Problem Set 2

Ran Tao

(Dated: February 18, 2021)

Abstract

I was asked to design a model, that can classify, if a certain treatment is recommended for the patient or not. I use Logistic Regression model with L1 regularization. This model is useful.

I. INTRODUCTION

My task from CEO: I was asked to design a model, that can classify, if a certain treatment is recommended for the patient or not. Then I was asked to train the model and access how useful this model is. Additionally, I must give recommendations which features are important to collect.

My task from Technical Manager: I was asked to evaluate my model by reporting accuracy and precision. I was also asked to report false positive, false negative and negative, and the AUC score. In addition, I was asked to add regularization to my logistic regression and show how the model can be interpreted by figuring out the most important relations between the variables and the expected outcome.

My task from senior developer: I was asked to give a very explained code (meaning that I should write comments and try to explain any non-trivial section).

II. DATA EXPLORATION

Firstly, I check the variable type, and find there are four categorical variable, namely, gender, blood test, family history, and GeneA. Beside, I also find family history has around 30% missing values.

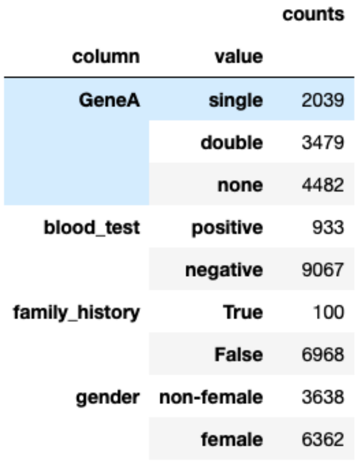
Secondly, I count frequency of each categorical data, see Fig.1. In this step, I first find blood test and family history are imbalanced. Then I have some feature engineering ideas: a) blood_test, family_history, and gender: create dummy variables b) GeneA: use one-hot encoding.

Thirdly, I compute the coefficient matrix of each features and find the relationship between age and Measure A is -0.97. This shows this two features are strongly correlated with each other. I decide to use regularization to deal with it.

III. DATA PREPROCESSING

Data cleaning:

When exploring the data set, I find family history has around 30% missing values. So my first step of data preprocessing is to deal with those missing values.



column	value	counts
GeneA	single	2039
	double	3479
	none	4482
blood_test	positive	933
	negative	9067
family_history	True	100
	False	6968
gender	non-female	3638
	female	6362

FIG. 1. Frequency of each categorical data

Firstly, I try to find whether these null values are Missing Completely at Random (MCAR). If so, we can just delete them. In this step, I first find whether missing values are correlated with any object feature. I count values of each object from row with and without null-values. It turns out the proportion of values of each object is basically the same. This shows that missing values are not correlated with any object feature. Then I try to compute the mean and variance of each numerical features and find the numbers are roughly the same for rows with and without missing values. This shows the missing values are also not correlated with numerical features. By far, we can know that missing values are Missing Completely at Random (MCAR). Therefore, I choose to simply delete those lines that have missing values.

Feature selection:

In the data exploring step, I find the feature: family history is extremely imbalanced. It has 6968 negative values but only 100 positive values. So I decide to check whether family history affect treatment. The result is shown in Fig.2.

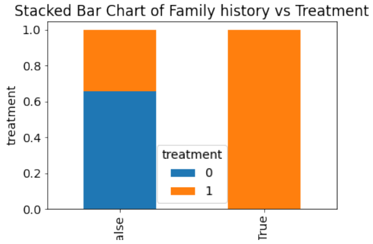


FIG. 2. Stacked Bar Chart of Family history vs Treatment

In this picture, when can know that the only 100 True value is all 1 in treament value. I think this will cause great bias. So I decide to delete this feature.

