KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Predicting the Probability of Heart Disease… A Comprehensive Guide to Model Building

Greg Balkcom (Master's), Dr. MinJae Woo (Instructor, Python Class)

INTRODUCTION

- Heart disease is the leading cause of death in the United States.
- Identifying heart disease early gives you the best chance of managing it well.
- Predictive models can help identify potential risk factors.
- This data set is combined from 5 different hospitals:
  - Cleveland: 303 observations
  - Hungarian: 294 observations
  - Switzerland: 123 observations
  - Long Beach VA: 200 observations
  - Statlog (Heart) Data Set: 270 observations
- Final dataset: 918 observations

Objectives:

1. Build predictive models for heart disease
2. Evaluate model performance
3. Determine important factors in predicting heart disease

METHODS

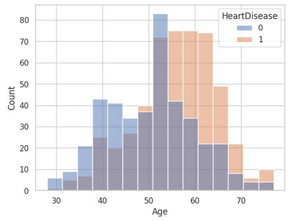
- Data set from www.kaggle.com
- 918 records with 11 predictive variables:
  - Age: Patient's age in years
  - Sex: Patient's gender
  - ChestPainType: Typical Angina, Atypical Angina, Non-Anginal Pain, or Asymptomatic
  - RestingBP: Resting Blood Pressure
  - Cholesterol: Blood Cholesterol Level
  - FastingBS: Fasting Blood Sugar (0 if ≤ 120, 1 if > 120)
  - RestingECG: Normal, ST (Irregular ST-T Wave), or LVH (Left Ventricular Hypertrophy).
  - MaxHR: Maximum Heart Rate
  - ExerciseAngina: Angina Induced by Exercise (Y or N)
  - Oldpeak: Numeric ST Value Measured in Depression
  - ST_Slope: Slope of Exercise ST Segment (Up, Flat, Down)
- Went through proper model building steps:
  1) Removed coded missing values, and imputed missing data using MICE imputation
  2) Visualized data to look for potential relationships
  3) Checked for multicollinearity with heatmap
  4) Normalized the data for machine learning models Used MinMax Scaler from SciKit Learn
  5) Divided data into train, validate, and test sections, and trained and validated the models:
  6) Tested performance of each model on test section and put results in a confusion matrix
  7) Evaluated model performance
     ROC: Area under the "Receiver Operating Curve"
     Accuracy: (True Pos + True Neg) / Total
     Precision: True Pos / (True Pos + False Pos)
     Recall: True Pos / (True Pos + False Neg)
     F1 Score: 2 * ((Prec*Rec) / (Prec+Rec))
  8) Identified important factors in predicting the probability of heart disease

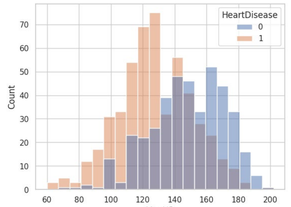
Steps in Model Building

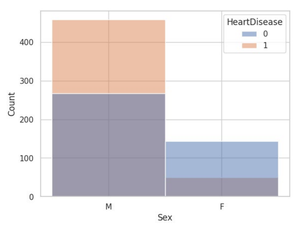
1) Asses Missingness and Impute new Values

Variable	Missing
Age	0
Sex	0
Chest Pain Type	0
Resting BP	1
Cholesterol	172
Fasting Blood Sugar	0
Resting ECG	0
Max Heart Rate	0
Exercise Angina	0
Oldpeak	0
ST Slope	0
Heart Disease	0

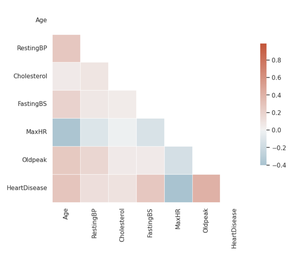
2) Explore Data and Look For Relationships





