Capstone Project 1 Report
Heart Disease Predictor

Brian Lubeck

December 9, 2017

Introduction

Heart disease is the build up of plaque in a person's arteries and is the leading cause of death in the United States. More than 600,000 Americans die from it each year, with one out of four deaths each year being from heart disease. As the plaque builds up, the arteries narrow, reducing blood flow to the heart. A doctor can perform several tests to diagnose heart disease, including chest X-rays, coronary angiograms, electrocardiograms (ECG or EKG) and exercise stress tests. Fortunately, heart disease is treatable. Eating a healthy diet, exercising regularly, maintaining a healthy weight and taking medications are four ways to treat it and to reduce the risk of developing it in the first place.

To accurately determine if a person has heart disease, we built and compared three classifiers - naive Bayes, logistic regression and random forest. This report presents the steps we took to build the classifiers and their performance. First we discuss and explore the dataset we used to build our classifiers. Then we discuss each classifier we built. Lastly, we end with some concluding thoughts. It is our hope that this report will inform doctors which medical tests to perform on a patient and what personal information to collect in order to accurately diagnose heart disease.

Data and Initial Exploration

We used the Cleveland heart disease dataset from the University of California, Irvine's Machine Learning Data repository[1] to build and test our classifiers. The dataset consists of medical information collected from patients at the Cleveland Clinic Foundation from 1988. The dataset has 14 fields and 303 observations. Each row represents a different patient. The field of primary interest is *num*. It indicates the degree to which the diameter of a person's arteries has narrowed. It has five possible values 0 to 4 with 0 being less than 50% narrowing and $1, 2, 3$ and 4 being greater than 50% narrowing. As in previous studies, we took a value of 1 or greater to mean that the patient has heart disease. We created the field HD which equals 1 if $NUM \geq 1$ and 0 otherwise. We used the other 13 fields in the dataset to predict the value of HD. See the appendix for a complete description of the fields.

The first steps in the analysis were to read in the data, assess its quality, and clean it as needed. The data was read into a pandas data frame object. Based on their descriptions, the fields were each given a numpy data type $np.float\_$ or $np.int\_$. Two of the fields CA and THAL were unable to be read in as integers because they had missing values, which had been marked as question marks in the data. Four records had missing values for CA, and two records had missing values for THAL. To handles the missing values, we simply replaced the question marks with -1.

Next, we looked at the distributions of the 14 data fields to assess if the data appeared reasonable and to detect any outliers. Table 1 displays

