PREDICTING IF A PERSON IS PRONE TO A HEART ATTACK

STA 567 Final Project

Kelvin Njuki

Date: May 6th, 2021

Introduction

Heart attack is among the cardiovascular diseases which are part of the chronic diseases. It occurs when there is a blockage in the flow of blood to the heart leading to insufficient oxygen in some parts of the human heart. This blockage mostly happens due to plaque raptures. If a quicker action to unblock the arteries is not taken, the section with the blockage starts to die. This may cause death to the affected person.

The following factors are commonly associated with increased risk of a heart attack, age of a person, high cholesterol levels in the body, high blood pressure, obesity, among others. It is advisable to go for a check-up often or engage in exercise to improve the health of your heart. If a heart attack is detected at an early stage, the affected individual can be treated. Therefore, this project aims at predicting the chances of getting a heart attack based on several factors.

The project used heart attack data that was extracted from kaggle.com URL: https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset. The data contains 14 variables (13 predictors and one response variable (`output`)). The description of each variable is outlined below.

- `age` - Age of the patient in years
- `sex` - Sex of the patient (1 = male; 0 = female)
- `cp` - Chest pain type ~ 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic
- `trtbps` - Resting blood pressure (in mm Hg on admission to the hospital)

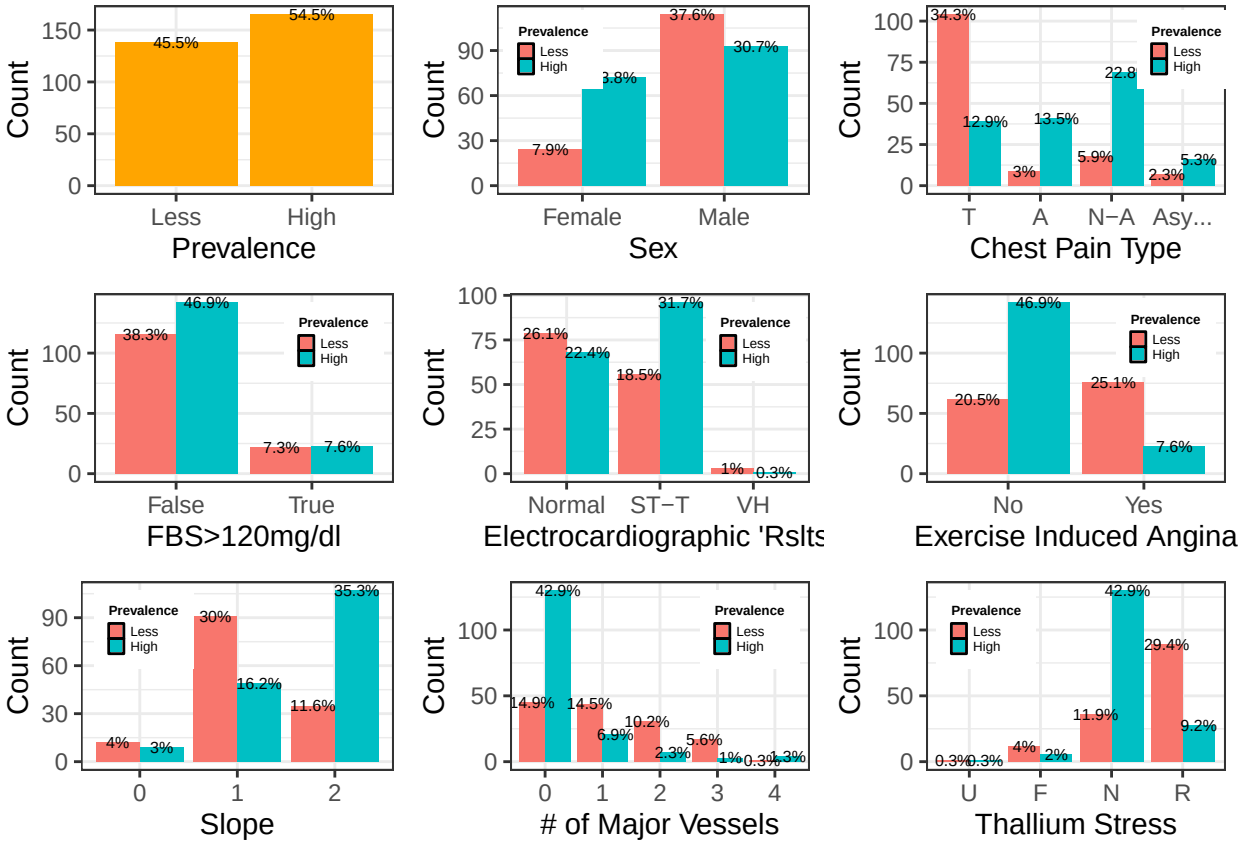
- `chol` - Cholestoral in mg/dl fetched via BMI sensor
- `fbs` - fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False
- `restecg` - Resting electrocardiographic results ~ 0 = Normal, 1 = having ST-T, 2 = hypertrophy
- `thalachh` - Maximum heart rate achieved
- `exng` - Exercise induced angina ~ 1 = Yes, 0 = No
- `oldpeak` - ST depression induced by exercise relative to rest
- `slp` - The slope of the peak exercise ST segment (0 = upsloping; 1 = flat; 2 = downsloping)
- `caa` - Number of major vessels (0-3) colored by flourosopy
- `thall` - Thallium Stress Test result ~ 0 = unknown; 1 = fixed defect; 2 = normal; 3 = reversible defect
- `output` - Target variable 0(less chance of heart attack) and 1(more chance of heart attack)

