Ninad Dixit **Capstone project report**

Heart failure prediction in patients

Summary:

A classifier model was developed to predict heart failure in patients whose medical records were collected from 5 locations. Exploratory data analysis showed a significant relationship between heart disease and ST slope, Oldpeak, and exercise angina. Prediction performance of several algorithms was tested – random forest classifier was the best performer with 89% accuracy. Extreme gradient boost classifier was a close second followed by K-nearest neighbors classifier and logistic regression. Several strategies are suggested to improve the model performance, e.g., choosing a better metric and threshold.

Problem Identification:

Cardiovascular diseases (CVDs) led to nearly 17.9 million deaths in 2019, which accounted for 32% of all global deaths. Nearly 85% of them were because of heart attack and stroke. More information is available here: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Several health issues may serve as precursors to CVDs and heart failure/stroke. A model that can predict heart failure in patients can lead to early management of the problem and help in reducing the mortality rate. Thus, the questions is: can we develop a model that will accurately predict heart failure in suspected patients?

Data wrangling:

The dataset was downloaded from a Kaggle competition.[1] This dataset consists of 11 features and is a combination of 5 datasets as follows:

1. Cleveland: 303 observations
2. Hungarian: 294 observations
3. Switzerland: 123 observations
4. Long Beach VA: 200 observations
5. Stalog (Heart) Data Set: 270 observations

Out of 1190 observations, 272 were duplicates and were removed. Thus, the dataset used in this project consisted of 918 unique observations.

Attribute information:

Definitions of dataset features are as follows [1]:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [ST depression induced by exercise relative to rest]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

A quick analysis of the dataset yielded following observations:

1. Majority of the patients were over 40 years old, and the oldest one was 77 years old.
2. Men accounted for nearly 80% of total patients.
3. Many patients showed higher than normal values for resting blood pressure, fasting blood sugar, and desired maximum heart rate.
4. Approximately 55% of patients were detected with heart disease and were susceptible to heart failure.

Exploratory Data Analysis:

This dataset consists of several numerical and categorical features. Relationship of numerical variables with target variable (‘HeartDisease’) was explored using correlation analysis.

