Project 1 Writeup

Ally Wardell

12/01/2021

Introduction

According to the CDC, the leading cause of death in the United States is heart disease, followed by cancer (2021). Due to this fact, heart disease has been prevalent in research, and should be continually examined, so researchers can gain insight into how to better diagnose and treat various heart diseases. The inter-workings of the heart are complicated, yet intriguing. As computational power has progressively developed, techniques to analyze data have progressed, as well as methodologies to use data to predict disease. Specifically, machine learning and various prediction algorithms have proven to be instrumental in cardiology advancements. (Cuocolo et al., 2019). Random forest algorithms have shown to be useful in prediction of cardiovascular disease using variables such as age, sex, chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercised induced angina, major vessels, and thalessemia. Additionally, support vector machines have been used to effectively classify cardiac arrhythmias using a heart rate variability signal (Asl et al., 2008).

Dataset Description

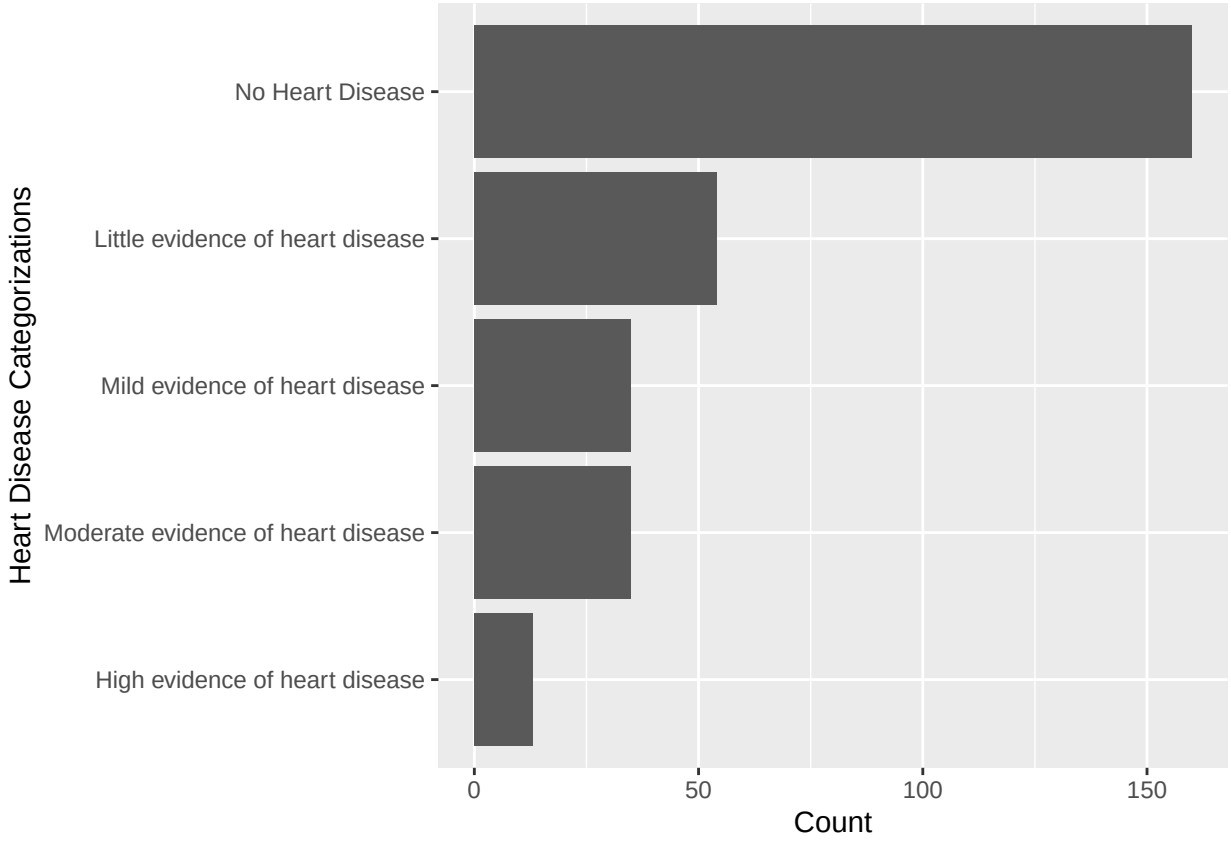
The data utilized contains 297 observations. Each observation is an individual subject, accompanied by the measurements of the covariates in the aforementioned paragraph. Consider Table 1 below. The mean age for those who have no presence of heart disease is slightly lower than the mean age for the group with heart disease present (53 vs. 57 years). Additionally, the mean resting blood pressure (129 mmHg vs. 135 mmHg) and cholesterol (243mg vs. 252mg) lower in the group of subjects who did not have any presence of heart disease detected. Conversely, the mean maximum heart rate was higher in the group who had no evidence of heart disease (259bpm vs. 139bpm).

