PFIZER BAI HEART DISEASE CASE STUDY

Charles Kelley

July 2023

INSTRUCTIONS

Objectives

We are interested in understanding what are the main contributing factors toward heart disease?

A. Use a predictive model to identify the main contributing factors towards heart disease

B. Put together a mock-up visual to share the results of your analysis. Be prepared to discuss which model(s) you used and why

Please put together your answers (slides) and any supporting code or diagrams/visualizations or however else you feel would be the most appropriate way to express your answers.

Please e-mail this to us.

Disclaimer

This analysis was time limited to 24-48 hours. There is only time to try so many approaches and document them for presentation in that short a period. As such, simplicity was favored over complexity and the research and planning phase of the analysis was shortened.

STUDY STEPS

1) Exploratory Data Analysis

Examination, visualization, and summarization of the data to understand its main characteristics and underlying structures before performing formal analysis

2) Problem Framing

Defining and establishing the boundaries and context of analytical problem to be solved and identification of stakeholders, resources, constraints, and the scope of outcomes

3) Contextual Research

Gathering and analyzing key information in the business and technical problem domains to inform the design and development of the problem solution

4) Solution Development

Data preparation, documentation, feature selection, model development, performance assessment, and deployment or reporting pipeline

5) Solution Selection & Delivery

Communicating results in a comprehensible way to stakeholders, and possibly deploying the solution to production while ensuring continuous monitoring and adjustments based on feedback

I) EXPLORATORY DATA ANALYSIS

Heart Disease Patient Dataset

Overview

Number of variables	14
Number of numeric variables	5
Number of categorical variables	9
Number of observations	303
Missing cells	0
Duplicate rows	1

Notable Variable Characteristics

- **male_binary** - has a high imbalance toward male sex with 207 observations indicated as male and 96 as female
- **exercise_st_depression** - has an extreme positive skew with most values being zero but with range of [0.0 - 6.2] and a skewness statistic of ~1.27 and kurtosis of ~1.57

All numeric variables aside from exercise_st_depression have a relatively normal distribution with any moderate positive or negative skewness due to one or a few outlier observations

Variable Correlation

		heart_disease_binary
01	heart_disease_binary	1.000
02	**thalassemia**	**0.522**
03	**chest_pain_type**	**0.510**
04	**major_vessels_colored**	**0.483**
05	**exercise_st_depression***	**0.433**
06	**execise_angina_binary**	**0.427**
07	**max_heart_rate***	**0.406**
08	**exercise_st_slope**	**0.388**
09	male_binary	0.268
10	age*	0.260
11	rest_ecg_results	0.163
12	serum_cholesterol*	0.061
13	rest_blood_pressure*	0.000
14	fast_bg_above120_binary	0.000

* Trailing asterisk indicates numeric variable

Bold text indicates variable has a moderate to high correlation with heart disease diagnoses where the threshold of 0.3 was selected based on common norms.

2) PROBLEM FRAMING

Objectives

Model Task - Binary Classification

Identify the contributing factors that are associated with a heart disease diagnosis.

Model Selection - Classifier w/Variable Importance

A binary classification model where the target is the `heart_disease_binary` variable and all other dataset variables can be assessed as for their importance in predicting whether a patient diagnosed with heart disease—most decision trees and regression based models support variable importance.

Model Interpretation - Variable Importance

The key assumption to this analysis is that a variable's contributions to classification model are separable and can be used to relatively rank their contribution to the heart disease diagnosis classification

Key Considerations

Model Explainability

Can we identify the variables that most associated with a heart disease diagnosis?

Time Constraint

The hard cap of 24-48 hours to complete this analysis limits the number of solution options that can be investigated, and limits the options to more traditional solutions that are simple to implement in practice.

3) CONTEXTUAL RESEARCH

Heart Disease

Developing a heart disease binary classification model necessitates understanding the medical domain, including heart disease:

- Risk factors
- Symptoms
- Diagnostic criteria
- Clinical data structures

This would involve gathering data from sources like electronic health records, lifestyle and genetics data, and medical literature, and collaborating with healthcare professionals for their expert insights.

Notable Literature

The following paper discusses the creation of a classification system for heart disease diagnosis and uses variables very similar to our heart disease patient dataset

- Heart disA Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method

Classification Methods

In addition to understanding the heart disease domain, a good heart disease predictive model requires a deep understanding of classification models including:

- Model assumptions
- Performance metrics
- Model variety - such as logistic regression, decision trees, or support vector machines

It would also be best to know which model types have been most successful in similar problem domains or contexts, along with an understanding of the ethical considerations in health-related predictions.

Notable Literature

The following paper is helpful in deciding what types of classification modeling methods to use—dimensionality reduction for classification in this case

- Comparison of PCA and RFE-RF Algorithm in Bankruptcy Prediction

4A) SOLUTION DEVELOPMENT

Recursive Feature Elimination

Recursive feature engineering (RFE) is a method used in machine learning where a model is trained on a dataset, the importance of each feature is determined, and the least important features are progressively eliminated in subsequent iterations of model training.