[1] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

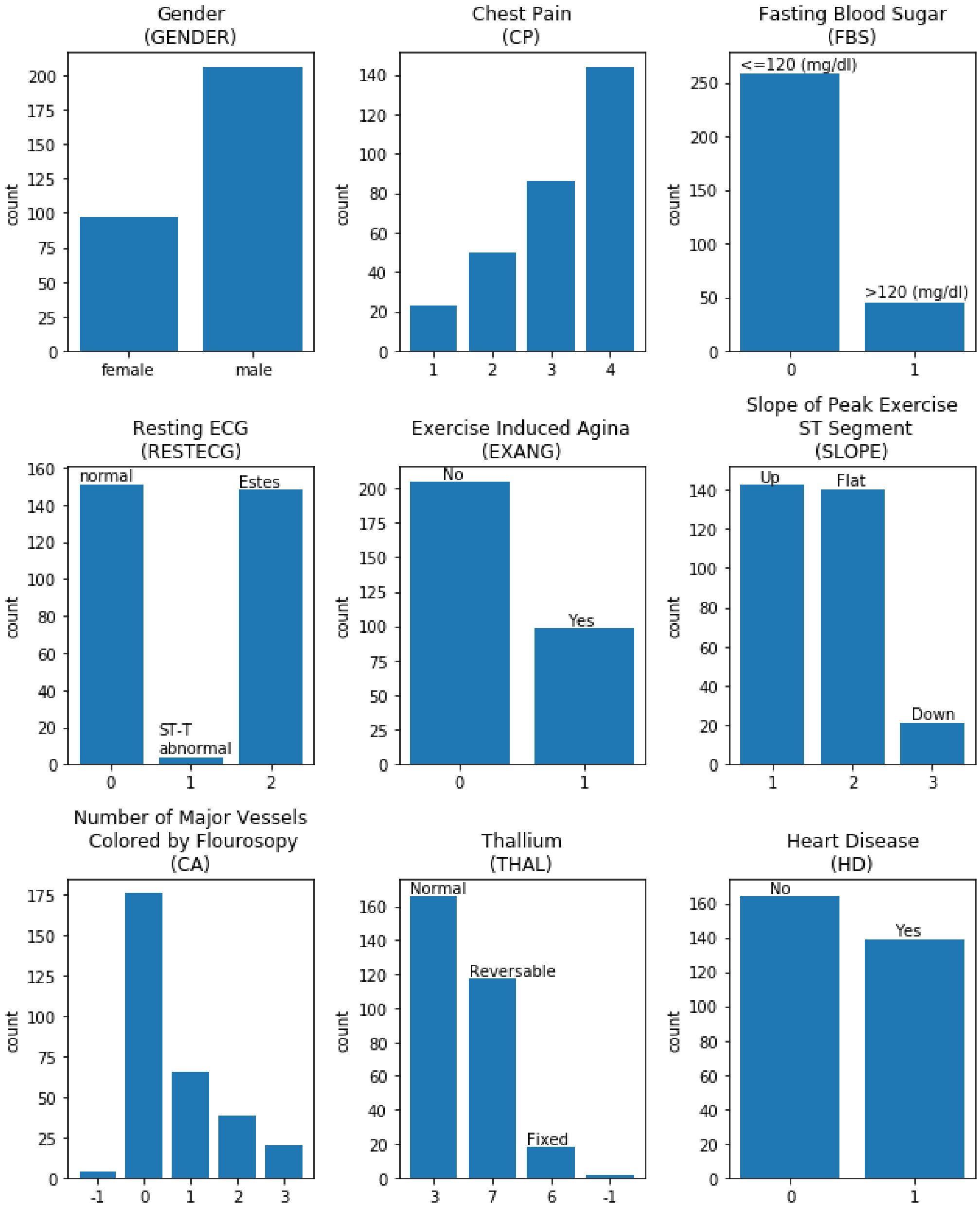
quantiles for the continuous variables, and Figure 1 displays histograms for the discrete variables. The distributions for AGE and GENDER appeared reasonable. The minimum age in the dataset is 29 and the 10^{th} percentile is 42. People younger than 40 tend not to have heart disease, and more men than women are more affected by heart disease. There were not any outliers for AGE either. Assessing the distributions of the remaining variables required medical knowledge. For example, a cardiologist would know what the results of an electrocardiographic exam typically are for a person with heart disease. In general, having an understanding of the data is required to assess the reasonableness of the data at a high-level and low-level.

Table 1: Quantiles of Continuous Data Fields

Field Name	Min	10th	25th	50th	75th	90th	Max
AGE	29.00	42.00	48.00	56.00	61.00	66.00	77.00
TRESTBPS	94.00	110.00	120.00	130.00	140.00	152.00	200.00
CHOL	126.00	188.80	211.00	241.00	275.00	308.80	564.00
THALACH	71.00	116.00	133.50	153.00	166.00	176.60	202.00
OLDPEAK	0.00	0.00	0.00	0.80	1.60	2.80	6.20

The next step was to develop an understanding of the relationships among the data fields. Since we were interested in predicting if a patient has heart disease, we looked at the relationships between heart disease and the other data fields. Chest pain is often a sign of a heart attack and is usually the first concrete indication a person receives that something is wrong. Table 2 shows the relationship between heart disease (HD) and chest pain type (CP). As can be seen from Table 2, 21% of the observations have heart disease if chest pain type is 1,2 or 3 but 73% of the observations have heart disease if chest pain type is 4. Based on this, we should include chest pain type in our classifiers. It also raises the question if it is worth the cost to educate people

Figure 1: Frequency Charts of Discrete Data Fields



about chest pain symptoms and how to recognize the difference betweeen asymptomatic pain (4) and other types of pain (1,2 or 3).

Table 2: Contingency Table of Heart Disease and Chest Pain

	Heart Disease (HD)		
Chest Pain (CP)	No	Yes	Total
1 (typical angina)	16	7	23
2 (atypical angina)	41	9	50
3 (non-anginal pain)	68	18	39
4 (asymptomatic)	39	105	144
Total	164	139	303

