Used Car Price Prediction Model Report

Katie Chak
M.S. Communication Data Science
DSCI-552 Spring 2021

INTRODUCTION

This report will explore the medical data provided by a hospital client to evaluate how tests and other patient information can help us predict whether or not the patient needs treatment. This report uses logistic regression to classify patients that need and do not need treatments. The report will go through the process of data exploration, data cleaning, model selection using grid search, feature importance evaluation, and lastly, interpretation of results and conclusion.

I. DATA EXPLORATION

An initial overview of the data using .describe() shows that some patients' blood pressure value is -999. This probably indicates missing values. In order to keep the data clean, we will remove these values during data processing. Other blood pressure, no other missing data is found. There are a total of 7493 data points (rows) after taking out the null blood pressures.

With some visualization tools, it was insightful to see how each independent variable is distributed in different classes of treatment (0 and 1). Notably, there are significantly more females than non-females to get treatment (Figure 1). Moreover, patients who were recommended treatment have slightly higher blood pressure (Figure 2). Lastly, patients who were recommended treatment are also slightly younger.

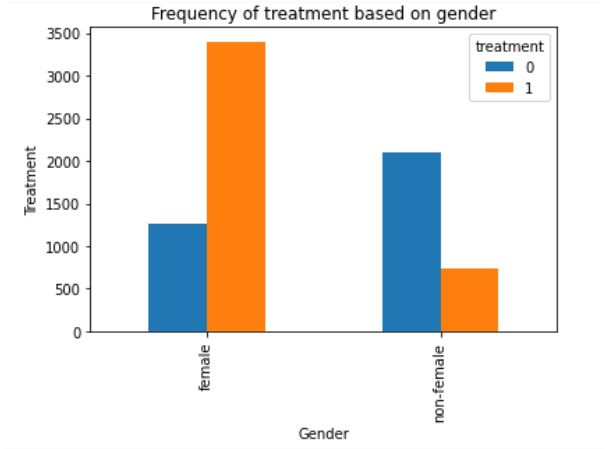

Figure 1

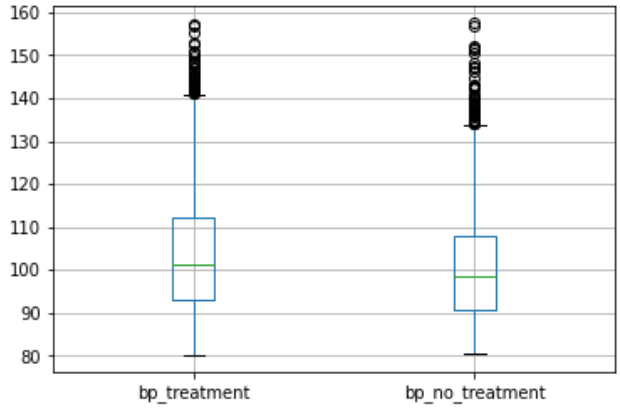

Figure 2

II. DATA PROCESSING

Firstly, we needed to transform all categorical variables into dummy variables using get_dummies() from the pandas library. This would enable the logistic regression model to properly process the categorical variables. Another processing that was done to the data set is to remove all unreasonable values. For example, it was found that some patients' blood pressure was -999, which was possibly a filler number for unavailable data. Therefore, it was necessary to remove all the rows that have -999 as blood pressure value in order to keep the raw data accurate and clean.

III. MODEL SELECTION

Since we are looking for an interpretable model. I have selected to use Logistic Regression from the sklearn library. However, to compare performance, I have also tried to classify treatment using Random Forest Classification, which is an ensemble learning method. To select the best model, we will use the auc-roc score to evaluate the overall performance of the model. The initial auc score for the logistic model after using 10-fold cross validation is 0.7925 with a standard deviation of 0.0148 (very consistent).

It looks like a pretty good score. However, when fitting the data to the random forest classification, the auc score is 0.9598 ± 0.0062. Although we will not be using random forest due to low interpretability, it is good to keep in mind for future models when interpretability is less of a concern.