Summary Statistics for Heart Failure Covariates				
Characteristic	**N**	**0, N = 160[1]**	**1, N = 137[1]**	**p-value[2]**
...1	297	145 (86)	155 (89)	0.4
Age	297	53 (10)	57 (8)	<0.001
Sex (M/F)	297	89 (56%)	112 (82%)	<0.001
Chest Pain	297			<0.001
1		16 (10%)	7 (5.1%)	
2		40 (25%)	9 (6.6%)	
3		65 (41%)	18 (13%)	
4		39 (24%)	103 (75%)	
Resting Blood Pressure	297	129 (16)	135 (19)	0.008
Cholesterol	297	243 (54)	252 (50)	0.2
Fasting Blood Sugar	297	23 (14%)	20 (15%)	>0.9
Resting ECG	297			0.008
0		92 (57%)	55 (40%)	
1		1 (0.6%)	3 (2.2%)	
2		67 (42%)	79 (58%)	
Maximum Heart Rate	297	159 (19)	139 (23)	<0.001
Exercise Induced Angina	297	23 (14%)	74 (54%)	<0.001
Depression	297	0.60 (0.79)	1.59 (1.31)	<0.001
Slope	297			<0.001
1		103 (64%)	36 (26%)	
2		48 (30%)	89 (65%)	
3		9 (5.6%)	12 (8.8%)	
Major Vessels	297			<0.001
0		129 (81%)	45 (33%)	
1		21 (13%)	44 (32%)	
2		7 (4.4%)	31 (23%)	
3		3 (1.9%)	17 (12%)	
Thalessemia	297			<0.001
3		127 (79%)	37 (27%)	
6		6 (3.8%)	12 (8.8%)	
7		27 (17%)	88 (64%)	