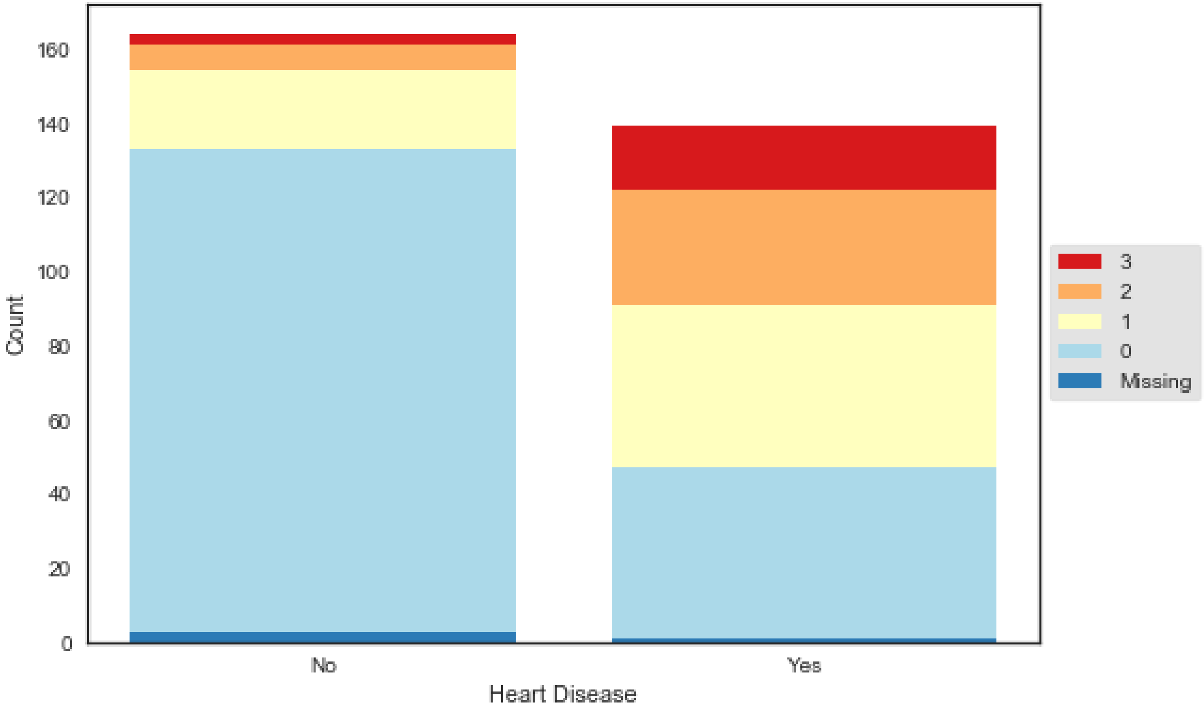
Next, we looked at how well a single medical test does at determining if a person has heart disease. Flouroscopy is medical imagining technique that uses X-rays to visualize parts of the body. One of its application is to visualize blood vessels and organs. Figure 2 shows the relationship between heart disease (HD) and the number of major vessels colored by flouroscopy (CA).

From Figure 2, we can see that having one or more major vessels colored by flouroscopy is indicative of heart disease. Thus flouroscopy is a useful test for a doctor to help diagnose heart disease.

Having explored the relationship between heart disease and two other data fields, we next explored how correlated all of the discrete data fields were with heart disease and each other. To answer this question, we calculated Cramer's V for each combination of discrete fields.[2] Figure 3 is a correlation matrix plot that shows the correlations compared to each other. As can

[2] Cramer's V is the χ^2 statistic normalized to lie between 0 and 1. A value of 0 means no relationship, while a value of 1 means a perfect relationship. One rule of thumb is that below 0.1 indicates a weak relationship, between 0.1 to 0.3 indicates a moderate

