Multivariate Logistic Regression for the Prediction of Coronary Heart Disease

Abstract
The purpose of this paper is to determine the factors with the greatest influence over the positive diagnosis of Coronary Heart Disease (CHD). A 1998 Data Report from the University of California, Irvine that combined data from four domestic and international medical institutions was used. The Data Report had 11 predictor variables recorded for each individual regarding their health. We found that age, sex, type of chest pain, cholesterol levels, fasting blood sugar levels, presence of exercise-induced chest pain (angina), and damage of heart tissue indicated by ECG readings (ST slope) have the most significant impact on the diagnosis of CHD. A Logistic Regression Model was built using these variables. This Model was compared to different models, including another constructed Penalized Regression model trained by the same dataset. Upon comparison, the Logistic Regression Model was ultimately chosen. To extend the model's accessibility and impact on the current rate of CHD diagnoses, an interactive web application was developed for users to make personalized predictions using the model.

Background and Introduction

Heart disease is often fatal, and it is also very common. Heart disease is the #1 leading cause of death for men and women worldwide, including within the United States. About 1 in 5 deaths in the US are due to heart disease (CDC).

80% of heart disease cases would be avoided if detected and prevented early (American Heart Association). However, currently, most people are not diagnosed until they are hospitalized for angina (severe chest pain), cardiac arrest, or heart attack (National Institute of Health). Proper treatment of heart disease may be avoided or unaddressed due to the large cost and time currently required for a personalized prediction of heart disease (UCSF Health). A patient-accessible and accurate prediction estimate of heart disease in an individual using relatively accessible health data is important and needed for early and widespread awareness of heart disease risk, prevention, and treatment.

Data and Exploratory Analysis

Data and Variables

The Hungarian Institute of Cardiology; the University Hospitals of Zurich and Basel Switzerland; and the V.A. Medical Center in Long Beach, California are four prominent hospitals serving patients of various ailments, including those related to the heart (UCI Machine Learning Repository). This data set consists of 918 patients selected through a simple random sample from the population: all patients from the four medical institutions. Out of the 76 original attributes from the data set, this study examined 11 total variables, explained below, seeking to predict whether a given patient has heart disease, the twelfth variable, as a probability.

Age is given in discrete years; the range is between 28 and 77 years. Sex is divided into two categories: Male (1) and Female (0). Chest Pain Type is divided into four categories: Typical Angina (TA, 0), Atypical Angina (ATA, 1), non-Anginal Pain (NAP, 2), and Asymptomatic (ASY, 3). Resting Blood Pressure is given in millimeters of mercury (mmHg), measured to the tenth of a millimeter; the range was between 0.0-200.0 mmHg. Cholesterol is given in milligrams per deciliter (mm/dl), measured to the tenth of a milligram; the range was between 0.0-603.0 mg/dl. Fasting Blood Sugar is divided into two categories: whether or not a patient's fasting blood sugar was greater than 120 milligrams per deciliter (1 or 0). Maximum Heart Rate is given by beats per minute, measured to the tenth of a beat; the range from 60.0-202.0 BPM. Exercised-Induced Angina was divided into two categories: whether or not they experienced it (1 or 0). Old Peak refers to ST depression induced by exercise relative to the rest of the patients; it had a range of -2.6000-6.2000. ST depression refers to a finding on an electrocardiogram wherein the trace in the ST segment is abnormally low below the baseline, indicating heart tissue damage. ST Slope refers to the slope of the peak exercise ST segment and is divided into three categories: Upsloping (0), Flat (1), and Downsloping (2). Lastly, Resting ECG Type refers to resting electrocardiographic results, and is divided into three categories: normal (0), having ST-T wave abnormality (1), and showing probable or definite left ventricular hypertrophy by Estes' criteria (2), which is another abnormality.