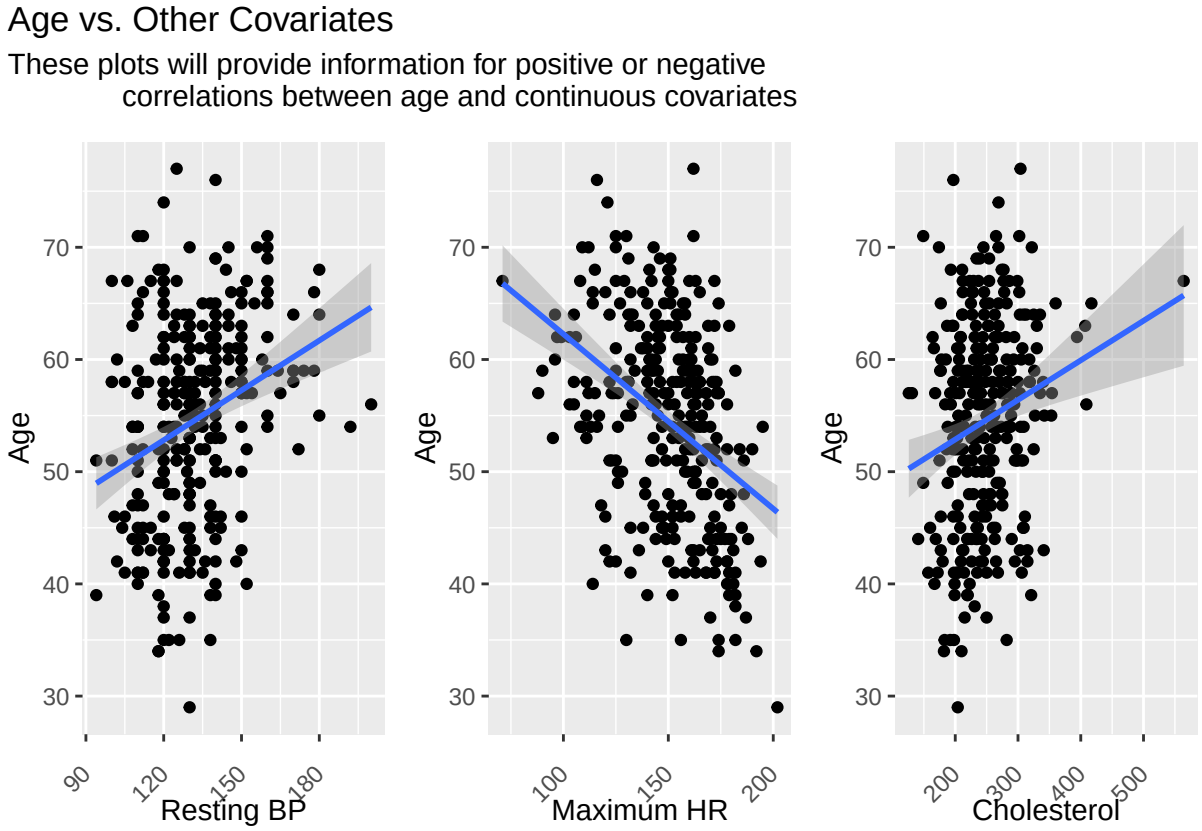
[1]Mean (SD); n (%)

[2]One-way ANOVA; Pearson's Chi-squared test

To further explore the data, Figure 2 depicts that as the severity of the categorization of heart disesase becomes more severe, the number of individuals representing these situations in the sample decrease.



It may be of interest to consider potential relationships of continuous predictors; such as, age and resting blood pressure.