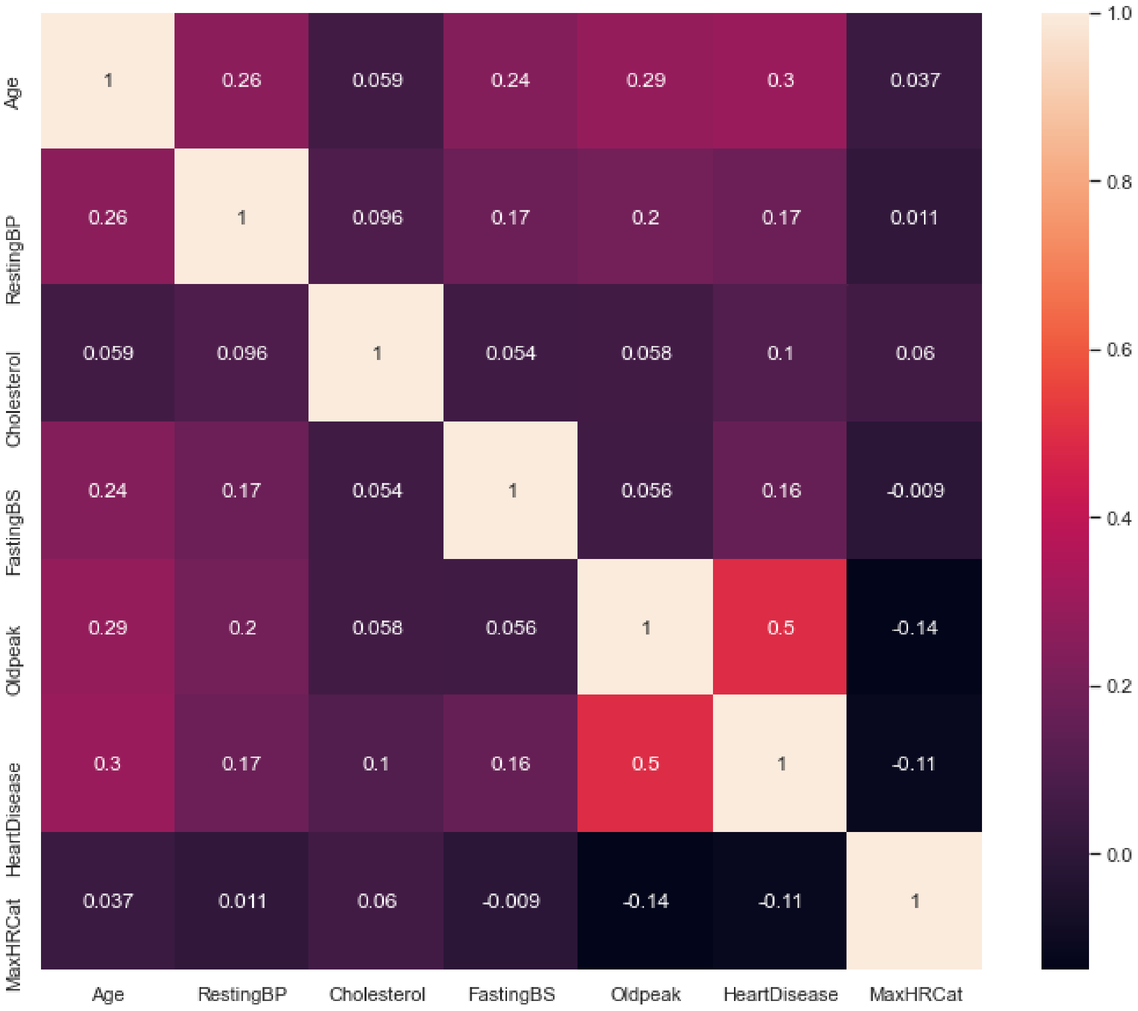

Fig. 1 Heatmap describing correlation between numerical features and target feature.

Among numerical variables, ‘Oldpeak’ (0.5) and ‘Age’ (0.3) showed a significant positive correlation with ‘HeartDisease’. 'Age' also had some degree of positive correlation with 'Oldpeak', 'FastingBS', and 'RestingBP'.

In many cases, plotting individual features against heart disease diagnosis provided a better understanding of their relationship.

Oldpeak:

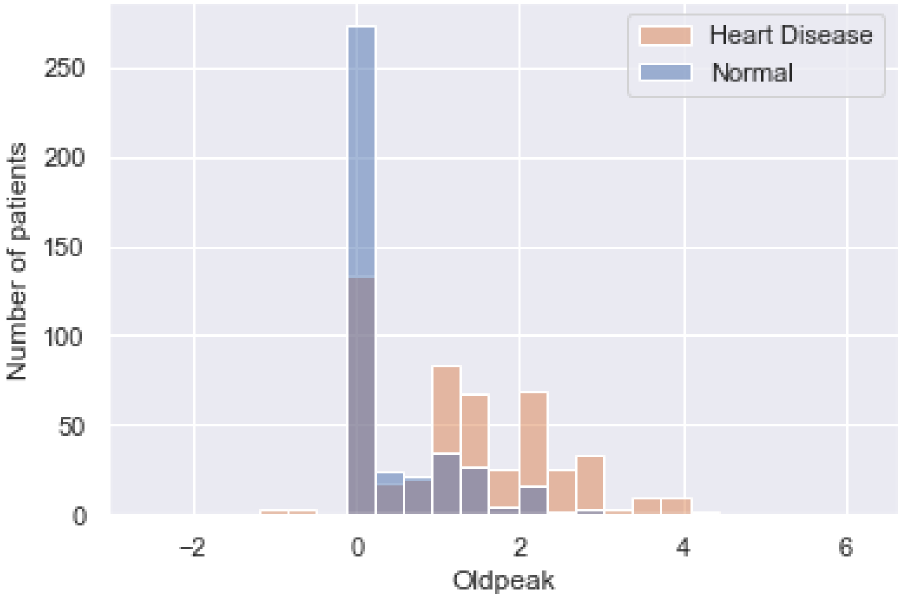

Fig.2 Relationship between heart disease diagnosis and Oldpeak values.

Nearly 50% of total patients' records showed either no or small ST depression (-1 < Oldpeak < 1). When the value was outside this range, patients were more likely to be diagnosed with heart disease. Such ST depression can be attributed a number of causes including heart attack (acute myocardial infarction) and temporary tightening of one or more arteries' walls (coronary vasospasm), which all indicate improper functioning of heart. Although ~ 50% of patients had 'Oldpeak' value between 0 and 1, few individuals exhibited values as high as 6.2.