Exploratory Data Analysis

Fig 1. Side by side boxplots were used to compare symptoms among those with and without heart disease. 410 patients did not have heart disease, while 508 patients did; they were split into two groups by this criterion. The patient group with heart disease had a higher median age than those without heart disease, though the distributions were relatively similar in shape and spread. Medians and distributions were very similar when comparing resting blood pressure in patients with versus without heart disease. Median cholesterol levels were very similar for patient groups with and without heart disease, but while the distribution was symmetrical and spread smaller in those without heart disease, it was left-skewed, and spread was much greater for those with heart disease. The patient group without heart disease had a higher median

maximum heart rate than those with heart disease, while the shape and spread of the two groups were quite similar. Finally, the mean Old Peak—described above—is greater in those with heart disease than those without, and also has greater spread and is slightly left-skewed. For those without heart disease, median Old Peak is 0 and the distribution is heavily right-skewed towards 0.

Fig 2. Mosaic plots were also used to compare symptoms among individuals with and without heart disease. The proportion of those with heart disease was greater in males, in those with heart disease was greater in those with a fasting blood sugar level of over 120 milligrams per deciliter, and those who experienced exercise-induced angina. The majority of those with heart disease had a flat ST slope at the peak exercise ST segment, while it was upsloping for those without heart disease. People with heart disease mainly experienced asymptomatic chest pain, while the type of chest pain felt by those without heart disease was more uniformly distributed, with very little experiencing typical angina. Finally, the Resting ECG types were mostly similar across the groups with and without heart disease, with most patients having normal ECG type.

Model and Results

Analytical Methods

A multiple logistic regression model was fitted to our data, with presence of Heart Disease as the response variable. A logistic regression model was employed because our response variable is binary, and our predictors include both categorical and quantitative variables. Then, we employed stepwise selection procedures to select the best predictors for prediction of CHD. An intermediate model was calculated, and the number of predictors in the model were reduced from the original 11 predictors. Predictors Resting ECG, Resting Blood Pressure, Maximum Heart Rate, and Peak ECG Readings were dropped, as there proved to be no significant difference between mean proportions of heart disease of those variable groups - leaving 7 predictors in our final model. Examination of empirical logit plots (**Fig. 3**) demonstrated that the relationships between the predictor variables and log odds were relatively linear, meaning no transformation of our regression model was necessary.

Final Model:

Our final model requires information about age, sex, cholesterol levels, fasting blood sugar, exercise-induced angina, ST slope, and type of chest pain. **Table 1** summarizes the Estimate, Standard Error, and P-value of the significant variables. A likelihood ratio test was performed to compare our final model to an intercept-only model to determine the effectiveness of our prediction model, and yielded a p-value of 2.2 x 10^{-16}. Our variable plot demonstrates that there is no significant collinearity between predictor variables, with the highest correlation coefficient of 0.43 between exercise-induced angina and ST_Slope (**Fig. 4**). Calculations of our confusion matrix using the probability threshold of 50% show that 86.38% of the observations in our dataset were predicted accurately, with specificity of 82.93%, and sensitivity of 89.17%. (**Fig. 5**), and thus our model may be used for prediction of Coronary Heart Disease. The Wald test testing significance of each predictor in our final model given the other predictors showed that all predictors were, indeed, had a statistically significant relationship with Heart Disease.

Receiver Operating Characteristics (ROC):

In addition to looking at model statistics, our team also assessed our model by splitting the dataset into two: one for training and one for testing our model. Then, to test our model's effectiveness at predicting CHD, we examined the area under the ROC curve, which yielded the predictive value of 85.8%, specificity of 82.4%, and sensitivity of 89.3% (**Fig. 6**). This result once again reaffirms the effectiveness of our model.

Comparison with Penalized Logistic Regression:

To further examine the predictive ability using our current dataset, a penalized logistic regression model was built using the lasso method in the glmnet library to select variables. Using area under the ROC curve as the evaluation metric, a tuning parameter value of 0.001373 yielded the best model with mean area under ROC curve of 0.8965 (**Fig. 7**). The existing model was opted for because the Penalized Logistic Regression model was only a marginal improvement to the existing model, and the existing model has fewer variables in the model, hence easier to interpret.