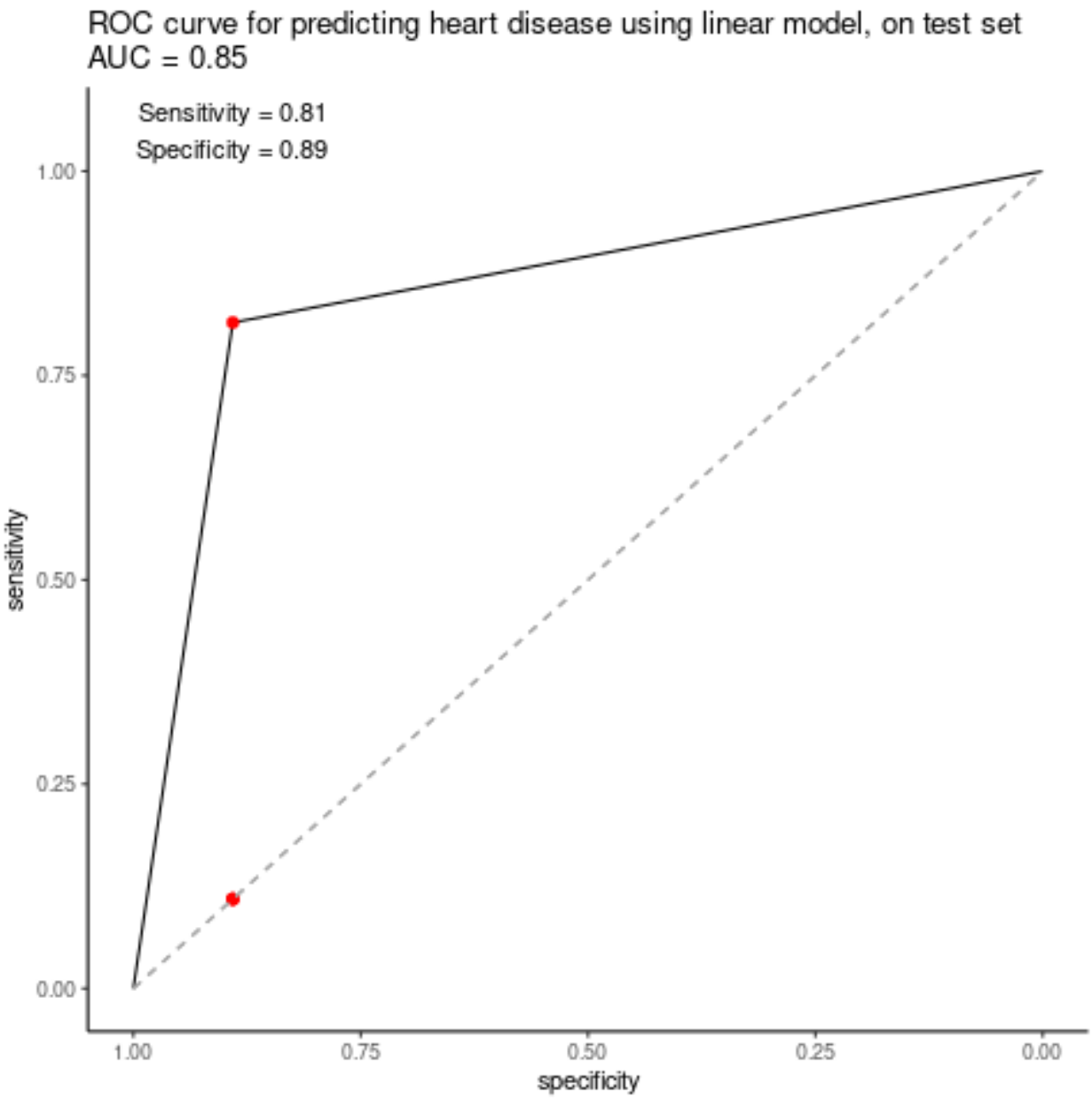
Finally, view the shiny app for visualizations regarding the data on the covariates, faceted by the 5 levels of heart disease presence.

