Homework 2 Report

Jae Young Kim

Applied Data Science, University of Southern California,

Los Angeles, California 90089, USA

(Dated: February 25, 2021)

Abstract

Through this report, ps2_public data set was explored and used to train models. Through exploration, important features were found. Gender, age and TestB were the best features. From the additional features we can use, TestB and GeneE were recommended to collect. To fit an easily interpretable model, logistic regression model is chosen. Among many regression models, L1 regularized logistic regression model scored the best score. It scored Accuracy 0.81, Precision 0.82, F1 0.82 score and ROC AUC score 0.89.

I. INTRODUCTION

To classify the treatment of the patients, I explored ps2_public data set and trained a prediction model. While exploring data, I tried to find interesting trend of the data from the stacked bar plots and histograms. I used stacked bar plot to see the distribution of categorical features such as gender, Family history, blood test, GeneC, GeneD, GeneE and GeneF. To prove it with statistics, T test is implemented and statistics were interpreted. For numerical variables such as age, blood pressure, TestA and TestB, I drew histogram and kernel density plot to check whether there is obvious difference. Through the exploration, I found that age, blood pressure, TestA and TestB all showed low p value result which means they are all informative at predicting the treatment class. For blood pressure, as I found -999 values and those values were replace with mean value of training set. When I checked the correlation matrix, TestA had high correlation with age which means one of TestA and age can be discarded. As it expensive to collect TestA, we don't have to collect TestA. For TestB, as the distribution showed symmetry by zero, TestB was squared.

For the model selection, as easily interpretable models were recommended, I tried logistic regression models. I tried two versions, simple logistic regression model and L1 regularized logistic regression model. To do hyperparameter search, grid search was implemented. After the hyperparameters were chosen, final score was calculated after training the models with train set and validation set and tested it with the test set.

II. DATA PREPROCESSING

To preprocess data, first and foremost, the rows with null values for each column were counted. As a result, family history column had 2607 null values. As it is about 35% of our whole data, I decided not to drop the rows with null values. I considered nan values in family history can be considered as a category for people who were not tested. Therefore, dummy variable that maps family story false to -1, true to 1 and nan to 0 was created.

Then categorical features were preprocessed. As all the categorical features(gender, blood test, GeneC) had only two different values, they are changed into numerical variable with 0 and 1.

Then outlier detection was done. Box plots were plotted to check it. In blood pressure

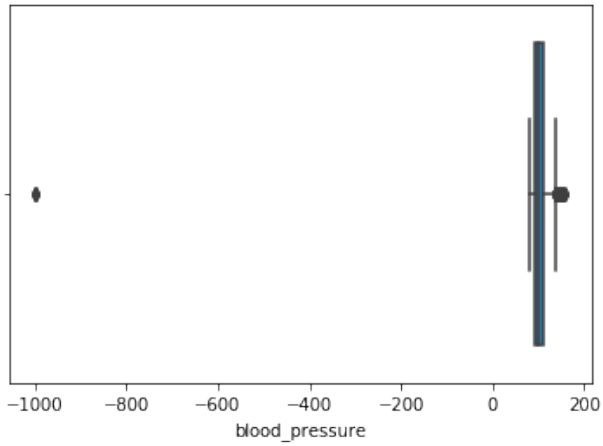

FIG. 1. Box Plot of Blood Pressure

box plot, outliers were detected.

The outliers in Fig.1 had blood pressure value -999. As -999 is often considered as an alternative of null value, those values were interpolated to mean value of blood pressure of the training set, 102.05.

III. DATA EXPLORATION

Data exploration is implemented to catch intuition from the data itself. As our task is classification task of 0, 1, two separate histograms and kernel density estimation of treatment 0 and 1 are drawn for continuous variables and two separate stacked bar plots were drawn for categorical variables. Then T test is done for each feature to check how useful they are.

First of all, Gender seems to be informative for determining treatment. From Fig.2, it is obvious that the proportion is different. Female had much higher proportion of treatment 1 and male had higher proportion of treatment 0. Statistically, the pvalue of t test between two groups of treatment 0 and treatment1 is 0. Which means it is reasonable to reject the null hypothesis H0 : treatment 0 and 1 has no difference in the proportion of gender. In other words, gender is an important factor.

In case of TestA, the histogram and kernel density estimation(KDE) in Fig.3 showed that there is difference for two groups, treatment 0 and 1.