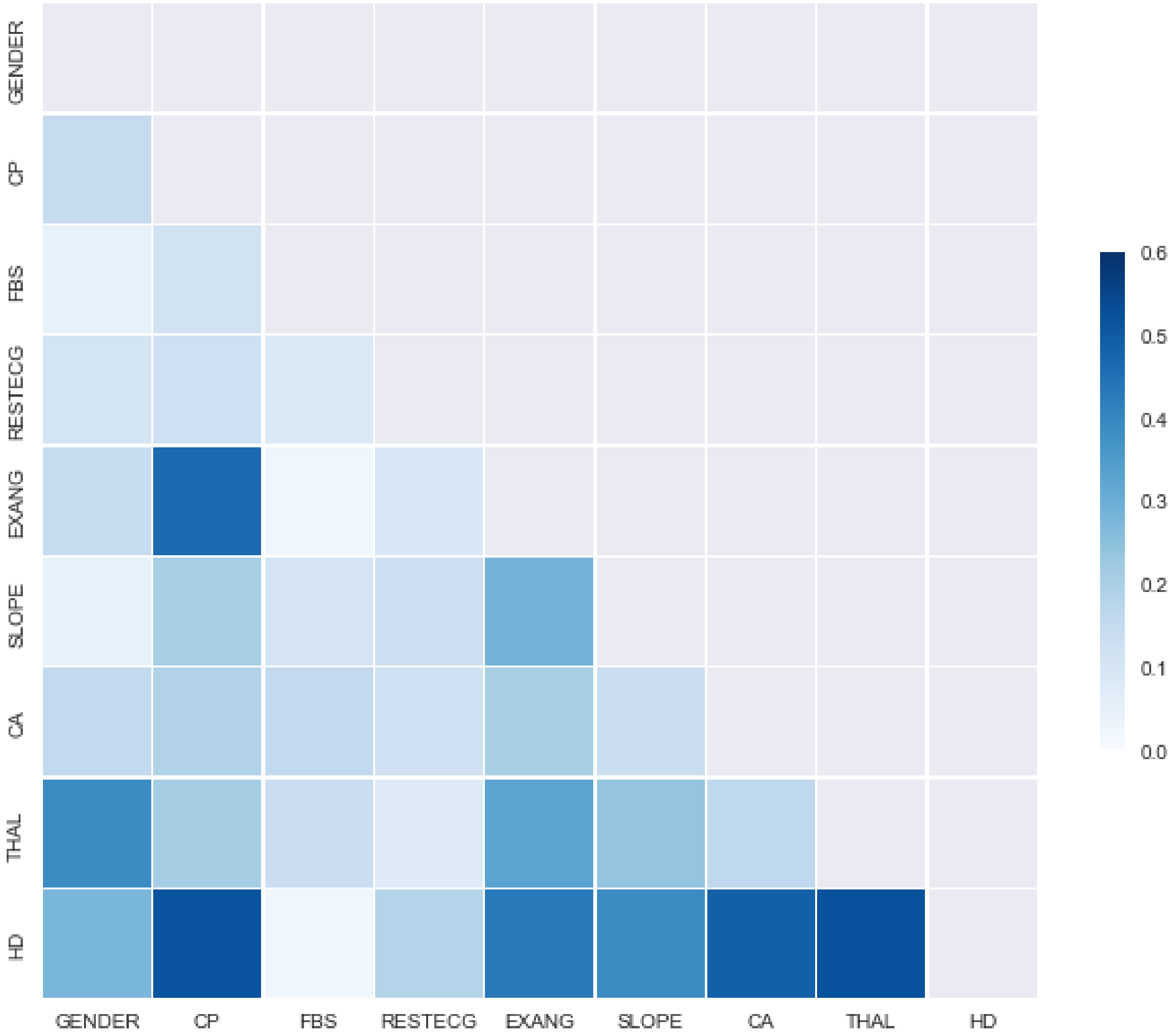
Figure 2: Flouroscopy versus Heart Disease



be seen from Figure 3, chest pain type (CP), thallium stress test results (THAL) and flouroscopy (CA) have the strongest associations with heart disease among discrete fields. A thallium stress test is a nuclear imaging test that shows how well blood flows into the heart while exercising or resting. In addition, it appears that chest pain, thallium stress test and flouroscopy have a weakly moderate relationship. When building a model it is desirable to have predictor variables that are not strongly correlated with each other so that their individual effects can be more easily determined. On the other hand, a person's fasting blood sugar (FBS) has essentially no relationship with heart disease. Thus, this field can be safely excluded from our heart disease prediction model. Knowing how correlated each medical test is with heart

relationship and above 0.3 indicates a strong relationship

Figure 3: Correlation Matrix - Discrete Variables



disease can help doctors and patients determine if the additional information is worth the additional cost of performing the medical test.

Lastly, before we build any formal prediction models, we explore the association among the continuous variables. To do this, we calculated the sample