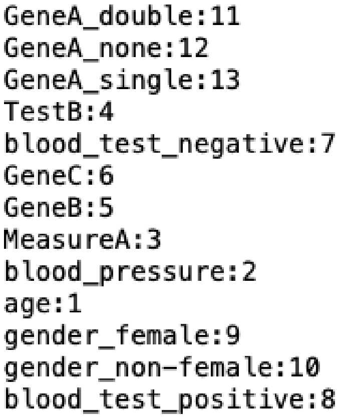
Feature engineering:

When exploring the data set, I decide to create dummy variables to deal with family history, blood test and gender, and use one-hot encoding to deal with GeneA.

IV. MODEL SELECTION AND FEATURE IMPORTANCE

Feature Importance:

I use RFE method from sklearn.feature_selection to measure the feature importance. I firstly create an estimator using LogisticRegression method from sklearn.linear_model. Then I create a selector by using RFE method. I set n_features_to_select to 1 and step to 1. Then I use this selector to fit my data. Then I show the ranking of feature importance. The result is shown in Fig.3.



```
GeneA_double:11
GeneA_none:12
GeneA_single:13
TestB:4
blood_test_negative:7
GeneC:6
GeneB:5
MeasureA:3
blood_pressure:2
age:1
gender_female:9
gender_non-female:10
blood_test_positive:8
```

FIG. 3. The ranking of feature importance

Model Selection:

In this step, I first split my data into training, validation, and test set. Then I fit my model using Logistic Regression. Then I fit other two models by using L1 and L2 regularization. Then I evaluate the accuracy, confusion matrix, precision score, recall score and f1 score of each three models. It turns out these scores of these three models are basically the same. However, the L1 regularization turns the coefficient of many useless features to zero or nearly zero. It significantly increase the model interpretability. Therefore, I choose to use the L1 regularization model as my final model.

V. MODEL EVALUATION

I firstly combine the training set and the validation set as my new training set. Then I use Logistic Regression Model with L1 penalty to fit our data. Then I calculate the accuracy of my model. It's 0.71. Then I compute the confusion matrix. The True negative is 790. The False Positive is 131. The False Negative is 273. The True Positive is 220. Then I compute the precision score, recall score, and f1 score. They are 0.63, 0.45, 0.53. Lastly, I draw ROC curve and compute the AUC score. It's 0.73. The ROC curve and AUC score is shown in Fig.4.

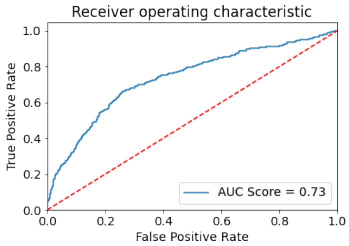
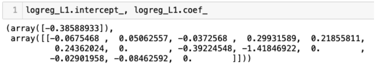


FIG. 4. ROC Curve and AUC score

VI. INTERPRETATION

The intercept and coefficient of my model is shown in Fig.5.



```
logreg_L1.intercept_, logreg_L1.coef_
```

```
(array([-0.38588933]),
 array([[-0.0675468 ,  0.05062557, -0.0372568 ,  0.29931589,  0.21855811,
          0.24362024,  0.        , -0.39224548, -1.41846922,  0.        ,
         -0.02901958, -0.08462592,  0.        ]]))
```

FIG. 5. Intercept and coefficient of my model.

The most important relations between the variables and the expected outcome is age. The coefficient is -0.07. That means with each passing year, the logit(log(p/1-p)) decreases 0.07, the amount of change of probability that the treatment is advised will depend on the value of age.

VII. CONCLUSIONS

The accuracy of my model is 0.71. The True negative is 790. The False Positive is 131. The False Negative is 273. The True Positive is 220. The precision score, recall score, and f1 score are 0.63, 0.45, 0.53. The AUC score is 0.73.

Measure A is important, but it may not worth to collect it because it's strongly correlated with age. Test B and GeneB are important and are worth to collect. GeneA and GeneC are not important.

DATA AVAILABILITY

Data is available at https://github.com/usc-dsci552-32415D-spring2021/problem-set-02-rantao-usc/blob/main/ps2_available_dataset.csv

CODE AVAILABILITY

Code is available at https://github.com/usc-dsci552-32415D-spring2021/problem-set-02-rantao-usc/blob/main/Untitled.ipynb