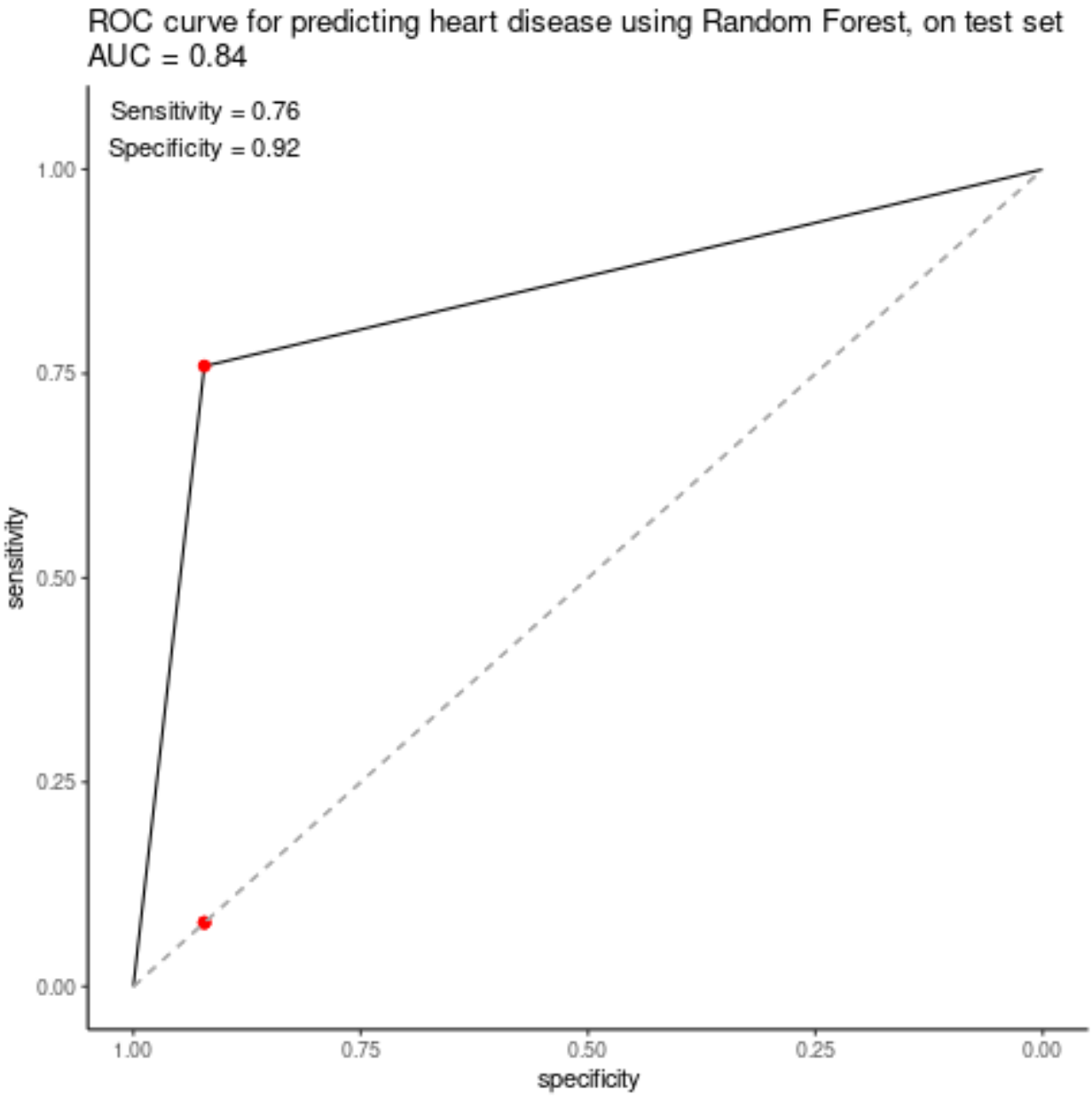
Machine Learning Modeling Methods: Initial ROC Visulaization