RFE as A Wrapper Algorithm

Different machine learning algorithms can be used in the core of the method and wrapped by the RFE algorithm and used to help select features.

This supports both the evaluation of classification models and feature importance simultaneously—perfect for our task of building a heart disease classifier to understand variable contribution.

RFE Algorithms Assessed for Feature Selection

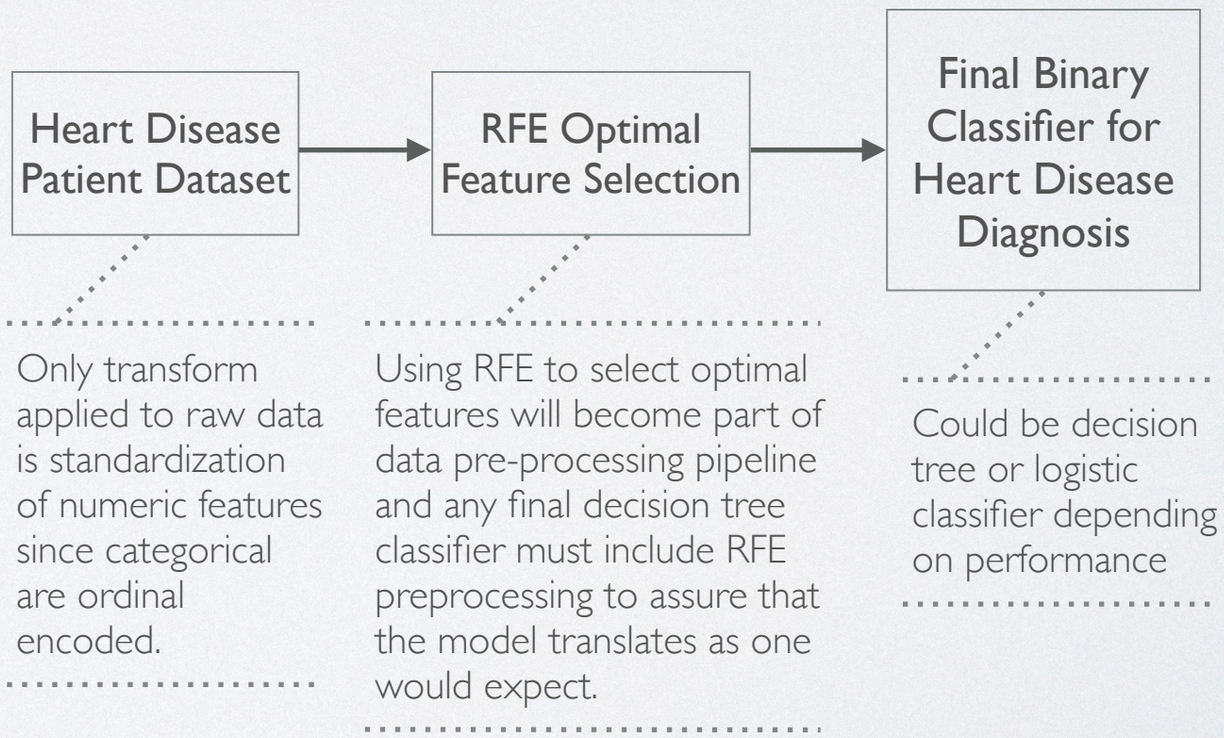
- Logistic Regression (logistic)
- Perceptron
- Classification and Regression Tree (cart)
- Random Forest (forest)
- Gradient Boost Machine (gbm)

Binary Classifier Selection

Decision tree and logistic regression classifiers are good candidates for a final binary classification model due to their simplicity and explainability

The better model based on the performance and explainability requirements should be used. Generally speaking decision trees are explainable more straightforward to explain, but they often favor variance over bias and could be lest translatable.

Full Model Pipeline



4B) SOLUTION DEVELOPMENT

Model Pipeline Feature Assessment

	logistic	perceptron	cart	forest	gbm
age	6	5	✅	✅	✅
exercise_st_depression	✅	✅	✅	✅	✅
rest_blood_pressure	4	2	✅	✅	✅
serum_cholesterol	5	3	✅	✅	✅
max_heart_rate	3	✅	✅	✅	✅
male_binary	✅	✅	✅	✅	✅
chest_pain_type	✅	✅	✅	✅	✅
fast_bg_above120_binary	7	✅	3	✅	2
rest_ecg_results	2	4	2	✅	✅
execise_angina_binary	✅	✅	✅	✅	✅
exercise_st_slope	✅	✅	✅	✅	✅
major_vessels_colored	✅	✅	✅	✅	✅
thalassemia	✅	✅	✅	✅	✅
Features Selected	**7**	**9**	**11**	**13**	**12**

✅ = Variable Selected; Else # = Variable Importance Rank

RFE Feature Selection

Fewer features usually leads to better explainability if accuracy doesn't suffer. In this case, using a logistic regression model as the core model of the RFE algorithm leads to the fewest features in the final model pipeline.