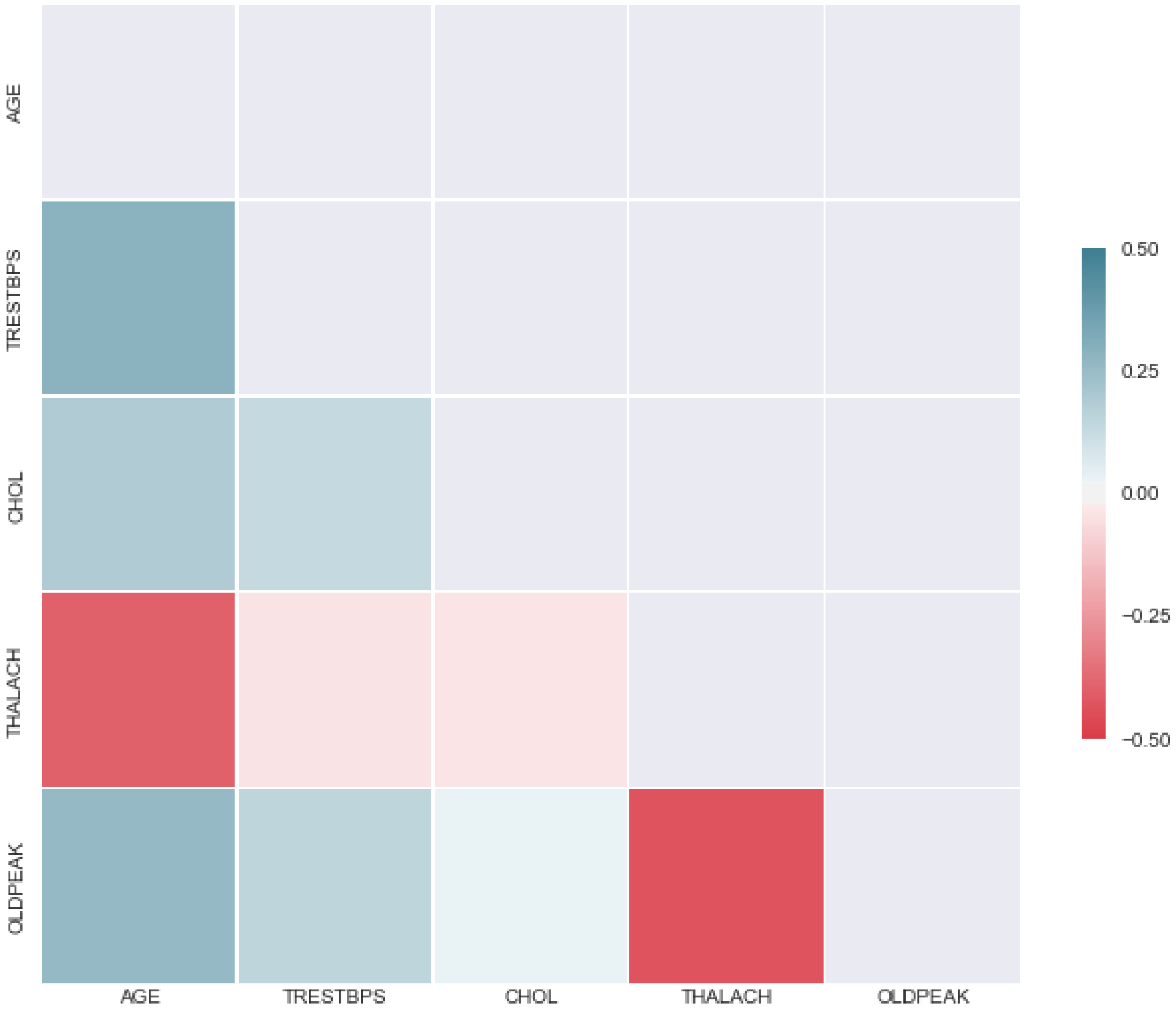
Spearman rank correlation for each pair of continuous variables. Like, Pearson correlation, the Spearman correlation is a number between -1 and 1. A value of 1 indicates the two variables are an increasing function of each other, and a value of -1 indicates the two variables are a decreasing function of each other. As can be seen from Figure 4, there is a moderate decreasing relationship between age (AGE) and maximum heart rate achieved (THALACH), which is to be expected.

Predictive Models

In this section, we describe the three binary classification models, naive Bayes, logistic regression and random forest, that we built to predict heart disease based on the other fields in our dataset. Our goal was to maximize the recall of each model. We used the recall metric $P(Classifier = 1|HD = 1)$ because it is important to identify everyone who potentially has heart disease since it is a serious medical condition. Other metrics such as accuracy, precision and F1-score are important, and this section reports these statistics as well.[3] However, even if the classifier has poor precision, the consequences of incorrectly classifying someone who does do not have heart disease are not serious because the first steps in treating heart disease are usually to eat a healthy diet and to begin exercising regularly - activities that maintain and promote good health to begin with. Before prescribing medication that has potentially serious side effects, a doctor could conduct further, specialized medical tests beyond the medical tests included in our dataset or utilize a classifier tuned to maximize precision. To train and test our classifiers, we stratified the dataset by HD and then randomly assigned 70% of the observations in each stratum to the training dataset and the remaining 30% of observations to the testing dataset. We stratified on HD so that the propor-

[3] The F1-score $F1 = (2 \cdot precision \cdot recall)/(precision + recall)$.

Figure 4: Correlation Matrix - Continuous Variables



tion of observations with HD = 1 was the same between the two datasets.

Naive Bayes

The first classifier we built was our baseline classifier using the naive Bayes binary model. Let $X = (X_1, X_2, \ldots, X_n)$ be an n x 1 discrete random vector. Let Y be a binary random variable. Recall that, using Bayes' Theorem and the conditional independence assumption, the naive Bayes binary classifier is

$$\hat{y} = \underset{y \in \{0,1\}}{\arg\min} \, P(Y = y) \prod_{i=1}^{n} P(X = x_i | Y = y)$$

The probabilities are most commonly fit using maximum a posteriori estimation.

