PREDICTION OF HEART DISEASE FROM BIOMEDICAL INDICATORS

GROUP ASSIGNMENT REPORT, FOUNDATIONS OF DATA SCIENCE

JULY 29TH 2019

GROUP 11:

XIAO TIAN HE
KAREN PARRA
NICHOLAS ROMANCHUK
PATRICK DALEY
KHRIS WILHELM

EXECUTIVE SUMMARY

Automating medical diagnosis to reduce human error and speed up treatment has been a goal of many medical professionals for decades. In 1989, a researcher published a study using logistic regression to predict whether patients with who had been referred for angiograms were suffering from coronary artery disease. That research made use of all the information expected to be clinically relevant from a standard battery of tests. Some of this data may have been less relevant than was assumed, which would make some of the tests unnecessary, and could actually have reduced the discriminant function's accuracy.

We attempted to find out if it is possible to reduce the number of tests necessary to screen for a disease in a clinical setting without impacting diagnostic accuracy. Our hypothesis was that a model based on a carefully selected subset of test results can match or exceed the accuracy of a model built using all the available clinical information.

The data source for the analysis are the clinical and noninvasive test results for a group of patients undergoing angiography at the Cleveland Clinic in Cleveland, Ohio. Features from the dataset were selected for the model based on modified Kolmogorov-Smirnov goodness-of-fit tests to compare the distribution of each variable among patients with heart disease with that of patients without.

The training data was used to fit a generalized linear model (GLM) with logit (logistic regression) classifier to the selected and normalized variables. We also trained a support vector machine (SVM) using the same training data.

The estimated predictive accuracy of our support vector machine classifier is 81%, which is a modest improvement on the predictive power of the discriminant function developed in the original 1989 paper using all 13 'clinically significant' variables. Our classifier was built using only a subset of those variables, selected based on the difference in their distributions (K-S statistic) between patients with and without coronary disease. This demonstrates that careful tuning of machine learning models can reduce the number of tests that heart patients must be subjected to, without reducing the diagnostic accuracy of automated screening.

TABLE OF CONTENTS

INTRODUCTION

Automating medical diagnosis to reduce human error and speed up treatment has been a goal of many medical professionals for decades. In 1989, a researcher published a study using logistic regression to predict whether patients with who had been referred for angiograms were suffering from coronary artery disease.[1] That research made use of all the information expected to be clinically relevant from a standard battery of tests. It is conceivable, however, that some of this data may have been less relevant than assumed, which would make some of the tests unnecessary, and could actually have reduced the discriminant function's accuracy.

OBJECTIVES:

GOALS AND RESEARCH QUESTIONS:

Is it possible to reduce the number of tests necessary to screen for a disease in a clinical setting without impacting diagnostic accuracy? By quantitative analysis of the results of standard test results, we hope to identify which are actually critical to effective diagnosis, and which are unnecessary.

HYPOTHESIS:

The key testable hypothesis of this study is that a model based on a carefully selected subset of test results can match or exceed the accuracy of a model built using all the available clinical information.

DATA PREPARATION:

DATA SOURCE:

The data source for the analysis are the clinical and noninvasive test results for a group of patients undergoing angiography at the Cleveland Clinic in Cleveland, Ohio. The dataset was first published in the American Journal of Cardiology in 1989[2], and an anonymized subset (14 of the 76 original variables) was made publicly available on the UCI Machine Learning Repository[3] and the Kaggle[4] dataset repository by David Aha. This subset corresponds to the 13 predictor variables used in the initial study, and a 14th representing whether the patient's angiogram showed substantial (over 50%) narrowing.

[1] Detrano, R., 1989, "International application of a new probability algorithm for the diagnosis of coronary artery disease." *American Journal of Cardiology* 64.

[2] Ibid.

[3] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] https://www.kaggle.com/ronitf/heart-disease-uci

The dataset was downloaded from Kaggle in tabular format, and imported into a Pandas dataframe using Python 3. It contained records for 303 anonymous patients, consisting of both medical test results such as blood cholesterol and heart rate, and limited demographic information such as age and sex.