As the p value of T test was 3.37*e-59 and it is very small, it is reasonable to reject the null hypothesis H0 which means there is no difference in the mean value of two groups treatment 0 and 1. In other words, there is significant difference in the two groups, treatment

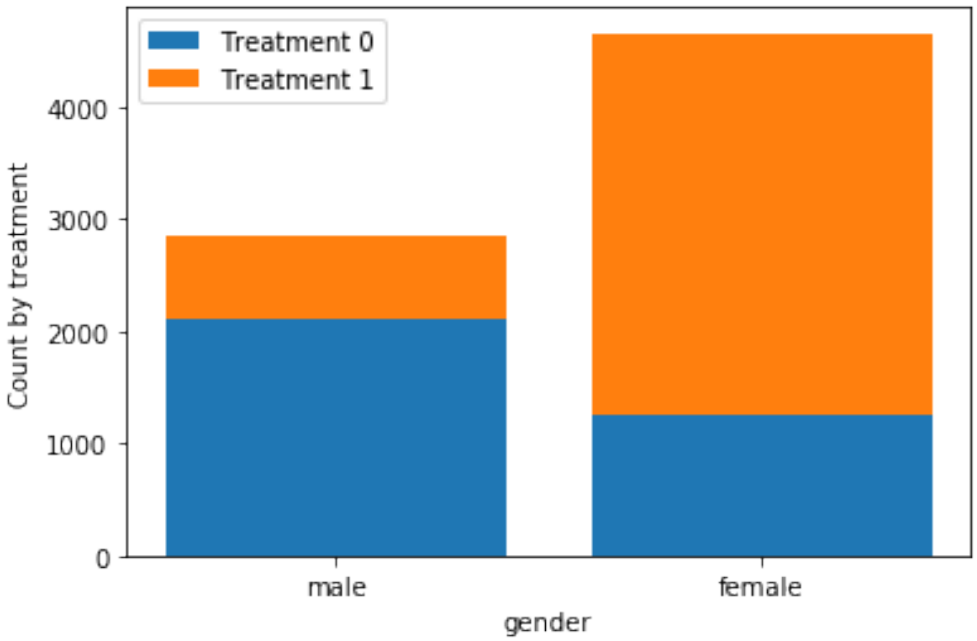

FIG. 2. Stacked bar plot of gender versus treatment

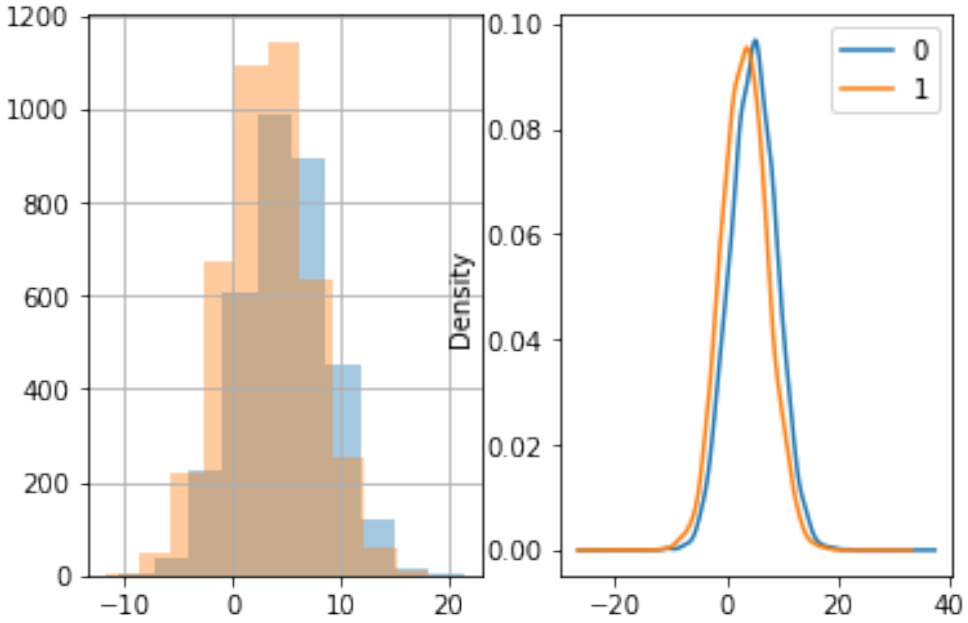

FIG. 3. Histogram and KDE of TestA grouped by treatment

0 and 1 in terms of TestA. However, when you see the correlation matrix of the features in Fig.4, TestA has significantly high positive correlation with age. In other words, if both age and TestA are used as input feature in regression model, there will be multi-colinearity problem and have less additional information from TestA. Therefore, it is decided not to