Exercise angina:

More than 80% of males and 60% of females who reported exercise angina were diagnosed with heart disease. On the other hand, about 40% of males and 17% of females who did not report exercise angina were diagnosed with heart disease. Since angina (chest pain) is observed when there is not enough blood flow through heart, it is a strong indicator of heart disease.

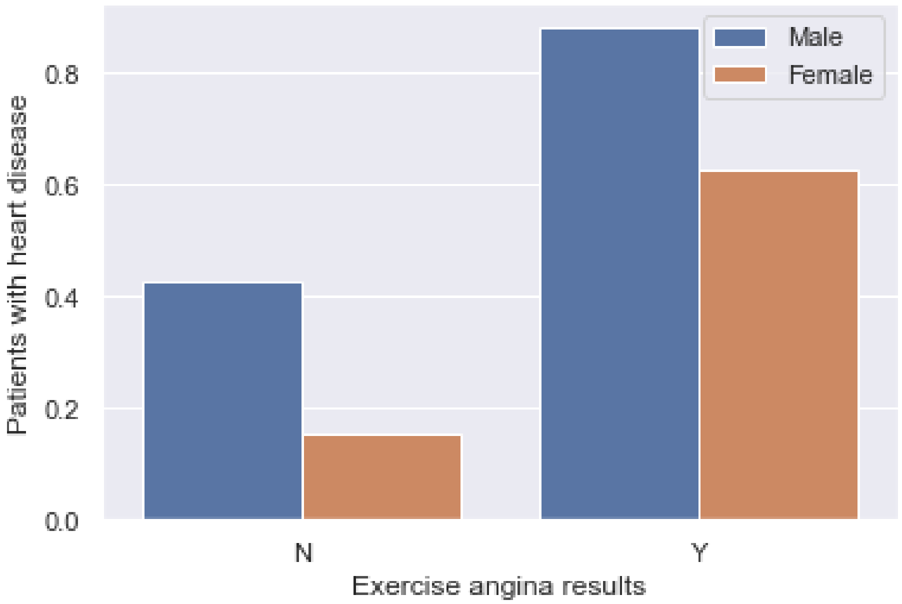

Fig.3 Relationship between heart disease diagnosis and exercise angina results.

Cholesterol level:

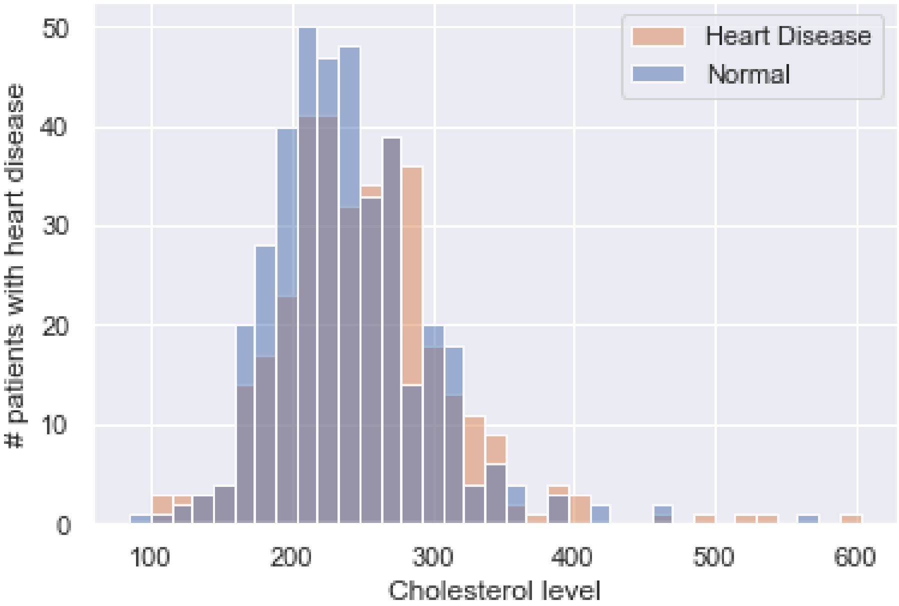

Fig.4 Relationship between heart disease diagnosis and cholesterol level in patients.