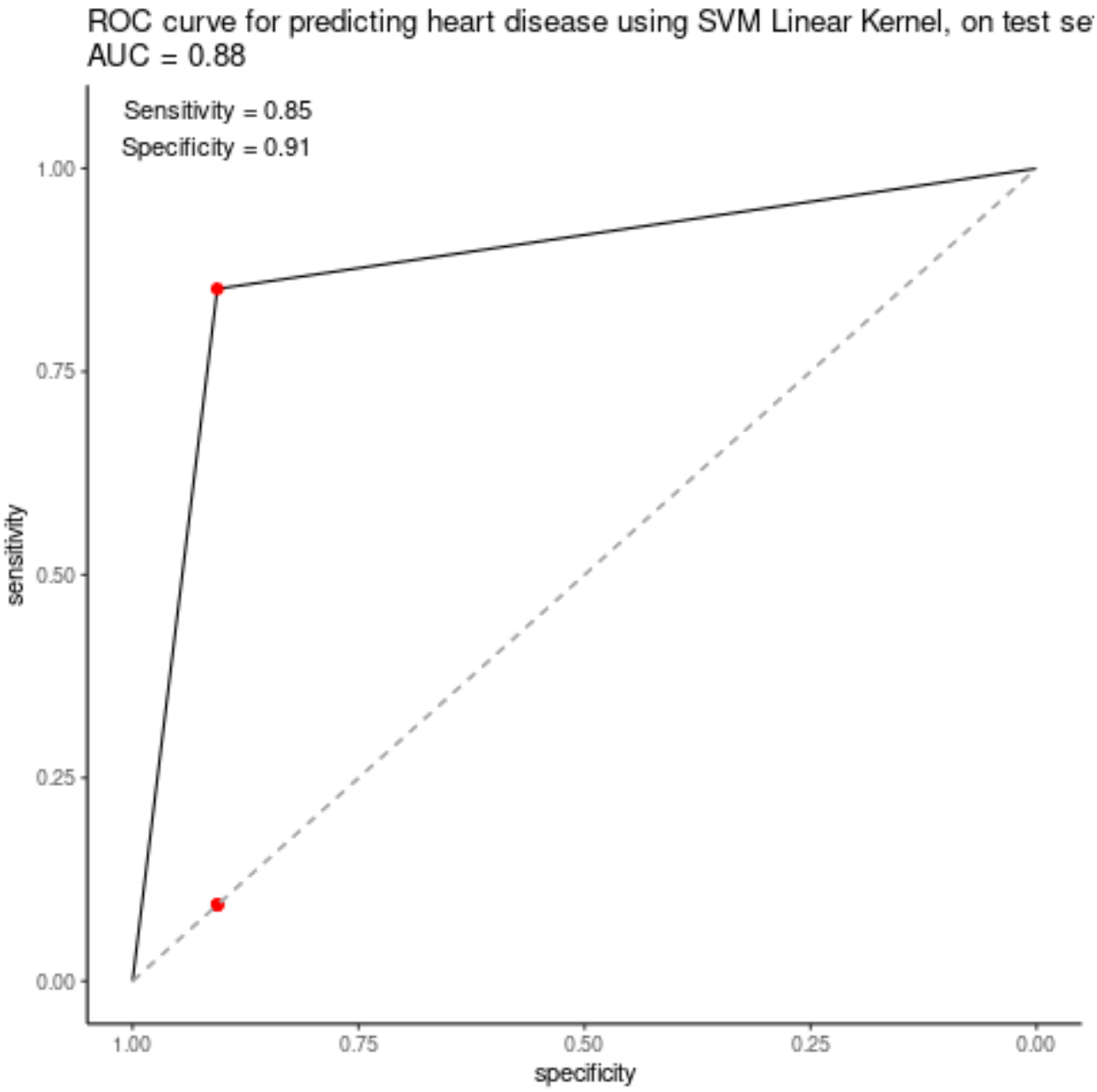
A random forest model (RF) was compared with two support vector machine models (SVM) and a linear model, with the goal of selecting the model with the better performance. To select the model with the best performance, observations forming the training and testing sets were chosen randomly. The random selection of participants to form training and testing sets will assist in obtaining better error estimation for categorizing the individuals into the two categories presence of heart disease vs no presence of heart disease detected.