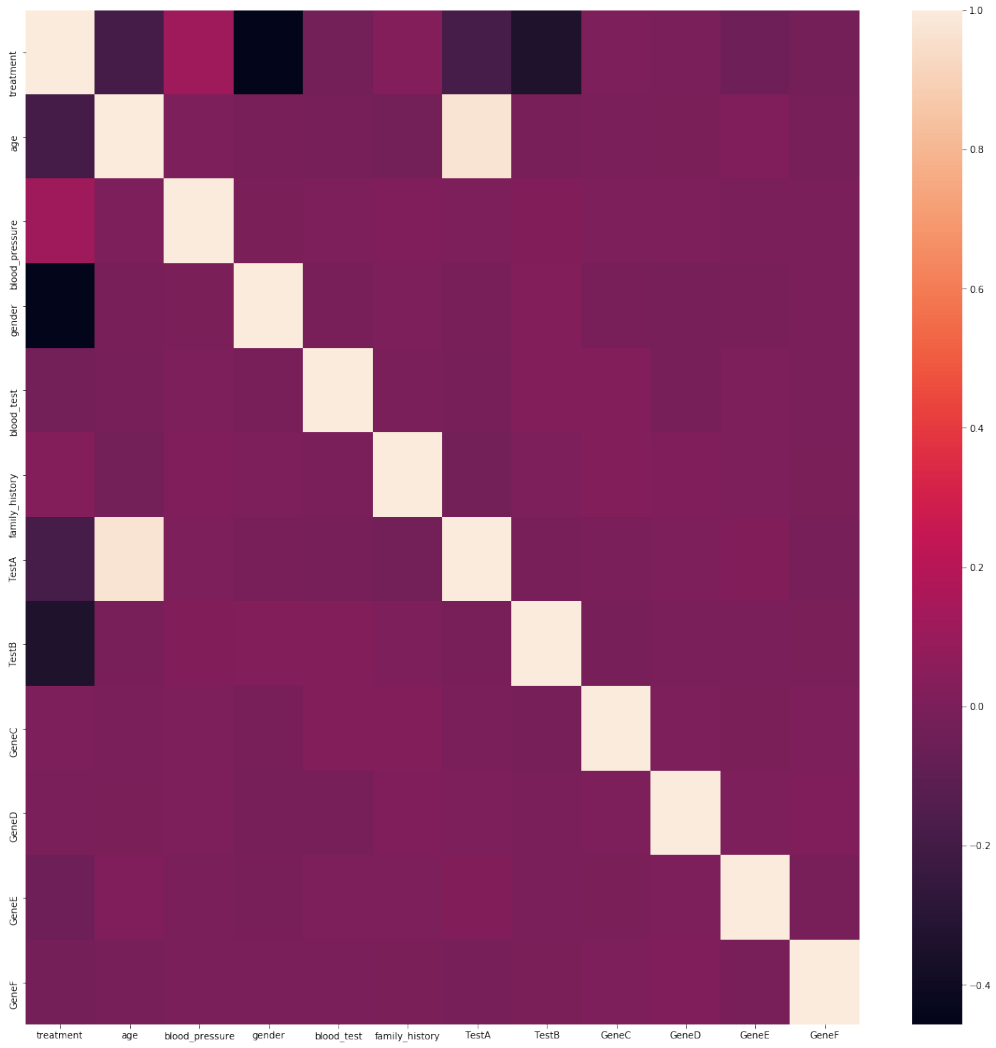

FIG. 4. Heatmap of Correlation Matrix

include TestA in the final model.

For TestB, it is found that there is obvious difference of distribution for treatment 0 and 1. When TestB value is near 0, there is high probability of treatment 1 but not for treatment 0. Through this, I thought the distance from TestB to 0 has important meaning. Therefore, I decided to transform TestB by squaring the value. After squaring TestB, the distribution changed as it is shown in Fig.6.