Interactive Online Application:

Given the accuracy of our model, an interactive web application of our model was created in order to make it widely accessible to the general population. The interactive application takes a patient's health indices as input and outputs the predicted probability, and 95% confidence interval of the probability, of that patient getting CHD (**Fig. 8**). Our interactive application can be found online at: **link to app**.

Discussion/Conclusions

Age, sex, type of chest pain, cholesterol levels, fasting blood sugar levels, presence of exercise-induced angina, and ST slope were important factors in the prediction of coronary heart disease. Age is highly correlated with the presence of coronary heart disease because the buildup of fats along artery walls and the hardening of blood vessels over time increase the likelihood of heart disease. Males are more likely to develop heart disease than females and are correlated with higher rates of coronary heart disease. Current literature suggests that this is because men have worse coping mechanisms (physiologically, behaviorally, emotionally) that lead to reduced adaptability to stressful situations as compared to females, increasing their risk for CHD (Weidner 2000). Additionally, chest pain is highly correlated with heart disease, and is often used as an early indicator of cardiovascular problems, even when other symptoms have not yet manifested. Cholesterol level is also correlated with CHD because cholesterol is responsible for the hardening and narrowing of arteries, which may cause hypertension. High levels of fasting blood sugar levels were highly correlated with CHD because high levels of sugars and fats in the bloodstream promote oxidative stress, which often damages blood vessels. This may impair proper function of myocardial tissues and lead to the onset of CHD. Exercise-induced angina is a highly correlated predictor of coronary heart disease, as well. Pain due to angina is primarily caused by poor blood flow to the heart, often related to buildups of thick fatty deposits or plaques that narrow arteries. This may restrict blood supply to the heart muscle and induce heart disease. Finally, ST Slope is highly correlated with coronary heart disease because it indicates the presence of damaged cardiovascular tissue.

The limitations of this study are the scope of the population, size of the dataset, and the data collection methods. Because the sample was drawn from the four medical centers, the results can only be generalized to patients of those medical centers. In order to obtain results that are applicable to more individuals, a representative sample of the larger population would be needed. The dataset was limited to approximately 913 patients, which is relatively small in comparison to the number of heart disease cases present throughout the United States. Additionally, data was collected primarily from developed countries, where more advanced healthcare systems and better environmental conditions exist to support both healthy and ill populations. These health networks may look significantly different in developing nations, where diagnosis, treatment, safe working conditions, and clean water, to name a few, may not be readily accessible. Thus, our data may not be representative of all populations. A way to improve our study would be to broaden our sample size to include more diverse populations from different geographic and socioeconomic backgrounds. Heart disease is the leading cause of death for people living in the United States, and is an illness that afflicts the families and friends of many Americans today. More research, early prediction, and treatment of heart disease are all crucial for the prevention of deaths and care for aging populations.

References

1. American Heart Association. “What Is Cardiovascular Disease?” Www.heart.org, American Heart Association, 31 May 2017, www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease.

2. Centers for Disease Control and Prevention. “Heart Disease Facts.” Centers for Disease Control and Prevention, 14 Oct. 2022, www.cdc.gov/heartdisease/facts.htm.

3. National Heart, Lung, and Blood Institute. “Coronary Heart Disease - What Is Coronary Heart Disease? | NHLBI, NIH.” Www.nhlbi.nih.gov, 24 Mar. 2022, www.nhlbi.nih.gov/health/coronary-heart-disease.

4. UCI Machine Learning Repository. “Heart Disease Data Set,” https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

5. UCSF Health. “Diagnosing Heart Disease.” *Ucsfhealth.org*, UCSF Health, 24 June 2022, https://www.ucsfhealth.org/education/diagnosing-heart-disease.