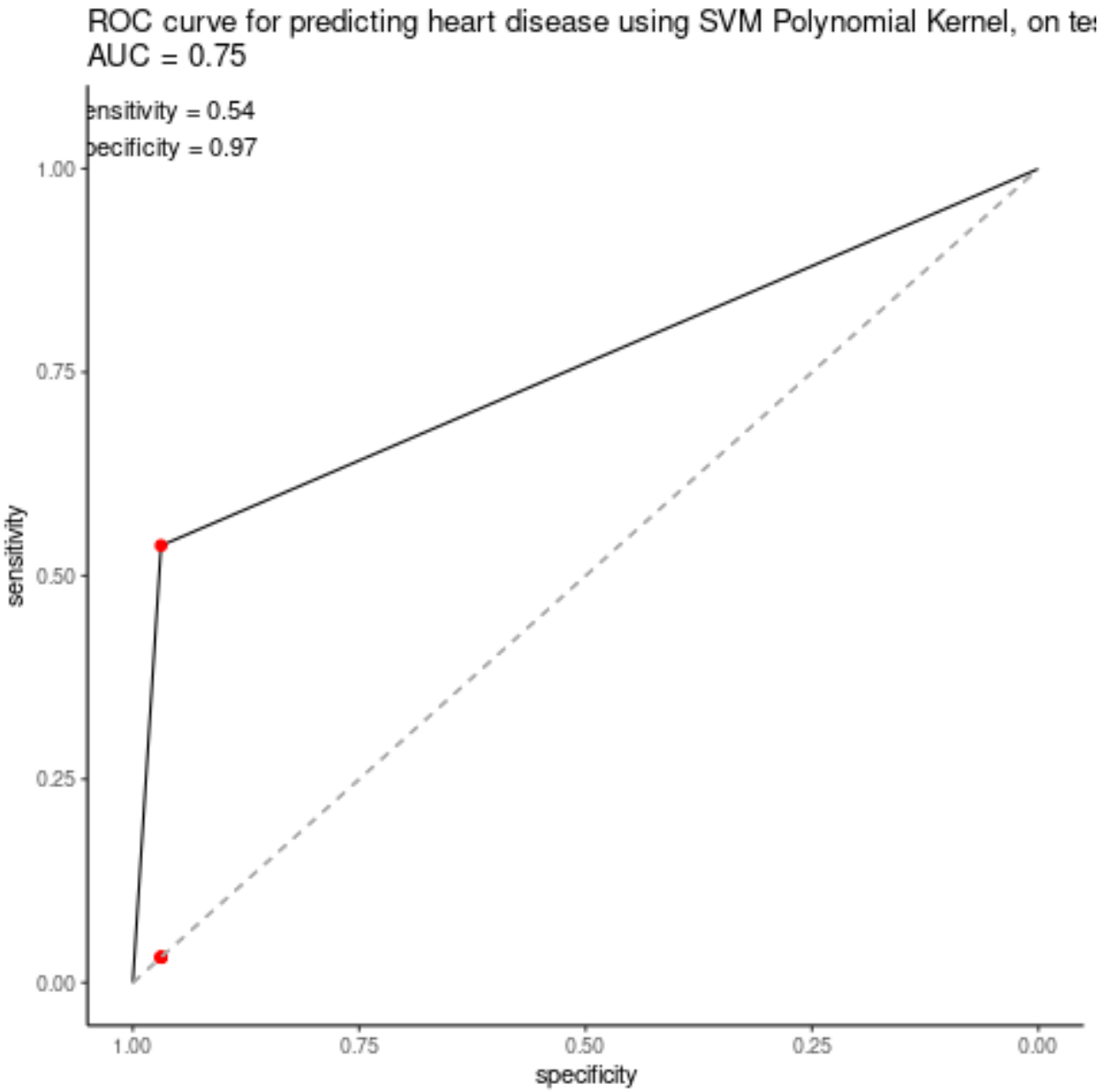
The random forest model utilized a grid search technique to select the best parameter values. The grid search considered 50, 250, and 500 trees in combination with 4, 7, and 13 predictors at the random forest splits. The best tuning parameters for a given training set were selected using the out-of-bag mean squared error in the training set.

The ROC Curves for the various machine learning methods are below:


ROC curve for predicting heart disease using linear model, on test set
AUC = 0.85
Sensitivity = 0.81
Specificity = 0.89
1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity


ROC curve for predicting heart disease using Random Forest, on test set
AUC = 0.84
Sensitivity = 0.76
Specificity = 0.92
1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity


ROC curve for predicting heart disease using SVM Linear Kernel, on test se
AUC = 0.88
Sensitivity = 0.85
Specificity = 0.91
1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity



Machine Learning Modeling Methods:Cross-Validation

Additionally, the aforementioned algorithms were re-run with five-fold cross validation to train and test the RF algorithm and subsequently produced corresponding CV errors and standard errors.