DATA QUALITY:

The dataset was checked for missing values, skewness, and outliers, any of which can have an impact on the accuracy of models derived from the dataset:

Missing values: if any of the records in the dataset were missing information for one or more variables, they could impact the model and reduce its effectiveness. The dataset was checked, but none of the 303 records had missing values.

Skewness: Skewness is the degree to which the distribution of a variable is asymmetrical. Many statistical tests work much better on normally distributed data. A distribution that is strongly skewed has the majority of its values occur towards the top or bottom of its range. While a logistic regression does require that its predictor variables be normally distributed, a highly skewed variable might not improve the model's predictive strength. However, all continuous variables were within an acceptable range (less than ±2).

Outliers: Likewise, records with values much higher or lower than normal can impact the effectiveness of a model. Of the 303 records in the dataset, 9 were found to have a value for at least one continuous variable that was an outlier (more than 3 standard deviations from the mean). It was decided that these would not impact the accuracy of the model.

DATA PREPARATION

In order to measure the actual effectiveness of a classifier on new (previously unseen) data, it is necessary to test it on known data that was not used to train the classifier. There are several common approaches to this, but the method selected was a train-test split. The order of the records in the dataset was randomized, and 70% of the records were assigned to be training data for the model, while 30% was saved for testing the model's accuracy.

ANALYSIS:

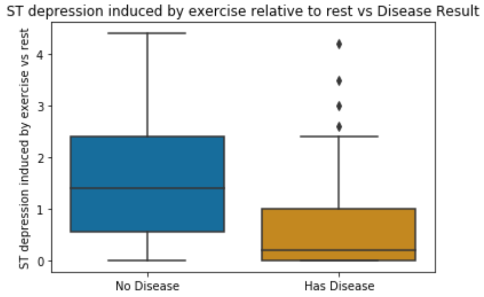
INITIAL DESCRIPTIVE ANALYSIS

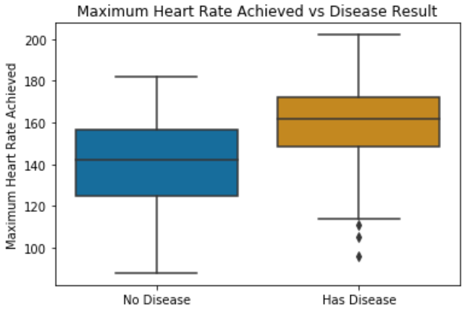
The data was loaded into python using the ***pandas*** package. The first 5 rows were plotted to ensure that the column headers and indexes were loaded correctly. This was performed by using the ***.head()*** method. Next the data set was checked for any missing values, using the ***.isnull().any()*** methods. Since the data has already been 'processed' as mentioned above, no missing values were identified. Next, the skew of each attribute was accessed via ***.skew()***, with ±2 used as the threshold to determine asymmetry. Summary statistics were then plotted for each of the attributes to confirm no missing values and identify basic descriptive information about the data set. This descriptive summary was done through the ***.describe()*** method.

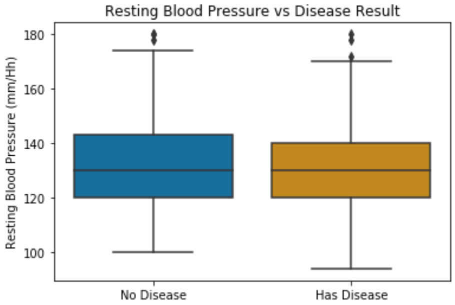
Table 1: Variables in heart disease dataset

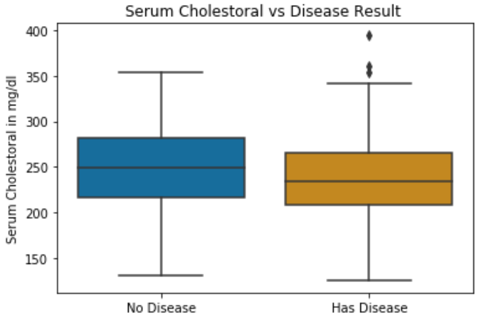
Variable	**Description**
age	Patient's age in years
sex	Patient's sex
cp	Chest pain type (4 values)
trestbps	Resting blood pressure
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar > 120 mg/dl
restecg	Resting electrocardiographic results (values 0,1,2)
thalach	Maximum heart rate achieved
exang	Exercise induced angina (yes or no)
oldpeak	ST depression induced by exercise relative to rest ('ST' refers to a portion of the ECG plot)
slope	The slope of the peak exercise ST segment in the ECG plot
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	Presence of a blood disorder called thalassemia (normal, fixed defect, or reversible defect)
target	Presence (1) or absence (0) of heart disease