6. Weidner G. (2000). Why do men get more heart disease than women? An international perspective. Journal of American college health : J of ACH, 48(6), 291–294.

Appendix:

Fig. 1: Side-by-side boxplots for exploratory data analysis

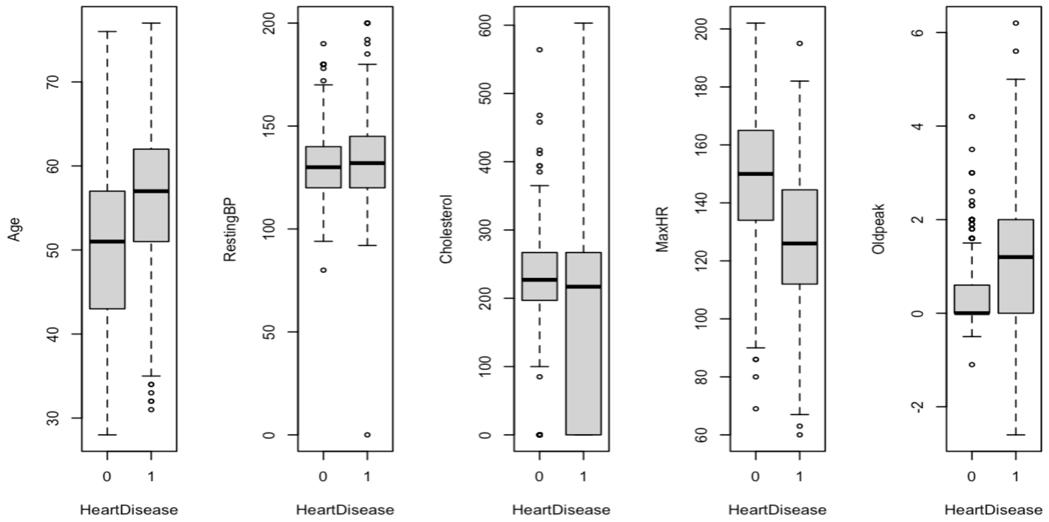


Fig. 2: Mosaic plots for exploratory data analysis

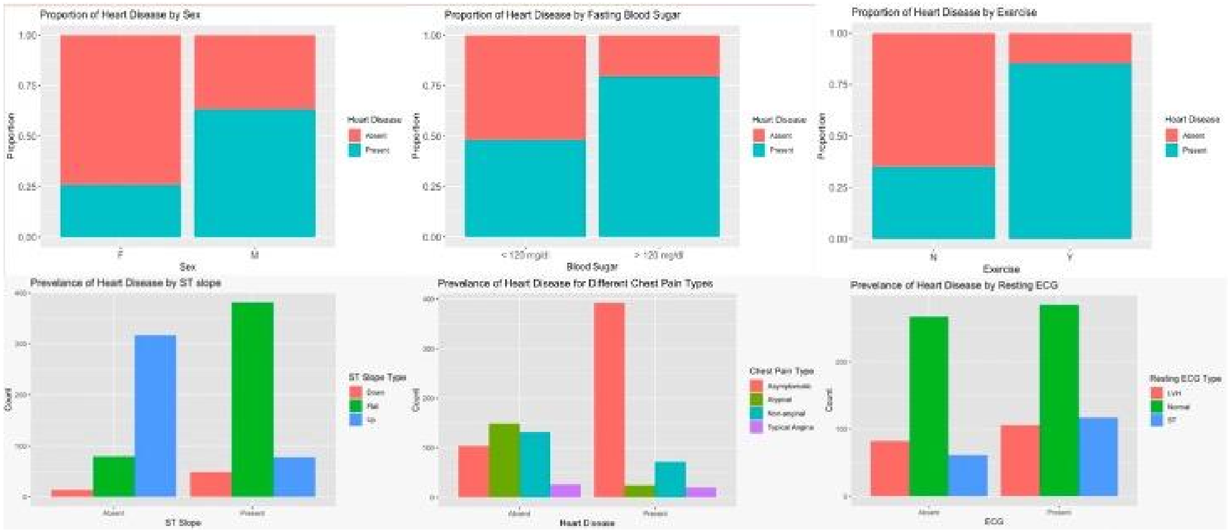


Fig. 3: Empirical logit plot

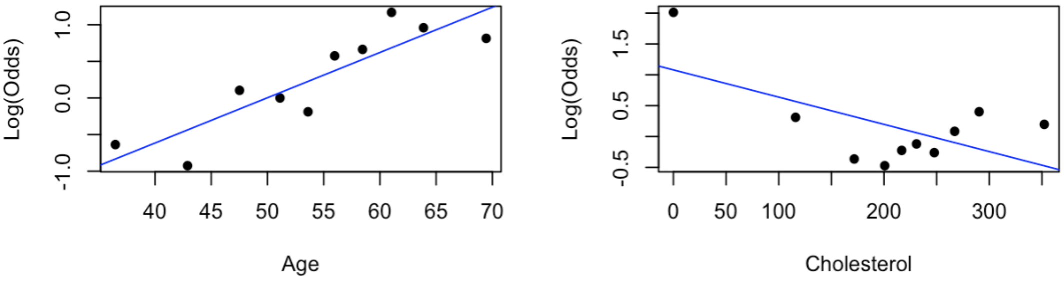


Fig. 4: Collinearity plot

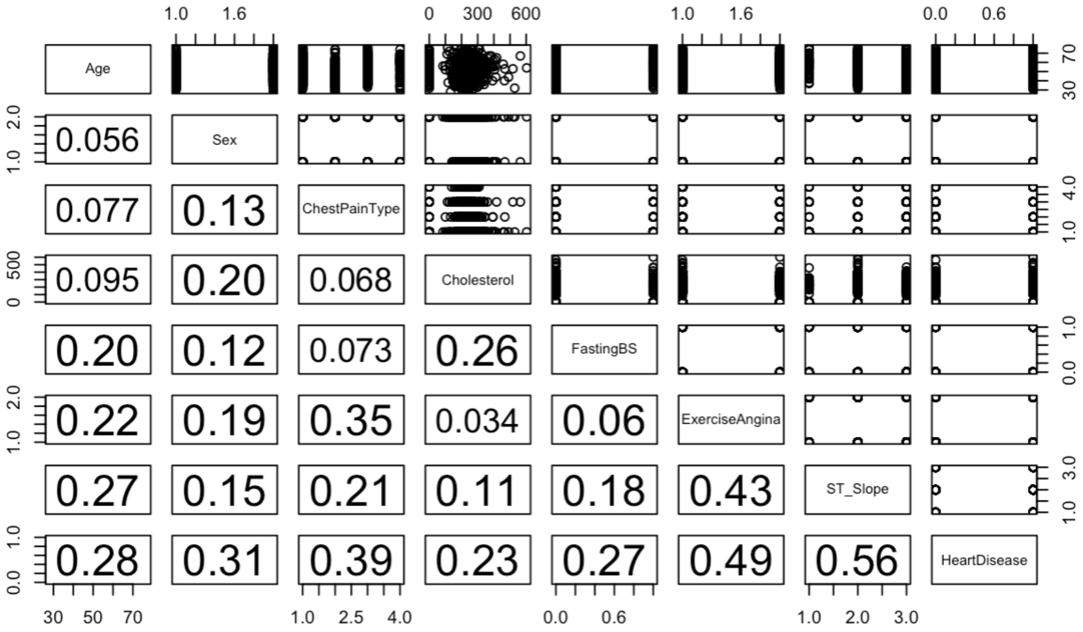