The support vector machine models were trained with a linear kernel and a radial basis kernel, respectively. The grid of parameters that were tested included the epsilon parameter with values 0, 0.25, 0.50, 0.75, and 1.00, coupled with the cost parameter with values of 1 to 5. Five-fold cross validation was utilized to train and test the SVM algorithm and subsequently produced corresponding CV errors and standard errors. The main goal was to obtain a predictive model with the best cross validation errors and cross validation standard errors. All model tuning and testing was completed in R version 4.0.3 (R Core Team, 2020).

Results

For analysis, I compare the random forest method, the support vector machine method with a linear and polynomial kernel (respectively), and a regression method to determine which algorithm most accurately predicts heart disease.

The support vector machine (SVM) model with the polynomial kernel performed the best with a 91.9% prediction accuracy for the no heart disease detected indication. This machine learning model was followed by the support vector machine (SVM) model with the linear kernel (accuracy = 86.2%), the linear model (accuracy = 85.6%), and lastly the random forest model (accuracy = 85.0%).

When trying to accurately predict the presence of heart disease, the linear model performed the best (accuracy = 80.3%). This model was followed by the support vector machine model with the linear kernel (accuracy = 79.6%), the random forest model (accuracy = 77.4%), and the support vector machine model with the polynomial kernel (accuracy = 71.6%).

The support vector machine model with the polynomial kernel performed the best for predicting no heart disease but the worst for predicting the presence of heart disease accurately. Practically, the main goal would be to predict the presence of heart disease as closely to reality as possible, thus making the linear model or the support vector machine model with the linear kernel more practical in this situation.

Random Forest results:

Random Forest Sensitivity/specificty		
Heart.Disease.Category	CV.sensitivity.specificity	CV.sensitivity.specificity.SE
0	0.8687500	0.05134899
1	0.7812169	0.08430192

SVM with linear kernel results:

Support Vector Machine with a Linear Kernel and Linear Model /nSensitivity/Specificity		
Method	CV.sensitivity.specificity	CV.sensitivity.specificity.SE
svm	0.8625000	0.07194290
svm	0.7962963	0.06343692
lm	0.8562500	0.02795085
lm	0.8031746	0.03972000

SVM with polynomial kernel results:

Support Vector Machine with a Polynomial Kernel and Linear Model /nSensitivity/Specificity		
Method	CV.sensitivity.specificity	CV.sensitivity.specificity.SE
svm	0.9187500	0.02795085
svm	0.7156085	0.04438614
lm	0.8562500	0.02795085
lm	0.8031746	0.03972000

Concluding Remarks

All of the possible variables are all easily attained through blood tests or a cardiology report. The purpose of this study would be to accurately predict mortality caused by heart failure. If an accurate prediction algorithm can be achieved by using easily attainable measures, the algorithm could be useful as an assistance mechanism for physicians to consult when considering a treatment plan. Additionally, physicians may be able to use the algorithm to take additional precautionary measures for patients who have a higher risk of death from heart disease.

The next step of this project would be to make plots for the ROC curves for the cross validation steps.

References

Asl , B. M., Mohebbi, M., & Setarehdan, S. K. (2008, June). Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artificial intelligence in medicine. https://pubmed.ncbi.nlm.nih.gov/18585905/).

Centers for Disease Control and Prevention. (2021, April 9). FastStats - deaths and mortality. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/fastats/deaths.htm.

Cuocolo, R., Perillo, T., De Rosa, E., Ugga, L., & Petretta, M. (2019, August). Current applications of big data and machine learning in cardiology. Journal of geriatric cardiology : JGC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748901/).

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna,

Su, X., Xu, Y., Tan, Z., Wang, X., Yang, P., Su, Y., Jiang, Y., Qin, S., & Shang, L. (2020, September). Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model. Journal of clinical laboratory analysis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7521325/.