To fine-tune the model, I compared the 4 different penalty methods "l1", "l2", "elastic net", and "none". These penalties are regularization methods to panalize certain coefficients to zero or near-zero, which prevents models from overfitting. The GridSearchCV results show that the best auc-roc score is produced by using l1 penalty, the auc-roc score would be slightly better (0.7999) than our first model.

IV. MODEL EVALUATION

To evaluate our model, we used the classification_report module from sklearn to calculate different scores. The precision score indicates the percentage of positives overall detected positive values. The recall score indicates the percentage of correctly predicted positive out of all positive samples. And the F-1 score is a combined score that takes into account both precision and recall.

As we can see from Figure 3, after fitting our data to the logistic model, our model has a high precision (average 0.74), high recall for predicting treatment (0.82), and high F-1 scores (0.78 for treatment and 0.69 for no treatment).

	precision	recall	f1-score	support
0	0.75	0.64	0.69	681
1	0.73	0.82	0.78	819
accuracy			0.74	1500
macro avg	0.74	0.73	0.73	1500
weighted avg	0.74	0.74	0.74	1500

Figure 3

The confusion matrix (Figure 4) shows more detail about how our model predicted the cases. This model performs worse in detecting patients that do not need treatment (more false positives), which could potentially be problematic as it would direct healthy patients to receive unnecessary treatment. Therefore, despite having this predictive model at hand, we should still ask human doctors to double-check results before confirming any treatment.

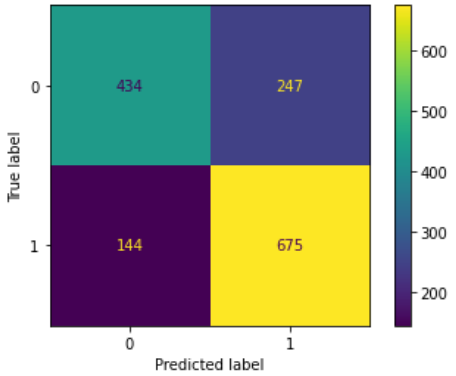

Figure 4

The AUC-ROC graph shown below also further shows that our model has moderate accuracy with AUC score of 0.73.

Figure 5

V. FEATURE IMPORTANCE

As the chart below shows, the features that are most important in predicting treatment are gender, family history, TestB and GeneE, blood test, and GeneF. Therefore, since assessments for TestA, TestB, and genes are expensive, it is useful to know that there are certain tests and genes (like Test A, GeneC, and GeneD) that are unnecessary in determining treatment.