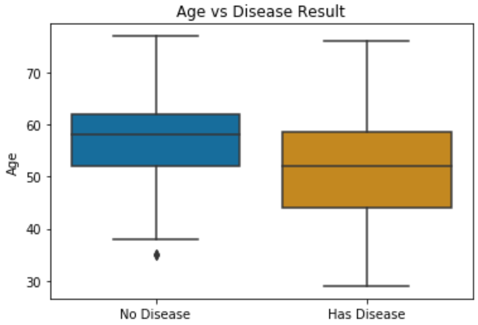
Comparing the values of the continuous variables with the presence or absence of heart disease shows that while some differ greatly between the two groups, some measures in the dataset seem to have very little ability to discriminate between people with and without heart disease:











FEATURE SELECTION

Features from the dataset were selected for the model based on modified Kolmogorov-Smirnov goodness-of-fit tests to compare the distribution of each variable among patients with heart disease with that of patients without. This test both determines whether the two distributions are significantly different, and also provides a measure of the degree of difference in the form of the KS statistic.[5] A level of 20% (or 0.2) was selected as the minimum for a variable to be included in the model.

Table 2: KS statistics for feature selection

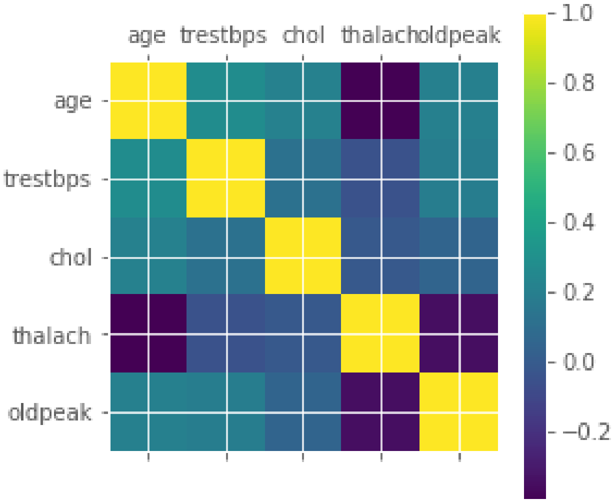
Continuous		Categorical	
age	0.119368	sex	0.262451
chol	0.051647	cp	0.51726
oldpeak	0.307510	fbs	0.020026
thalach	0.137549	restecg	0.176021
trestbps	0.073386	exang	0.411331
		slope	0.394862
		ca	0.461792
		agegroup	0.106719
		OldPeak2	0.403426
		RestBloodPressure	0.006061
		MaxHeartRate	0.101186
		propension	0.**22**5955

The variables that were retained based on their K-S statistic are highlighted in green in Table 2, above.

A correlation matrix was also generated for the continuous variables, to identify any redundant indicators. Strong correlations between different indicators can indicate that they are redundant, and may reduce the predictive accuracy of the model.

[5] Dufour, J.-M., and Farhat, A. (2001). *Exact nonparametric two-sample homogeneity tests for possibly discrete distributions.* Cahier 2001-23 (working paper), Département de sciences économiques - Université de Montréal.

Figure 1: Correlation matrix of continuous variables



While many of these variables do have positive or negative correlations, meaning that they are at least somewhat related, the strongest is the (negative) correlation between **thalach** and **age** of r = -0.39. This means that a change in one of these two variables can predict about 15% of the change in the other. This is not a strong enough relationship to be of concern, so no variables needed to be excluded due to autocorrelation.

FEATURE TRANSFORMATION

DERIVED VARIABLES

While the variables are not strongly autocorrelated, many of them do interact. In some cases, the influence of uncorrelated variables may be important. One such is the link between age (**age**) and the maximum heart achieved during exercise (**thalach**). While **thalach** by itself was excluded from the model because its distribution did not vary strongly enough between people with and without heart disease, it was used to generate a new variable, **Propension**, which identified individuals who were both over the age of 58 and had a high heartrate during exercise.

