Prediction of Heart Disease using Machine Learning

Rellan, Chayan
cr3194@columbia.edu

Palermo, Eric
ep2979@columbia.edu>

Imam, Fariha
fi2183@columbia.edu

April 24, 2021

Abstract

Cardiovascular disorders have been the most frequent cause of death in recent years. Heart disease is caused by a variety of factors, including an unhealthy lifestyle, lack of physical activity, obesity or opioid and alcohol use, both of which lead to social issues too. The purpose of our project is to develop a heart disease prediction system that predicts whether a patient is suffering from heart disease or not.

Machine Learning models like Logistic Regression, Random Forest and Gradient Boosting are used for heart disease predictions.

Keywords - Logistic Regression, Random Forests, Gradient Boosting, Heart disease prediction

1 Introduction

Machine Learning is commonly used in the area of healthcare in numerous areas of medicine, such as identifying rare disorders, interpreting trends to predict a rare disorder, and so on. According to a World Health Organization estimate, heart attacks and strokes account for 17.5 million deaths worldwide. The aim of this project is to use the UCI dataset to determine whether or not the patient has a heart condition. The dataset originally included 76 attributes that were gathered from four separate datasets, and we used fourteen of them in our analysis. The dataset was extracted from the ML related projects website Kaggle and the size of the dataset downloaded is 303 rows and 14 columns.

The dataset contains 14 attributes such as:

- age - in years
- sex - (1=male, 0=female)
- cp - chest pain type
- trestbps - resting blood pressure in mm Hg
- chol - serum cholesterol in mm/dl
- fbs - fasting blood sugar > 120mm/dl
- restecg - resting electrocardiographic results
- thalach - maximum heart rate achieved
- exang - exercise induced angina
- oldpeak - ST depression induced by exercise relative to rest
- slope - The slope of the peak exercise segment
- ca - Number of major vessels (0-3) colored by fluoroscopy.
- Thal - 3=normal, 6=fixed, 7=reversible defect
- Target - have disease or not (1=yes, 0=no)

2 Data Preprocessing and Feature Engineering

To avoid errors in the models' predictions, the correlation between each of the predictor variables was calculated. The majority of variables had little to no correlation with each other, as indicated by a Pearson correlation coefficients of less than 0.10. The maximum heart rate achieved ("thalach") was weakly correlated with other predictor variables, with several correlation coefficient values around 0.35. However, "thalach" was still used as a predictor variable because removing it did not consistently improve the Brier scores of the models in predicting the test data.

To avoid overfitting the model to unimportant variables, the correlation between each predictor variable and the target was also assessed. Cholesterol (“chol”) and fasting blood sugar (“fbs”) had very little correlation with the target variable, with correlation coefficients of -0.09 and -0.02, respectively. They were thus dropped from consideration.

The remaining predictor variables were then subdivided into categorical and continuous variables. Categorical variables were preprocessed using OneHotEncoder to transform them into a DataFrame easily processable by machine learning models. The complexity of the model is also considered by fitting to a second-degree polynomial using PolynomialFeatures. Continuous variables were normalized using StandardScaler and also fit with a second-degree polynomial.

Finally, ColumnTransformer was used to specify which columns would be preprocessed as categorical variables versus continuous variables. The data set had no missing values, so imputation was unnecessary. The data was also already in tabular format, so no additional data wrangling was performed.

3 Proposed Models

Three classification models were selected to predict whether or not a patient has heart disease: a logistic regression model, a random forest classifier and a gradient boosting classifier. Logistic regression is a binary classification model that is used when the dependent variable can have one of two outcomes. This model is beneficial in explaining the relationship between a set of nominal, ordinal, or categorical independent variables and the binary dependent variable. Cross validation was performed to optimize the column transformer polynomial degree using 5-fold split.

A random forest model builds several binary decision trees and combines their results to provide a more stable prediction. The features in each binary tree are randomly selected, which allow specific trees in the forest to isolate more important features, thus implementing feature selection. For these reasons, the random forest classifier was chosen as another predictive model for the Cleveland dataset, and is expected to out perform the logistic regression model. The hyper-parameters that were optimized using 5-fold cross validation were the max depth, minimum sample split, and the polynomial degree of the column transformer.

The third model selected was gradient boosting. This machine-learning algorithm improves model predictions by training the predictor using the residual errors of the predecessor. Each successive model attempts to correct for the deficiencies of the models before it. The hyper-parameters that were optimized using 5-fold cross validation were the polynomial degree of the column transformer, the minimum samples split, the learning rate, and the number of boosting stages.

4 Result and Analysis

Logistic Regression could achieve a brier score of 0.1012 on the train set and 0.0991 on the test set. We see that the model has better generalization. Also, the optimal values for C (which inversely varies with regularization) and degree of the polynomial are 0.1 and 2 respectively.

On the other hand, Random Forests could achieve a brier score of 0.0844 on the train set and 0.1081 on the test set. The optimal value for degree of the polynomial, max depth (maximum depth of the trees in Random Forest), min samples split ((minimum samples in a leaf after the split), min samples leaf (minimum samples in a leaf) turned out to be 1, 4, 4, 5 respectively.