For our model, we used the fields age (AGE) , gender (GENDER), chest pain type (CP) and resting blood pressure (TRESTBPS) to predict heart disease (HD). We chose these fields and the naive Bayes model to construct our classifier because the naive Bayes model is easy to use and interpret and age, gender, resting blood pressure and chest pain are easy and cheap pieces information to collect. A person knows his or her age and gender. Most likely, a person knows his or her resting blood pressure and possibly his or her chest pain type. If not, then a doctor can easily determine these pieces of information during an office visit. Thus, our baseline classifier is a practical model that hopefully also has good predictive power.

The BernoulliNB class in Scikit-Learn requires that the features be binary-valued, so we first converted the continuous features age and resting blood pressure into five intervals where the endpoints of each interval are the $0, 20, 40, 60, 80$ and 100 quantiles, respectively. Table 3 below shows how age and resting blood pressure were converted into intervals. Then we converted the intervals into dummy variables to be fitted. We also converted gender and chest pain type into dummy variables.

Table 4 below shows the accuracy, precision, recall, F1-score and area under the ROC curve for our naive Bayes classifier on the training and testing

Table 3: Intervals for AGE and TRESTBPS

	Interval				
Field Name	1	2	3	4	5
AGE	[29, 45)	[45, 53)	[53, 58)	[58, 62)	[62, 77]
TRESTBPS	[94, 118)	[118, 125)	[125, 132)	[132, 142)	[142, 200]

datasets. Accompanying the heart disease dataset was a cost dataset that contained the dollar cost (in Canadian dollars) of obtaining each field in the heart disease dataset. The cost information was taken from the Ontario Health Insurance Program's fee schedule. Based on this dataset, the cost of our naive Bayes classifier is $4.

Table 4: Evaluation Metrics for Naive Bayes Classifier

Dataset	Accuracy	Precision	Recall	F1-Score	AUC
Training	0.811321	0.793814	0.793814	0.793814	0.873644
Testing	0.736264	0.695652	0.761905	0.727273	0.821429

Logistic Regression

The next model we built was a logistic regression classifier with $L2$ regularization. Logistic regression models the probability that a random binary variable is 1 given a random vector X. With $L2$ regularization, the model is

$$P(Y = 1|X = x, \beta) = \frac{1}{1 + e^{-(\beta^t x)}} + \lambda \frac{1}{2} \beta^t \beta$$

where $\lambda \geq 0$ is the regularization parameter. Note that $L2$ regularization can be derived by assuming that the coefficient vector $\beta \sim N(0, \tau^2 I)$ and using maximum a posteriori (MAP) estimation to calculate β. The larger the value of λ, the closer the coefficients will be to 0.

For our logistic regression model, we used the features chest pain type (CP), thallium stress test results (THAL), max heart rate during the thallium stress test (THALACH), exercise induced angina (EXANG) and serum cholesterol in mg/dl (CHOL). Since CP and THAL are categorical variables without any ordering, we converted them to dummy variables. We wanted a more sophisticated model than our baseline model so we included the medical test THAL, which has a strong correlation with HD as shown in Figure 3, to see what the improvement in performance was. To determine a reasonable value for λ, we used 10-fold cross validation on the training dataset and recall as our evaluation metric. We considered $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$ first. Recall was maximized at $\lambda = 1$. We then searched over the interval $[0.5, 2]$ in increments of 0.025, from which we determined $\lambda = 1.05$ to be the best value.

Table 5 below shows the accuracy, precision, recall, F1-score and area under the ROC curve for our logistic regression classifier on the training and testing datasets. The cost of our logistic regression classifier is \$199.47. Compared to our naive Bayes model, every evaluation metric is higher, but at a much higher financial cost.

Table 5: Evaluation Metrics for Logistic Regression Classifier

Dataset	Accuracy	Precision	Recall	F1-Score	AUC
Training	0.830189	0.827957	0.793814	0.810526	0.891618
Testing	0.769231	0.733333	0.785714	0.758621	0.858601

Random Forest

The final model we built was a random forest classifier. A random forest is a collection of individual decision trees. To classify a new vector X, the vector is feed through each classification tree and receives a class label. The random forest chooses the class label with the greatest count.