Note On Model Tuning

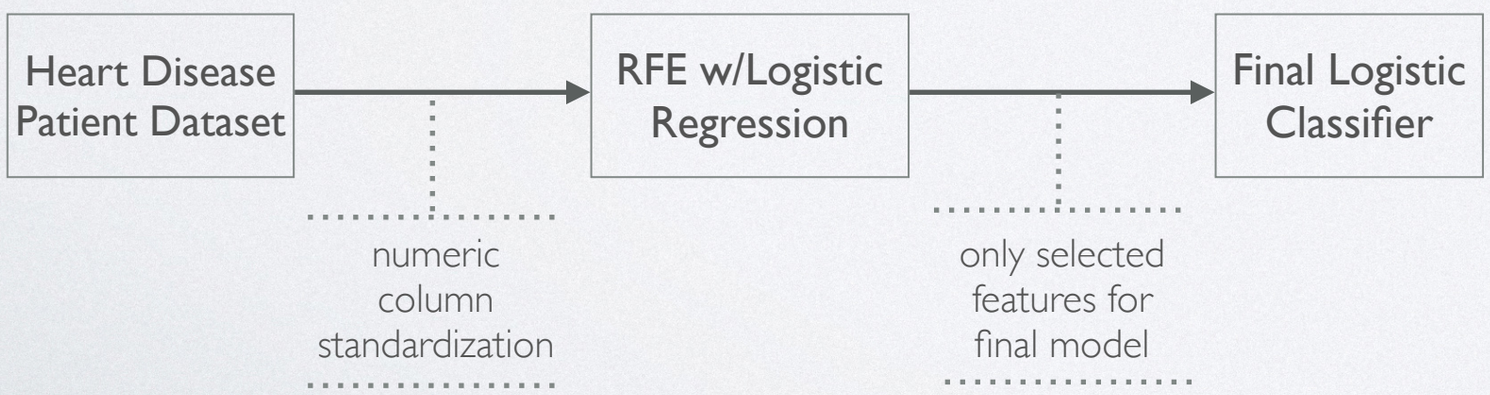
The default versions of all of these models were used. There may be considerable room for improvement via iteration and hyperparameter tuning.

5) SOLUTION SELECTION & DELIVERY

Full Model Pipeline For Heart Disease Classifier

		RFE Core Model				
		logistic	perceptron	cart	forest	gbm
RFE Core Model Features Selected		**7**	9	11	13	12
Decision Tree Final Classifier	Pipeline Accuracy Mean	0.770	0.744	0.734	0.750	0.746
	Pipeline Accuracy Standard Deviation	0.055	0.068	0.068	0.059	0.057
Logistic Regression Final Classifier	Pipeline Accuracy Mean	0.816	0.820	0.809	**0.821**	0.814
	Pipeline Accuracy Standard Deviation	0.040	**0.031**	0.043	0.039	0.041

Selected Full Model Pipeline



Pipeline Selection

Starting from a baseline of the logistic regression as the core for RFE since it uses the least features, we see that the logistic regression final classifier also performs better than the decision tree.

Additionally, changing the RFE core model to anything besides logistic regression does not provide a large lift in model accuracy. As such, logistic regression based RFE and final classification is the preferred pipeline

Note: Model pipeline accuracy was assessed using 5 fold cross-validation with the accuracy mean and standard deviation derived from the combined series of scores.

Final Logistic Classifier Model

FEATURE	COEFF
exercise_st_depression	-0.7166
male_binary	-1.1630
chest_pain_type	0.8423
execise_angina_binary	-1.04280
exercise_st_slope	0.6698
major_vessels_colored	-0.7783
thalassemia	-0.8149
intercept	1.9831

REFERENCE: PATIENT DATA DICTIONARY

Column map of variable names in provided raw patient heart disease dataset
along with alias names used for analysis

VARIABLE	ALIAS	DESCRIPTION
age	age	age in years
sex	male_binary	(1 = male; 0 = female)
cp	chest_pain_type	chest pain type
trestbps	rest_blood_pressure	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum_cholesterol	serum cholesterol in mg/dl
fbs	fast_bg_above120_binary	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	rest_ecg_results	resting electrocardiographic results
thalach	max_heart_rate	maximum heart rate achieved
exang	execise_angina_binary	exercise induced angina (1 = yes; 0 = no)
oldpeak	exercise_st_depression	ST depression induced by exercise relative to rest
slope	exercise_st_slope	the slope of the peak exercise ST segment
ca	major_vessels_colored	number of major vessels (0-3) colored by flouroscopy
thal	thalassemia	3 = normal; 2 = fixed defect; 1 = reversable defect
target	heart_disease_binary	Flag if the patient has heart disease (1) or not (0)