Fig. 5: Confusion Matrix

Predicted Y	**0**	**1**
0	340	55
1	70	453

Fig. 6: ROC Curve of Logistic Regression Model

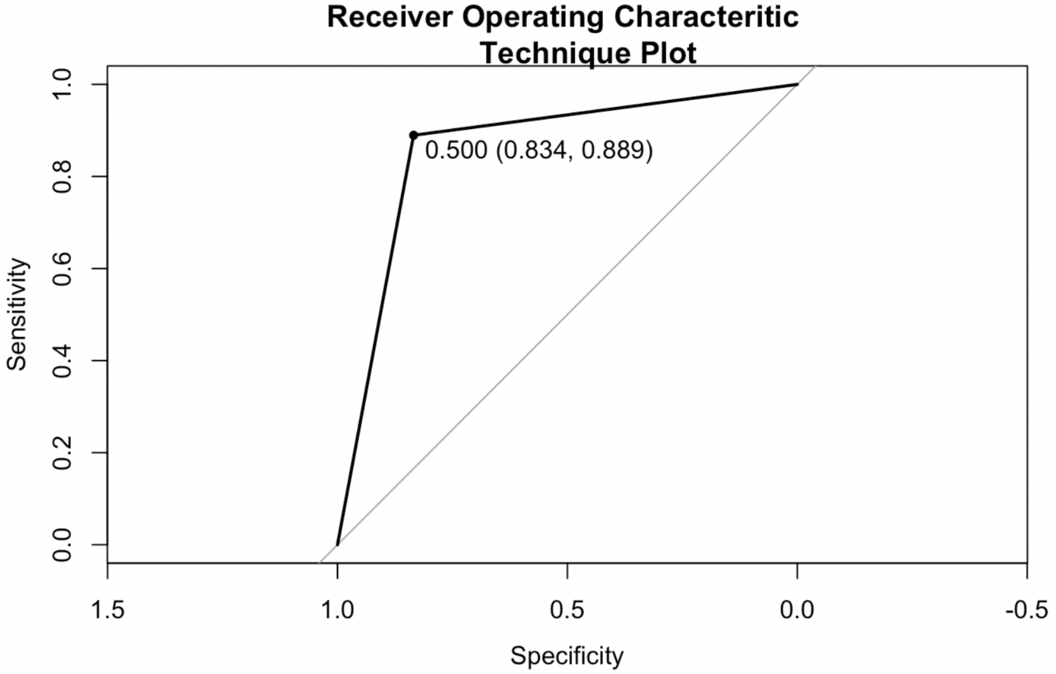
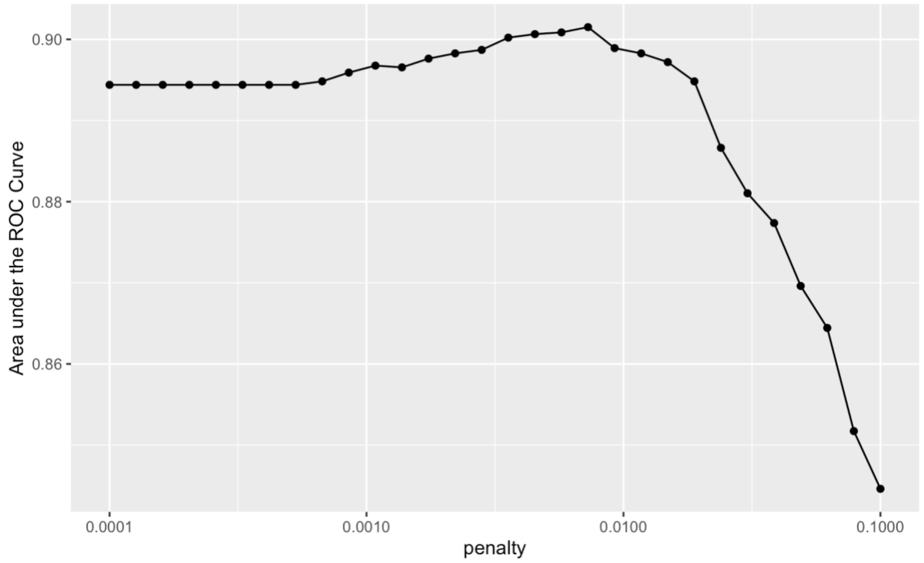
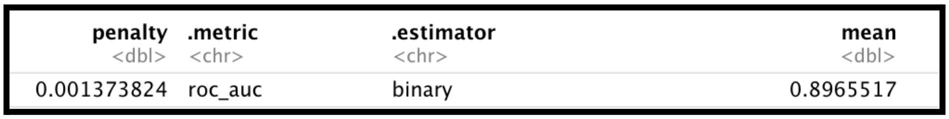