The rest of the project is organized as follows, Exploratory Data Analysis, Model Building, and Conclusion.

Exploratory Data Analysis

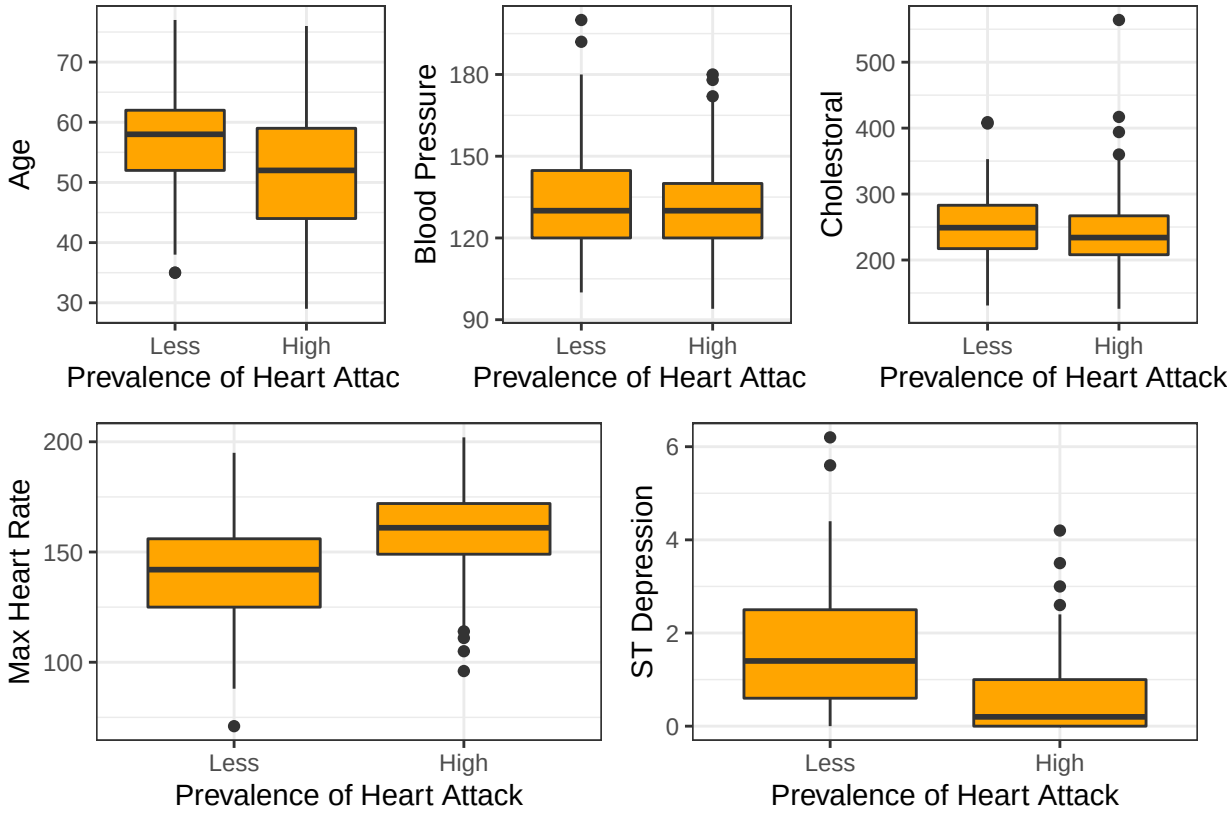
Exploratory data analysis was necessary to gain insight into the data. The data has 9 categorical variables (including the response variable) and five quantitative variables. To assess the association between categorical predictors and the response, grouped bar plots were generated for each categorical predictor and the response variable. Comparative boxplots were constructed to compare the distribution of quantitative variables across the levels of the response variable.