Before squaring TestB feature values, the pvalue of T test between treatment 0 and 1 was 2.56*e-12. However after squaring it, the pvalue became 2.05*e-178. The pvalue was sufficiently small before squaring. However, after squaring the p value decreased significantly. In addition, as we are going to use logistic regression model and it is a kind of linear model, this distribution change is effective. Trivially, as p value is very small, TestB is an important

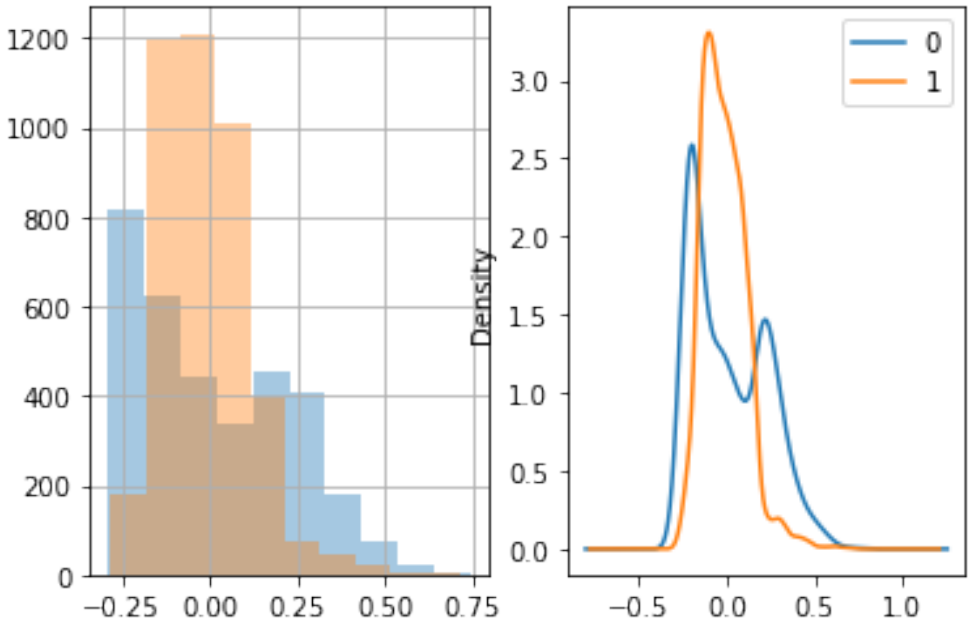

FIG. 5. Histogram and KDE of TestB grouped by treatment

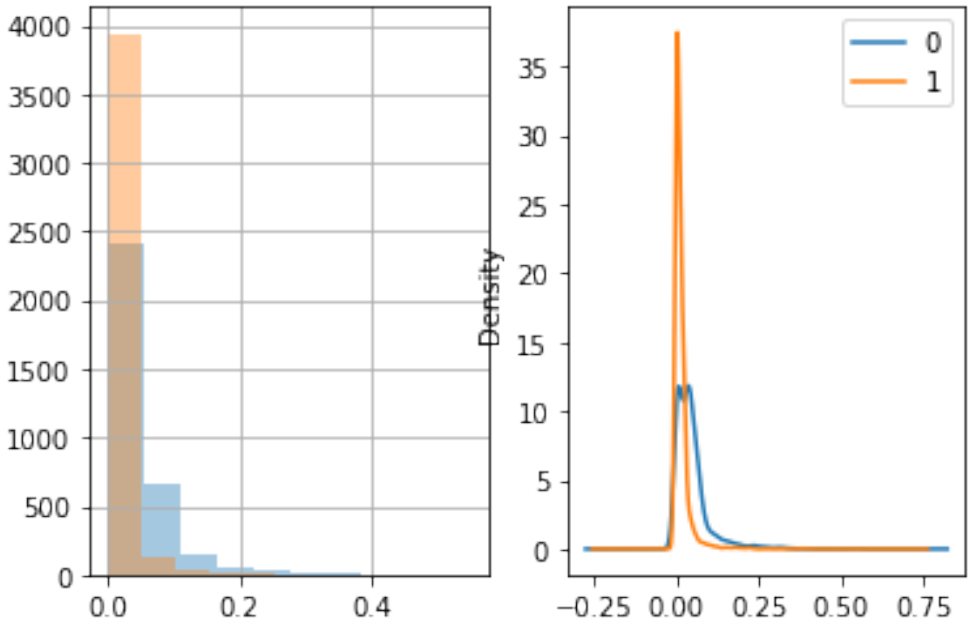

FIG. 6. Histogram and KDE of TestB after squaring and grouped by treatment

feature in determining treatment.

For GeneC, GeneD, GeneF, the pvalue of T test between treatment 0 and 1 are 0.45, 0.64 and 0.07 for each. They are large enough not to reject the null hypothesis H0 : The proportion of treatment 0,1 are not different when GeneC, GeneD, GeneF are different.

