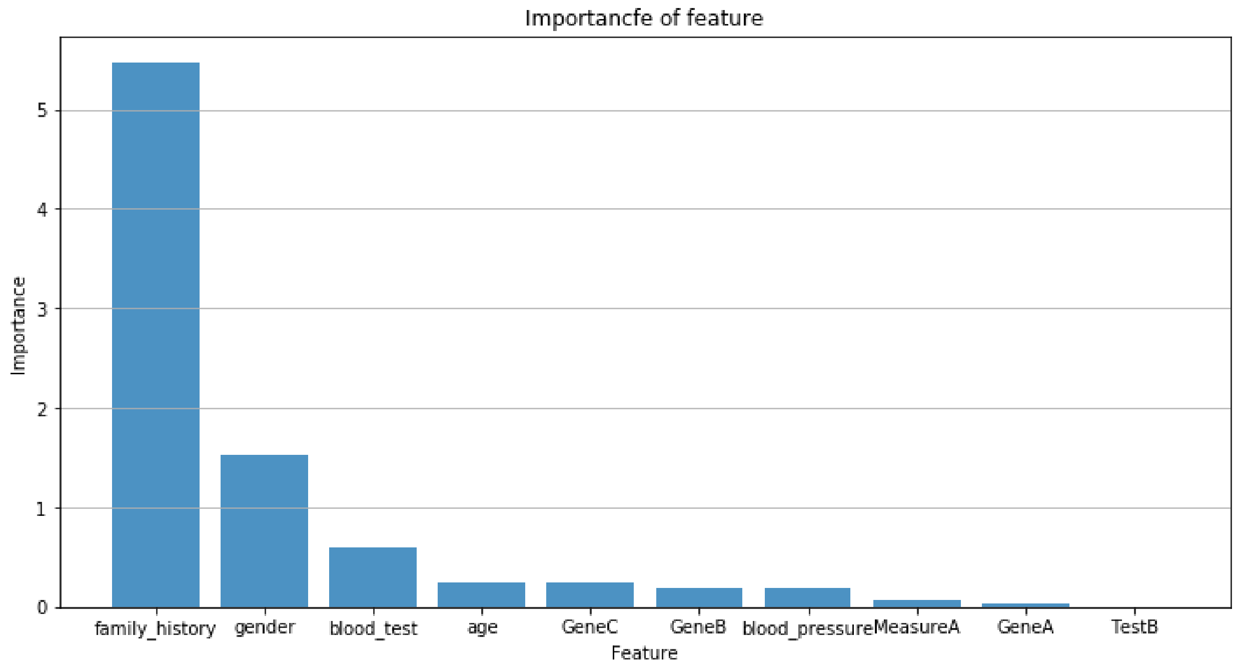
IV. MODEL SELECTION

It is a binary classification problem, so I chose to apply Logistic regression. Also, I used GridSearchCV trying to find the best estimator. In Logistic regression, the hyper parameter is 'C'.

V. MODEL EVALUATION

I used AUC score to evaluate the performance. And I got 0.816 for my Logistic regression model.

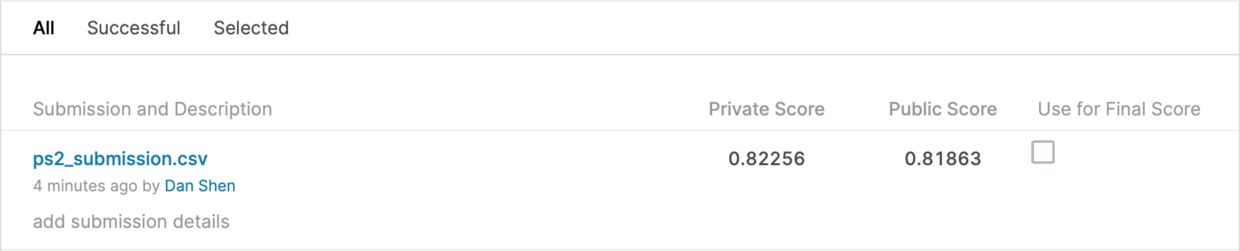
VI. FEATURE IMPORTANCE



As shown in the figure, family history is the most critical feature. And the additional 5 features (GeneC, GeneB, GeneA, MeasureA and TestB) are not so important.

VII. CONCLUSIONS

Family history and gender are top two factors to predict whether a patient could undergo this treatment or not. The additional 5 features are not very useful. So it is not that worthy accessing them considering they are really expensive and difficult to get. The model I designed got 0.816 for AUC score. And on Kaggle it gets 0.823 for Private Score, 0.819 for public score.



Submission and Description	Private Score	Public Score	Use for Final Score
ps2_submission.csv 4 minutes ago by Dan Shen add submission details	0.82256	0.81863	□

CODE AVAILABILITY

Code is available at https://github.com/usc-dsci552-32415D-spring2021/problem-set-02-TuffySd And I also uploaded the ipynb file on blackboard.