NORMALIZATION

Because most of the variables were measured in different units, large differences in the apparent scale exist. **oldpeak**, for example, range from 0 to 4, while **chol** (blood cholesterol) goes as high as 564. This sort of imbalance can degrade the regression's predictive accuracy by giving differences in some variables more weight than others. For this reason, all variables selected for the model underwent a logit transformation (the natural logarithm of the ratio of one value to the other), which effectively put them all on the same scale.

FITTING MODELS

The training data was used to fit a generalized linear model (GLM) with logit (logistic regression) classifier to the selected and normalized variables. While there are a number of alternative methods to develop classifiers, a logit was used in 1989 when the dataset was assembled, which makes it the best way to compare the impact of our feature selection and transformation on prediction accuracy.

$$logit\left(\frac{\pi}{1-\pi}\right) = x^T\beta$$

An alternative classifier is a support vector machine (SVM). This is are commonly-used machine-learning algorithms that can perform the same sort of classification task as the logistic regression, but may provide a higher or lower level of accuracy depending on the task and dataset.

$$k(x_i, x_j) = tanh(ax_ix_j - b)$$

The first iteration of the GLM was fitted using the set of variables previously selected based on the K-S statistic. For each of these, performance indicators were calculated to identify the degree to which each variable contributes to the model's ability to predict the target variable.

Variable	p value
Intercept	2.152368e-43
LN Propension	3.385478e-06
LN ca	2.231922e-06
LN slope	1.337189e-03
LN exang	1.820553e-04
LN cp	1.385082e-12
LN sex	4.001657e-01

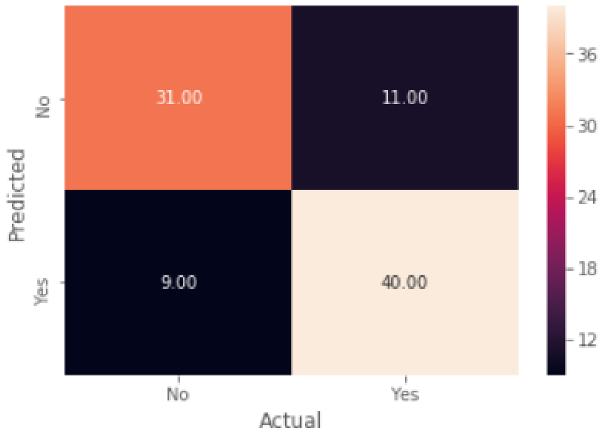
An threshold of α=0.05 was selected. For these purposes, this means that *p*-values of less than 0.05 indicate that the variable adds to the model's predictive ability. Sex was found to have a *p*=0.4, meaning it was not contributing to the model's accuracy in detecting heart disease. A second GLM was generated, excluding sex.

Variables	P Values
Intercept	1.419439e-51
LN Propension	4.102678e-06
LN ca	1.375774e-06
LN slope	1.507415e-03
LN exang	1.474819e-04
LN cp	1.260075e-14

TESTING MODELS

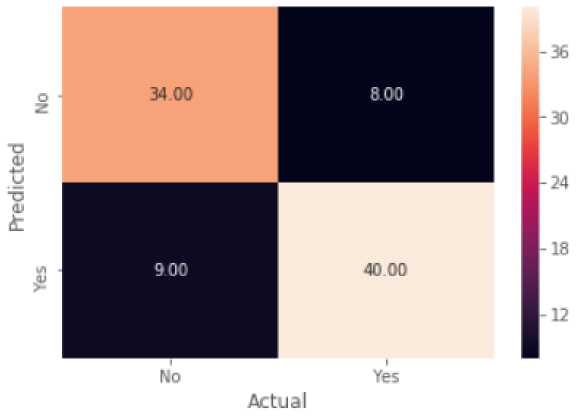
The accuracy of the model was tested by using it to make predictions for the 30% of records that were help back as testing data. Those predictions (yes or no for heart disease) were then compared with the actual presence or absence of heart disease for each patient.