Gradient Boosting could achieve a brier score of 0.0712 on the train set and 0.1231 on the test set. The model performed well on the train set but not on the test set. Hence, it did not generalize well and actually overfitted on the training set. The optimal values for polynomial degree, learning rate minimum samples split and number of estimators were 1, 0.05, 8 and 50 respectively where number of estimators in the number of trees used and learning rate tell us the contribution of each new base model (typically a shallow tree) that is added in the series.

Apart from the brier score, we actually tried to use other metrics for classification problems. These include Precision, Recall and F1-Score.

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$$

The classification report is as follows :

Class 0 - Have a Heart Disease	Precision	Recall	F1-Score
Logistic Regression	0.90	0.90	0.90
Random Forest Classifier	0.89	0.86	0.88
Gradient Boosting Classifier	0.78	0.86	0.82

5 Conclusions

Looking at the Brier Score as well as the classification report, we can clearly say that Logistic Regression outperformed the Random Forest and Gradient Boosting Classifier. This brings us to the conclusion that more complex models do not guarantee better prediction.

The results from each model were passed through the function predict_proba() to obtain the predicted probabilities. These probabilities were then combined with the original data set and a groupby.mean() was performed on a specific feature for each machine-learning model. This was done to develop understanding of how that exact feature relates to having a heart condition, as well as to compare models for that feature against one another. The figures below compare the predicted probabilities of the models against the actual number of people documented with a heart condition. The highlighted features are resting ECG results, chest pain type and fasting blood sugar.

Having a resting electrocardiographic test (ECG) with abnormal result showed a higher predicted probability of being diagnosed with a heart condition. A person with a normal resting ECG had a lower predicted probability of being diagnosed with a heart condition. However, the models matched the actual results (y) better on normal resting ECG.

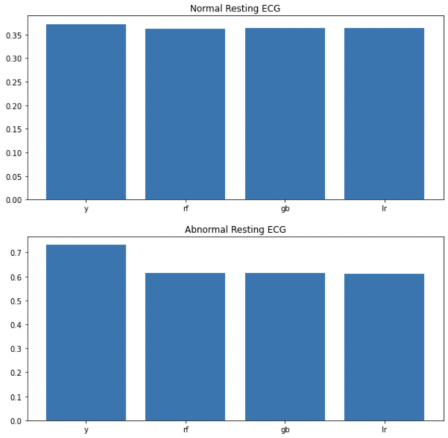

Figure 1: Resting ECG (normal/abnormal)

restecg	y	rf	gb	lr
0	0.371429	0.354379	0.352374	0.322713
1	0.730769	0.618928	0.576277	0.641501

The chest pain type also showed higher average probabilities for specific chest pain types patients reported. Chest pain type 0 relates to typical angina, type 2 relates to atypical angina, type 3 is non-anginal pain and type 4 is asymptomatic. Having a typical angina shows a lower predicted probability of being diagnosed with a heart condition, whereas type 1, type 2 and type 3 show a higher average probability. Logistic Regression produced closer results to the actual diagnosis.

cp	y	rf	gb	lr
0	0.212121	0.236283	0.246009	0.184990
1	1.000000	0.860359	0.829760	0.862404
2	0.923077	0.705408	0.651504	0.780723
3	0.666667	0.630753	0.543433	0.659709

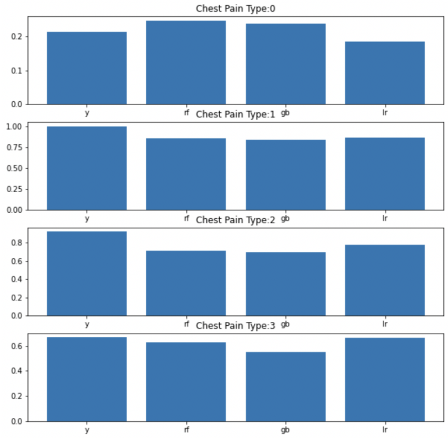

Figure 2: Chest Pain Type (Type 0,1,2)

Fasting blood sugar level did not show influence in having a heart condition, with the predicted probabilities being relatively split between the two groups.

fbs	y	rf	gb	lr
0	0.553191	0.478587	0.473183	0.464671
1	0.428571	0.397243	0.362543	0.411037

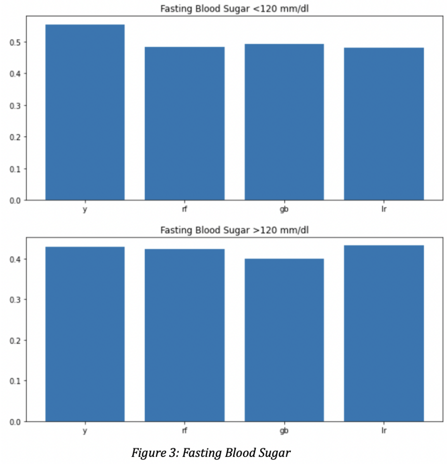

Figure 3: Fasting Blood Sugar (mm/dL)