Many patients with heart disease showed cholesterol levels < 240 (borderline high). However, as cholesterol levels increased, the number of patients with heart disease exceeded the ones without it. When cholesterol level in blood increases, it builds up in the walls of arteries and reduces the blood flow to heart. Thus, high cholesterol levels is another indicator of heart disease.

Although, there were no empty cells in the dataset, 171 patients had their cholesterol level reported as ‘0’. There are multiple strategies to replace ‘0’ (or a missing value) with an estimated value such as mean or median. However, high cholesterol level is a prominent cause of heart diseases. Thus, instead of assigning a potentially incorrect cholesterol value that may lead to wrong diagnosis, I decided to drop the patients’ data with ‘0’ cholesterol level. The remaining dataset consisted of 700+ entries, which would still provide a reliable predictive model. Few patients also reported very high cholesterol levels (>400). Such levels are likely due to 'Familial Hypercholesterolemia', which is a genetic disease and affects about 1 in 250 individuals. [2] Considering the number and medical records of the patients in this study, it’s possible that several patients whose records were used in this study had this condition.

ST slope:

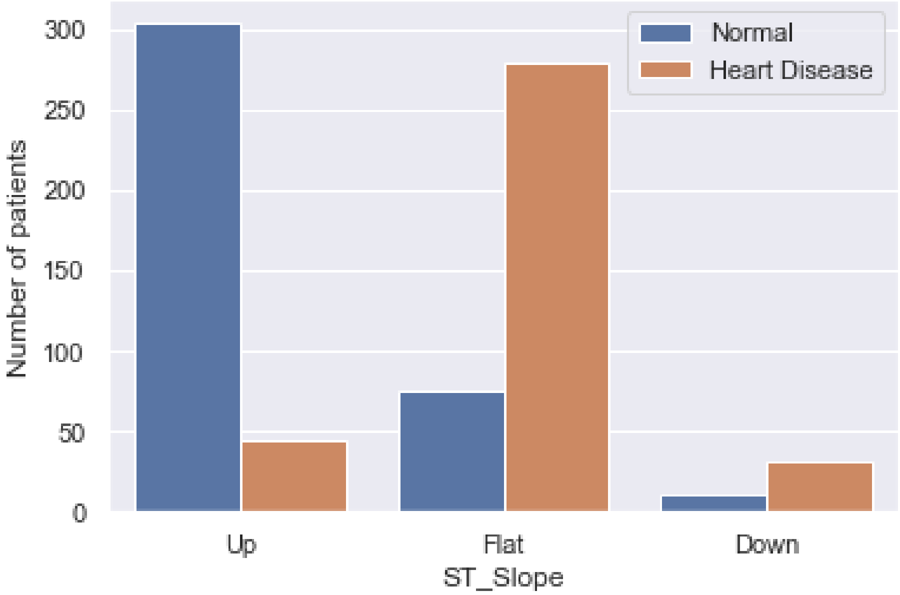

Fig.5 Relationship between heart disease diagnosis and ST slope in ECG.

A large majority of patients with an upward ST slope were diagnosed as normal. On the other hand, a large majority of patients with a flat or downward slope were diagnosed with heart disease. This indicates a strong correlation between ST slope and heart disease.

Model selection and performance:

Since the target feature (‘HeartDisease) contained two classes (types) of outcomes (0 and 1), predicting them became a classification problem. The following models were tuned and employed to predict the classes as accurately as possible:

1. Logistic regression
2. K-nearest neighbors classifier
3. Random forest classifier
4. Extreme gradient boosting (XGBoost) classifier

‘Accuracy’ was used as the performance metric to compare these models. For our purpose, accuracy is defined as:

$$accuracy = \frac{(no.\ of\ true\ positives) + (no.\ of\ true\ negatives)}{(total\ no.\ of\ observations)}$$

Performance of these models on cross-validation and test datasets is as follows:

Model	**Accuracy**		**ROC-AUC**
	Cross-validation	**Test**	
Logistic regression	0.85	0.87	-
K-nearest neighbors	0.84	0.87	-
Random forest	0.85	0.89	0.956
XGBoost	0.85	0.89	0.936

Table 1. Prediction accuracy of classification models

Random forest and XGBoost classifiers were the best performers – narrowly improving over logistic regression and K-nearest neighbors classifiers.