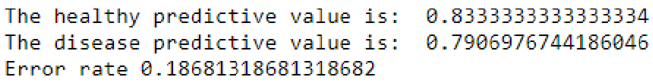
Figure 2: Confusion matrix for GLM



Out of the 91 patient records held back for testing, the GLM classifier correctly predicted 71 of them. There were 11 false negatives (people who actually had heart disease that was not detected by the classifier). There were also 9 false positives (people that the classifier predicted would have heart disease, but did not in reality). This works out to a 77.5% rate of correct predictions for healthy individuals, and 78.4% for individuals with heart disease.

Figure 3: Confusion matrix for SVM





```
The healthy predictive value is:  0.8333333333333334
The disease predictive value is:  0.7906976744186046
Error rate 0.18681318681318682
```

From the same 91 patient records, the SVM classifier correctly predicted 74 of them. There were 8 false negatives. There were also 9 false positives. This works out to an 83.3% rate of correct predictions for healthy individuals, and 79% for individuals with heart disease.

Because the result from a single test are not necessarily a good measure of exactly how well the model will do on average, other techniques are used to get a better idea of its actual accuracy compared to other models. One method is the ROC (receiver operating characteristic) curve:

Figure 4: ROC curve for GLM

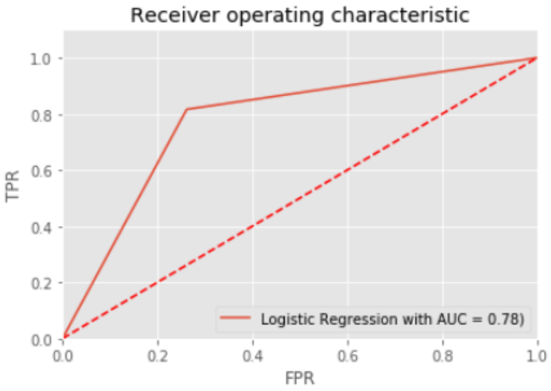
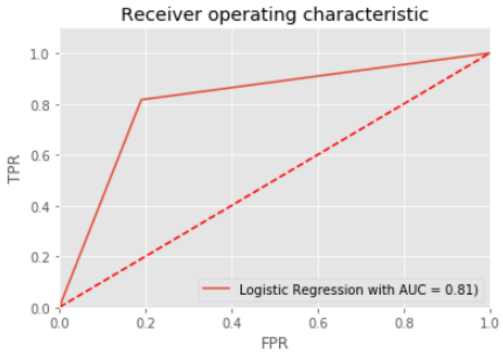


Figure 5: ROC curve for SVM



```
The GINI coefficient is:  0.6258503401360545
```

The area under the curve (AUC) of the GLM model was 0.78. The AUC represents the approximate accuracy: a score of 1 would mean perfectly accurate, while a score of 0.5 would mean it was only accurate half the time, equivalent to randomly guessing yes or no. This means that our classifier was quite successful at predicting whether a given patient had heart disease or not. The area under the curve for the SVM was somewhat better, at 0.81

CONCLUSIONS:

The estimated predictive accuracy of our support vector machine classifier is 81%, which is a modest improvement on the predictive power of the discriminant function developed in the original 1989 paper using all 13 ‘clinically significant’ variables. Our classifier was built using only a subset of those variables, selected based on the difference in their distributions (K-S statistic) between patients with and without coronary disease.

Careful analysis and selection of the source variables used to develop predictive algorithms has the potential to improve the effectiveness of existing automated clinical diagnosis systems. It could also result in the reduction of healthcare costs and the need for some invasive procedures by identifying which tests are actually necessary to effectively screen patients for a disease. The use of more modern algorithms such as support vector machines can also improve the accuracy in some cases, even when working from a smaller training set than the original study.

There are some caveats and directions for further work arising from this analysis. This classifier was developed based on a population who were not randomly selected, but had instead already been referred for angiograms. The model would probably have a lower rate of successful predictions if it was used on random members of the public.

There are possibilities for followup research projects, such as fully replicating the original study by testing our classifier on data from other groups of patients, or by adding those other values to the dataset and using cross-validation to see if a more widely-applicable model can be developed.