Therefore, GeneC, GeneD, GeneF are not informative in determining treatment.

For GeneE, the pvalue of T test between treatment 0 and 1 is 0.0001. Therefore, it is reasonable to reject the null hypothesis H0 : The proportion of treatment 0,1 is not different when GeneE is different. Furthermore, it is an informative feature.

Therefore, if it is possible to collect two features, I will select TestB and GeneE. But if I have to choose one feature, I will choose TestB.

IV. MODEL SELECTION AND EVALUATION

There are many models which can be applied to this data set. However, I tried logistic regression models to make the model simple and easily interpretable. I used simple logistic regression and l1 regularized version of logistic regression. To measure the performance of the models, accuracy, precision and F1-score are used.

To get the best hyperparameter of l1 regularization version of logistic regression, many models with different hyperparameters were tried. I brutally searched best parameter of L1 regularization from 0 to 100. 100 models with parameter from 0 to 100 with interval 1 were tried and then tried with interval 0.01 to find the best parameter in two decimals. Finally, hyperparameters of L1 regularization were 9.32 and 1.96 for model without feature drop and feature dropped model

Model	Simple LR without feature drop	L1 LR without feature drop
Accuracy	0.7267	0.8067
Precision	0.7633	0.8171
F1 Score	0.7871	0.8221
ROC AUC Score	0.8610	0.8932
False Positive	102	75
False Negative	76	70

Model	Simple LR with Feature selected	L1 LR with Feature selected
Accuracy	0.768	0.8067
Precision	0.7643	0.8171
F1 Score	0.7933	0.8221
ROC AUC Score	0.8607	0.8927
False Positive	103	75
False Negative	71	70

The result showed that L1 regularization have significant increase in the overall performance. However, feature selection did not show obvious performance increase. The best model was L1 regularized logistic regression model with accuracy 0.8067, precision 0.8171, f1 score 0.8221, ROC AUC score 0.8928 which showed 75 False Positive and 70 False Negative from 750 test data.

V. FEATURE IMPORTANCE AND INTERPRETATION

P value of T test with treatment for each feature can be a measure of feature importance. Fig.7 shows that gender is the most important feature. TestB(squared value of original TestB value), age, TestA follows next. However, as TestA had high correlation with age, it was dropped. GeneD, GeneC and GeneF showed relatively less importance.

When I measure feature importance with the coefficients of logistic regression model after normalizing feature values with min max scaler, the overall trend was similar. However, in this case, TestB showed the highest feature importance. Then age, gender followed. The coefficients are shown in Fig.8

VI. CONCLUSIONS

To sum up, gender, age and TestB were the features having the biggest impact on determining treatment of the patients. There was obvious trend that the women have been classified into treatment 1 more. In case of age, elder people were classified into treatment 9 and when the squared value of TestB is small, there is more possibility of treatment 1. From the 6 additional features that we can use, TestB, GeneE were informative. Therefore,

	feature	p_val
8	GeneD	6.410407e-01
7	GeneC	4.465529e-01
10	GeneF	6.728880e-02
3	blood_test	1.676388e-02
4	Family_History	8.474969e-03
9	GeneE	1.274084e-04
1	blood_pressure	8.596063e-23
5	TestA	3.371280e-59
0	age	1.856751e-63
6	TestB	2.049792e-178
2	gender	0.000000e+00

FIG. 7. Ranked P value of T test with treatment for each feature

if it is possible to collect two additional features, TestB and GeneE should be selected. But if it is necessary to select only one feature, TestB should be chosen.

Logistic regression with L1 regularization is selected as a final model. It scored the best for Accuracy(0.81), Precision(0.82), F1(0.82) score and ROC AUC score(0.89).

DATA AVAILABILITY

Data is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps2-jeayoung114/data

CODE AVAILABILITY

Code is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps2-jeayoung114/code/hw2

	Parameters	ABS_Parameters
TestB	-15.665257	15.665257
age	-3.438466	3.438466
gender	-2.528467	2.528467
blood_pressure	2.115507	2.115507
TestA	-1.374962	1.374962
family_history	0.345905	0.345905
blood_test	-0.326775	0.326775
GeneE	-0.272200	0.272200
GeneF	-0.176328	0.176328
GeneD	-0.043306	0.043306
GeneC	-0.012835	0.012835

FIG. 8. Ranked coefficient of Logistic regression