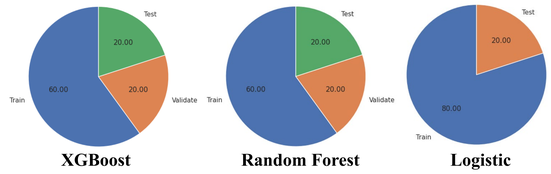


3) Check for Multicollinearity



4) Normalize the Data With SKLearn MinMax Scaler

5) Divide Data into Train, Validate, and Test Sections, then Build and Train Models

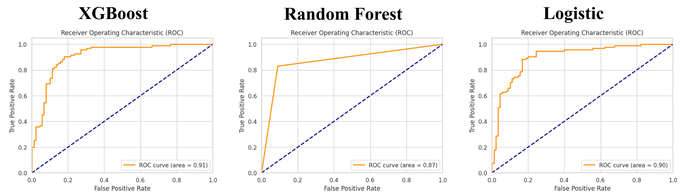


6) Test Final Models

Confusion Matrix By Model

XG Boost		Random Forest		Logistic Regression	
71	18	72	17	75	15
9	86	7	88	14	80

7) Evaluate Model Performance



Measure	Model: XG Boost	Random Forest	Logistic Regression
Accuracy	0.853	0.870	0.842
Precision	0.798	0.809	0.833
Recall	0.888	0.911	0.843
F1	0.840	0.857	0.838

For comparison, Mahmud et al. (2023) used a metamodel of Random Forest, Decision Tree, Gaussian Naive Bayes, and K-Nearest Neighbor with Accuracy of 0.87 on same data.

8) Identify Important Factors

Top 5 Important Variables by Model		
XG Boost	**Random Forest**	**Logistic Coefficient**
Cholesterol	ST_Slope_Up	Oldpeak
MaxHR	Sex_M	MaxHR
RestingBP	ChestPainType_NAP	ST_Slope_Up
Oldpeak	Oldpeak	ChestPainType_ATA
Age	RestingBP	ChestPainType_NAP

RESULTS

- Only two variables had coded missing values to be imputed: Cholesterol and Resting BP
- Missing values were < 20% of total, so no issues imputing.
- There were no strong correlations between variables, so all variables were included, and numeric variables were standardized on a 0 to 1 scale.
- 184 records were randomly selected for the 20% testing section
- Best Performing Models by Evaluation Techniques:
  - Area Under Curve: XG Boost at 0.91
  - Accuracy: Random Forest Classifier at 0.870
  - Precision: Logistic Regression at 0.833
  - Recall: Random Forest Classifier at 0.911
  - F1 Score: Random Forest Classifier at 0.857
- Random Forest Classifier was best on 3 of the 5 evaluation criteria; therefore, the Random Forest Classifier was the best performing model for this dataset.

DISCUSSION

I used each of the three models independently: XG Boost, Random Forest Classifier, and Logistic Regression. My best accuracy result was a score of 0.870 with the Random Forest Classifier. Mahmud et al. (2023) analyzed the same data using a metamodel of Random Forest, Naïve Bayes, and Decision Tree followed by a K-Nearest Neighbor analysis of the three previous models' prediction and the true presence of heart disease to get an accuracy of 0.87.

The three different models also seem to use different methodologies when rating variable importance in their classification schemes. When reviewing the top 5 most important variables for each of the 3 models, Only oldpeak shows up in all 3 lists. MaxHR, RestingBP, ST_Slope_Up, and ChestPainType_NAP show up in 2 of the 3 lists. These five variables must be important in predicting heart disease.

For the Logistic Regression model, I chose to use the variable's coefficient as opposed to its odds or p-value as a means to rate it against the importance of the other models. Odds relate only to the outcome of a "1," while p-values may be significant but the coefficient is very small.

LIMITATIONS

- Only 11 explanatory variables contained in the dataset.
- Maybe more variables would have given better performance.
- Other models, or other combinations of models may give better performance.
- Seem to be differences in variable importance by classification method.

Cited Reference:
Mahmud, I.; Kabir, M.M.; Mridha, M.F.; Alfarhood, S.; Safran, M.; Che, D. Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel. Diagnostics 2023, 13, 2540. https://doi.org/10.3390/diagnostics13152540