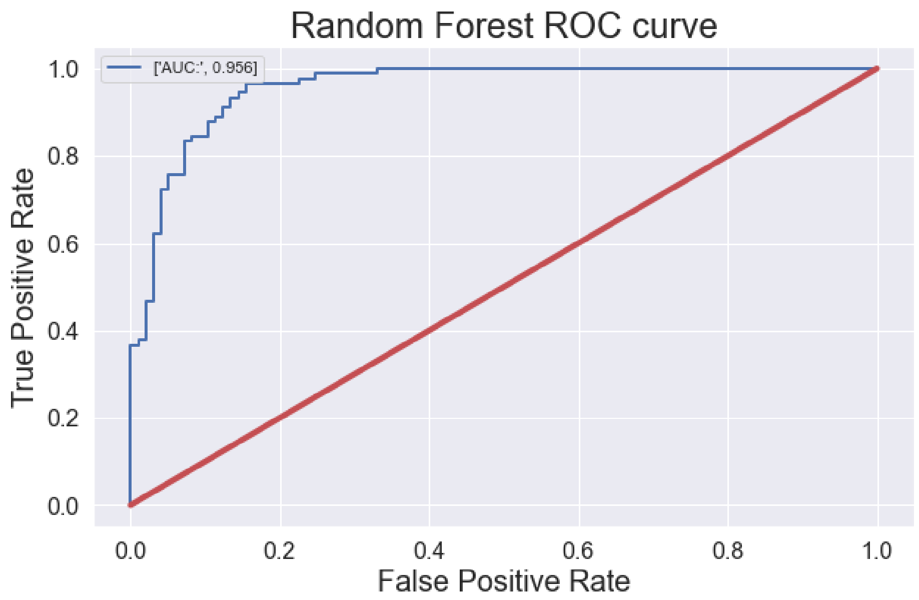

Fig.6 ROC curve for random forest classifier model.

ROC curve describes the ability of a model to identify true positives and false positives at various threshold values. Typically, classifier models have a threshold of 0.5. This means if the predicted probability of target feature ‘y’ is > 0.5, ‘y’ is assigned the value 1. Depending on the importance of finding more true positives or true negatives, this threshold might change. For example, an algorithm detecting spam emails should be more sensitive to false negative, i.e. emails that are spam but are not classified as such. Threshold adjustment is can be carried out in consultation with a subject matter expert and using ROC curve.

Area under the curve (AUC) is another metric that helps identifying better performing models. A truly random predictor model would have 50:50 chance of assigning '1' to the target feature. This model is represented by the straight line (x = y) in ROC curve, and AUC for this line is 0.5. AUC for an ideal model (100% true positive rate and 0% true negative rate) would be 1. Thus, a better model would have AUC close to 1. Since random forest classifier model has AUC of 0.956, it is expected to show good performance on new data.

According to random forest model, the top 5 important features affecting heart disease are: ST slope, exercise angina, Oldpeak, chest pain type, and age. This analysis is consistent with the observations made while exploring the data.

Conclusions and future work:

1. Random forest was the best model for this dataset. XGBoost classifier was close second followed by K-nearest neighbors and logistic regression.
2. ST slope, exercise angina, Oldpeak, chest pain type, and age had the most impact on heart disease diagnosis in the given set of patients.
3. There is scope for further improvement in the model performance. After consulting a subject matter expert, several strategies can be applied to achieve better predictions.
    a. Choose the right metric – evaluate whether accuracy is the appropriate metric for this study. Consider using precision, recall, or f1-score as a metric to tune the model.
    b. Choose the right threshold – this model uses default threshold of 0.5. Depending on the chosen metric, this threshold may be adjusted to provide improved diagnosis.
    c. Remove features with low importance – choosing only those features that have a significant correlation with the target feature may improve the model performance.
    d. Use other classifiers such as LightGBM and/or deep learning architectures.

Recommendations for client:

1. One of the main concerns with the model prediction is time. We can predict possibility of heart failure in patients, but cannot specify the time period over which the prediction are valid. Thus, client is recommended to use the latest medical records of patients for diagnosis.
2. Since the model accuracy is 89%, its predictions should be coupled with expert opinion before informing patients about their diagnosis.

References:

1. fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [12/2/2021] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.
2. Retrieved [12/3/2021] from https://rarediseases.org/rare-diseases/familial-hypercholesterolemia/.