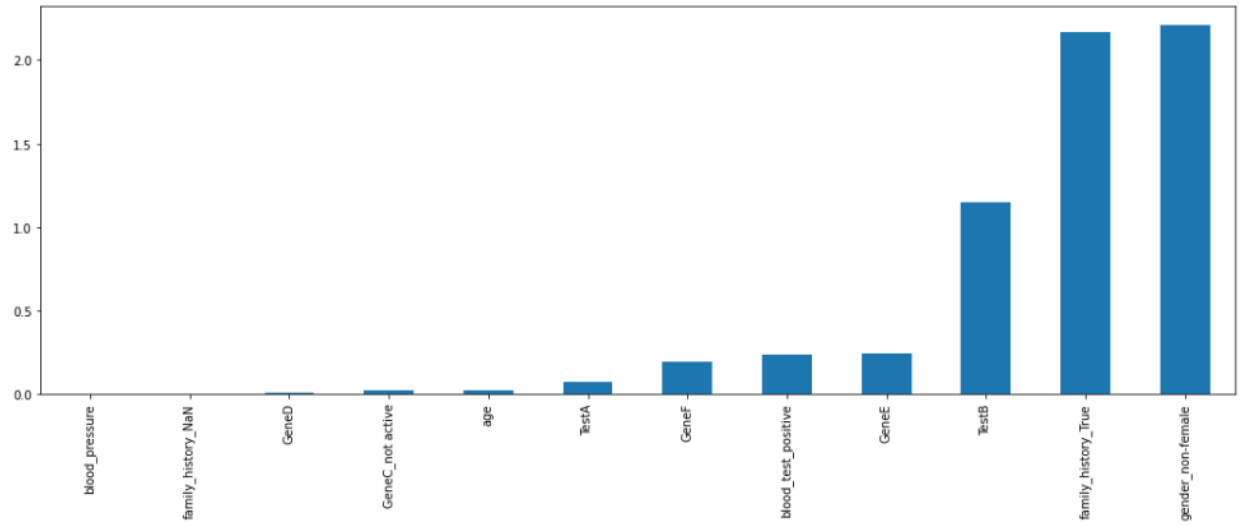

Figure 6

Specifically, the chart below shows the positive and negative correlations that each feature has with treatment. When the patient has a family history, it is a good indicator that the patient might need treatment, along with other test results. Moreover, being female also positively correlates with needing treatment. In terms of tests and genes, low scores in TestB, GeneE, GeneF, and blood tests all indicate the need for treatment.

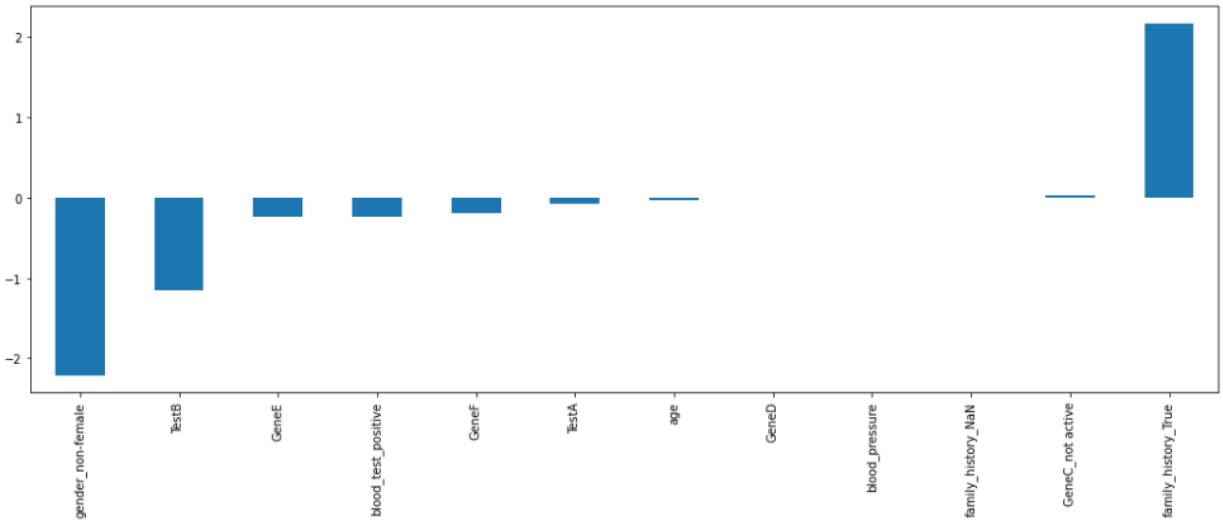

Figure 7

Note: It was found from calculating VIF scores that age and blood pressure have high multicollinearity. However, since neither of these variables is important in predicting treatment, this finding is insignificant to the report.

VI. INTERPRETATION

According to our logistic model and feature evaluation, our model has accurate performance (auc-roc score of 0.73). Although the logistic model is not good enough to completely replace human doctors in diagnosing treatment, this model is still a very good candidate in conducting preliminary assessments.

Moreover, our model also found that TestB is a highly relevant test in diagnosing patients, which the hospital should invest in. GeneE and GeneF are also useful indicators for needing treatment. On the contrary, the results show that GeneD and GeneC have very low importance in predicting treatment, which the hospital can consider not invest in these tests to lower cost.

VII. CONCLUSION

The research concludes that the use of the logistic model gives moderate accuracy in predicting treatment. It was also found while conducting testing that other classification models, like random forest classification, can produce more accurate results. We can bring this up with our hospital client to discuss if they want to consider a less interpretable but more accurate model.

Lastly, it was found that TestB, GeneE, and GeneF should continue to be used to predict treatment while GeneD and GeneC can be discarded to reduce cost.