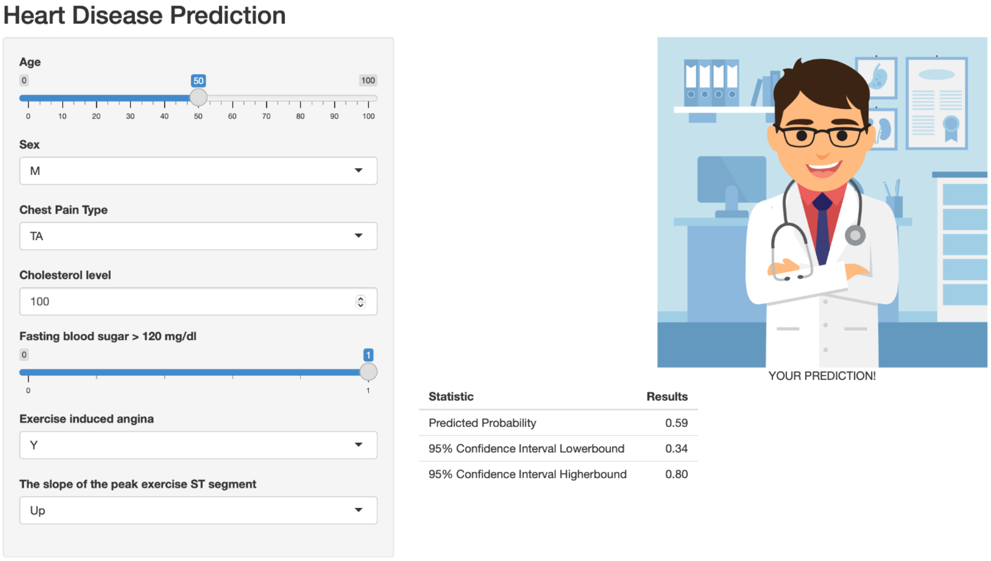
Fig. 7: Performance of Penalized Logistic Regression Model:





penalty <dbl>	.metric <chr>	.estimator <chr>	mean <dbl>
0.001373824	roc_auc	binary	0.8965517

Fig. 8: Interactive Online Application



Heart Disease Prediction

Age: 50

Sex: M

Chest Pain Type: TA

Cholesterol level: 100

Fasting blood sugar > 120 mg/dl: 1

Exercise induced angina: Y

The slope of the peak exercise ST segment: Up

Statistic	Results
Predicted Probability	0.59
95% Confidence Interval Lowerbound	0.34
95% Confidence Interval Higherbound	0.80

Table 1: Variables in Final Model

Variable	Estimate	Standard Error	P-value
Age	0.028509	0.011653	0.014422
Sex (Male)	1.4333221	0.274189	1.72×10^{-7}
Cholesterol	-0.003775	0.001033	0.000259
Fasting Blood Sugar	1.120845	0.272424	3.88×10^{-5}
Exercise Angina	1.054476	0.232360	5.68×10^{-6}
ST Slope Flat	1.039862	0.406439	0.010513
ST Slope Up	-1.625727	0.409904	7.31×10^{-5}
Chest Pain, Type TA	-1.420902	0.427256	0.000882
Chest Pain, Type NAP	-1.742923	0.260756	2.32×10^{-11}
Chest Pain, Type ATA	-1.941541	0.317548	9.71×10^{-10}