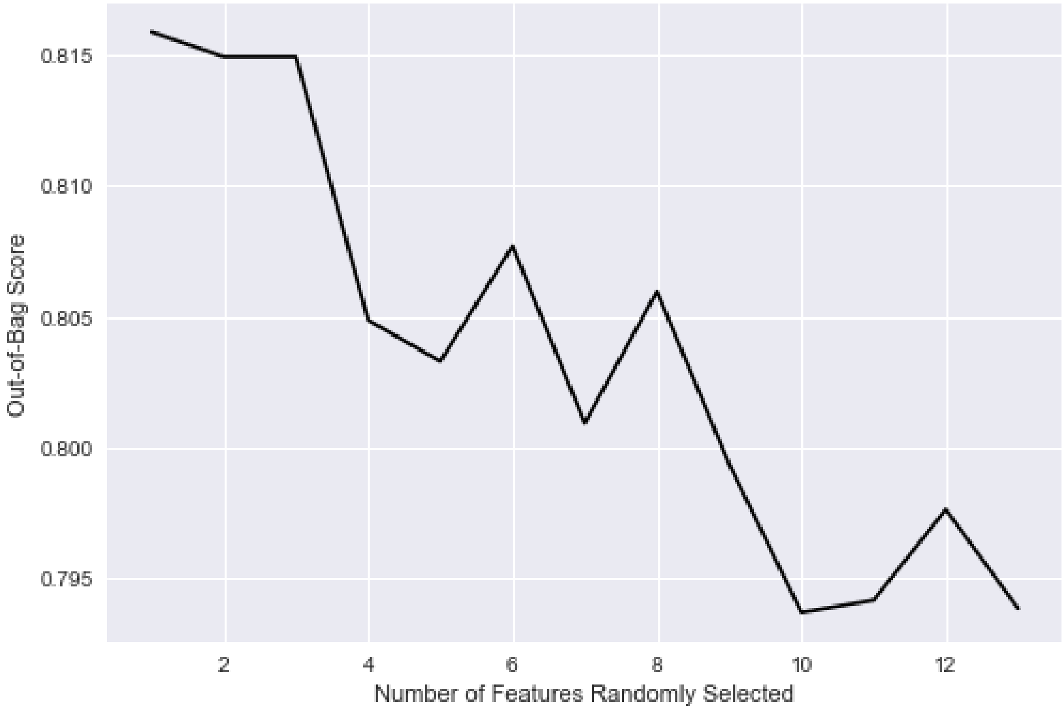
A random forest is fit to the data by fitting a selected number of individual decision trees. For each tree, N observations are chosen at random with replacement from the training dataset set also of size N. This manner of sampling, i.e. sampling with replacement from the empirical distribution, is called bootstrapping. To split a node, a subset of m features is randomly selected from the set of M features, where $1 \leq m \leq M$. These m features are then used to determine the best split. Each tree is grown to the largest extent possible.

For our random forest classifier, we used all of the features in the dataset because we wanted to capture all of the information in the data. The generalization error of the random forest is sensitive to the number of features randomly selected to determine the best split at each node. We first tuned this parameter to a reasonable value. Recall that our dataset consists of 13 independent features. For $m = 1$ to 13, we built 30 random forests consisting of 40 individual trees and calculated the average out-of-bag (OOB) score on the training dataset. Figure 5 below displays the results.

Based on Figure 5, we chose to use 2 features to build our random forest classifier. It is interesting to note that as the number of features increases, the out-of-bag score decreases. A random forest is a complex model and can result in over fitting if the number of features on which to split is too large. We were working with a small dataset consisting of only 303 observations and so over fitting was a concern. Table 6 below shows the accuracy, precision, recall, F1-score and area under the ROC curve for our random forest classifier on the training and testing datasets. The cost of our random forest classifier is $322.97. Our random forest classifier has the same recall as our logistic regression classifier but performs better on the other evaluation metrics.

Lastly, we graphically compared our three classifiers using an ROC chart.