i. *Exploring the association between each categorical variables and the response variable.*



Overall, the data suggest that people have a high prevalence of getting a heart attack. Males are more likely to get a heart attack than females. Patients with non-angina chest pain occupy the largest proportion of patients with a high chance of getting a heart attack. Those with typical chest pain are most unlikely to get a heart attack. Most of the patients with fasting blood sugar that is less than 120 mg/dl are unlikely to get a heart attack. Among the patients with a high chance of getting a heart attack, the majority have ST-T resting electrocardiographic results. People with a downsloping slope of the peak exercise ST-segment have the highest chance of an attack. Among those who have a high prevalence of an attack, patients with zero major vessels are the most. Thallium stress of 2 implies a patient has a high chance of a heart attack. Generally, sex, chest pain type, exercise-induced angina, slope, number of major vessels, and thallium stress have a high association with the prevalence of an attack.

ii. *Comparing the distribution of the quantitative variables from the two levels of response variable.*



It is surprising that, on average, patients with a high chance of heart attack have a median age of 52 years, while those with a low chance have a median age of 58 years. Intuitively, it is expected the older you are, the higher the chance of an attack, but that is not the case here. Similarly, people with high cholesterol are expected to be prone to a heart attack, but the data shows no significant difference. High blood pressure is one of the factors that can cause a heart attack. It is noted from the plots that the median blood pressure is the same for people with a low and high chance of a heart attack. There is a high likelihood of getting a heart attack if your maximum heart rate is high. Overall, age, maximum heart rate, and ST depression are most likely to influence the chances of getting a heart attack.

From the (i) and (ii) above, we can see sex, chest pain type, exercise-induced angina, slope, number of major vessels, thallium stress, age, maximum heart rate, and ST-depression are highly influence the prevalence of a heart attack.

Model Building

Different types of models were fitted in search of the best model for predicting a heart attack based on all predictors and a reduced number of predictors. These models include K-Nearest Neighbors, logistic regression, partial least squares, regression trees (regression trees, boosting, bagging, and random forests), and support vector machines (support vector classifier and support vector machine with polynomial and radial kernels). All the models were evaluated using a 10-folds cross-validation approach. For models with tuning parameters, the appropriate range was arrived at after fitting the model multiple times. All the models fitted were compared by their prediction accuracy to choose the best model. The comparison is shown in the table below.

Table 1: Comparing the models by their accuracy

Model	Accuracy
KNN	83.81
Logistic	82.51
PLSR	86.83
RegTree	77.56
Bagging	77.21
Boosting	83.56
RandomForest	83.13
SVC	85.76
svmP	80.19
svmR	84.49

Conclusion

The results suggest that the partial least squares regression model is the best in predicting the chances of a heart attack in a patient. The model has an accuracy of 86.83% and 7 components. Based on the exploratory data analysis, 10 predictors showed a high association with the response variable. The ten predictors were used to build a reduced partial least squares regression model. This model was compared with the full model reported above.

Table 3: Confusion matrix from a PLS model

	Less	High
Less	113	12
High	25	153

Table 2: Comparing full and reduced PLS models by their accuracy

Model	Accuracy
FullPLS	86.83
ReducedPLS	85.76

The partial least squares regression model still emerges as the best model compared to a PLS model with 10 predictors. Therefore, we conclude it is the best in predicting the prevalence of getting a heart attack. The prediction accuracy is 86.83% with 7 components in the model. Its confusion matrix is displayed below. Its confusion matrix is displayed below, which shows 0 (less chance) are more misclassified compared to 1 (high chance).

References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Class Notes, Homework and In-class Assignments.