Figure 5: Out-of-Bag Score vs. Number of Features

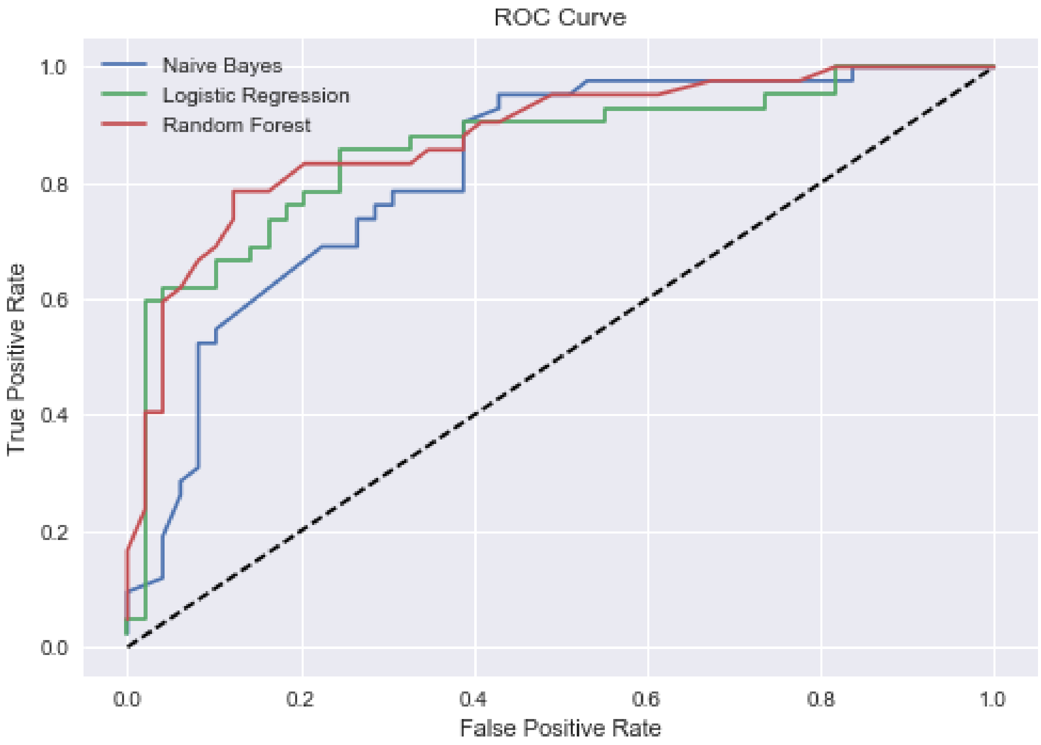


See Figure 6 below. The area under the curve equals the probability that a randomly chosen positive observation is ranked higher than a randomly chosen negative observation.

Table 6: Evaluation Metrics for Random Forest Classifier

Dataset	Accuracy	Precision	Recall	F1-Score	AUC
Training	1.000000	1.000000	1.000000	1.000000	1.000000
Testing	0.835165	0.846154	0.785714	0.814815	0.878523

Figure 6: ROC Chart



Conclusion

The goal of our analysis was to build a classifier that accurately determines if a patient has heart disease. In particular, we sought to maximize the recall of the classifier. Our random forest classifier was our best classifier in terms of recall and the other evaluation metrics. It has an accuracy of 83.52%, which is close to the best accuracy others have achieved using this dataset. The downside to our random forest classifier is that it is relatively expensive.

More than just building classifiers, our analysis also has public policy implications. The naive Bayes classifier uses common, everyday health features to predict heart disease. It has a recall of 76.19% and a precision of 69.57%. This shows that the age, gender, chest pain type and resting blood pressure are good indicators of heart disease. Furthermore, the naive Bayes model is a simple enough model that it can be explained to the general public. For example, a chart could be made of all the possible combinations of age, gender, chest pain type and resting blood pressure, along with the probability of having heart disease predicted by the model. This chart could then be used to educate people about risk factors that are associated with heart disease and that people can monitor on their own.

However several shortcomings of our analysis should be noted. The size of the dataset, 303 observations, is very small relative to the size of the general population and the number of people that die from it each year as stated in the introduction. In addition, the data was collected from patients at one hospital. Therefore, it is very possible that our dataset is not a representative sample of the general population or even those who seek medical treatment. Another shortcoming is that the data was collected in 1988. There have been many medical advances since then. New medical tests for heart disease should be included in any future analysis.

A Description of Data Fields

N	Field Name	Description
1.	AGE	Age in years
2.	GENDER	Gender (1 = male; 0 = female)
3.	CP	Chest pain type 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic
4.	TRESTBPS	Resting blood pressure in mm Hg on admission to the hospital
5.	CHOL	Serum cholestoral in mg/dl
6.	FBS	Fasting blood sugar (1 if > 120 mg/dl; 0 otherwise)
7.	RESTECG	Resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
8.	THALACH	Maximum heart rate achieved
9.	EXANG	Exercise induced angina (1 = yes; 0 = no)
10.	OLDPEAK	ST depression induced by exercise relative to rest
11.	SLOPE	Slope of the peak exercise ST segment 1 = upsloping 2 = flat 3 = downsloping
12.	CA	Number of major vessels (0-3) colored by flouroscopy
13.	THAL	3 = normal; 6 = fixed defect; 7 = reversible defect

N	Field Name	Description
14.	NUM	Diagnosis of heart disease (angiographic disease status)
		0 = less than 50% diameter narrowing
		1,2,3,4 = greater than 50% diameter narrowing
15.	HD	1 if NUM ≥ 1, 0 otherwise