Capstone Final Report
Michael Dyer

Background Info and Findings

I have chosen, for my capstone a dataset from Kaggle that contains 11 features (Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Pressure, Resting ECG, Max Heart Rate, Exercise Angina, Oldpeak, St_Slope, and, ultimately, Heart Failure). I will use these features to help predict whether or not a patient contracts heart failure.

I was able to develop a KNN model that can predict with 90% accuracy whether an individual will develop Heart Failure. Since this problem is centered around medical diagnoses, I would be more interested in the Recall (false negative) score. This model also had the highest Recall score at 94%.

Problem Statement

Almost 18 million people per year are killed by cardiovascular diseases. This accounts for 31% of worldwide deaths. Heart Failure itself is a common event of those who have CVD's. The problem to be addressed in this project is, how can we use given health metrics (listed above) to create a model to help with early detection and management.

Data Wrangling

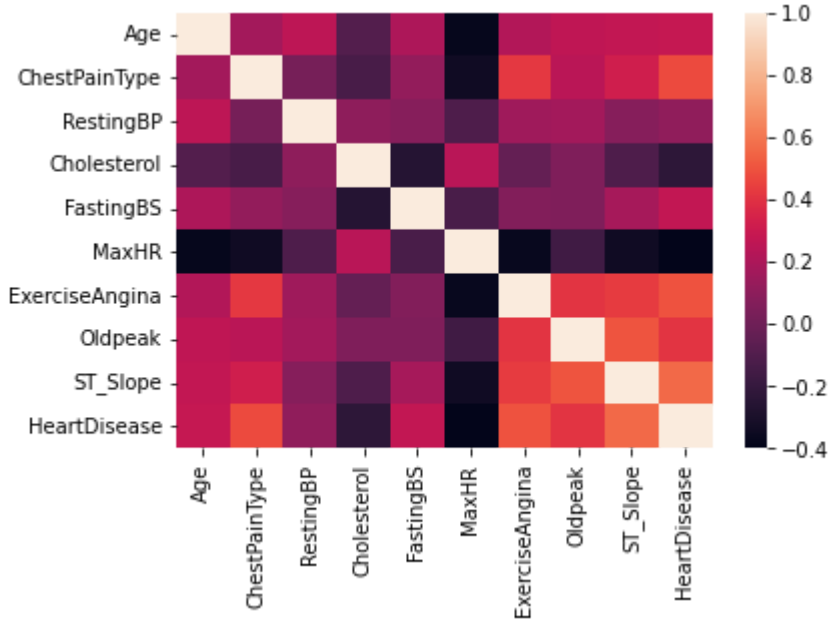
The raw dataset came from Kaggle. The dataset included 918 rows and 12 columns. There was no 'missing' data in the form of NaN, but two columns had issues. In the RestingBP column, one individual had a '0' present, and in the Cholesterol column, there were 172 individuals that had '0' for their Cholesterol. From there I found summary statistics for each column, and replaced any categorical values with numerical ones (Exercise Angina went from 'N' or 'Y' to '0' or '1'.) I imputed the missing Cholesterol values with the average, since less than 20% were missing. Lastly I determined outliers.

Exploratory Data Analysis

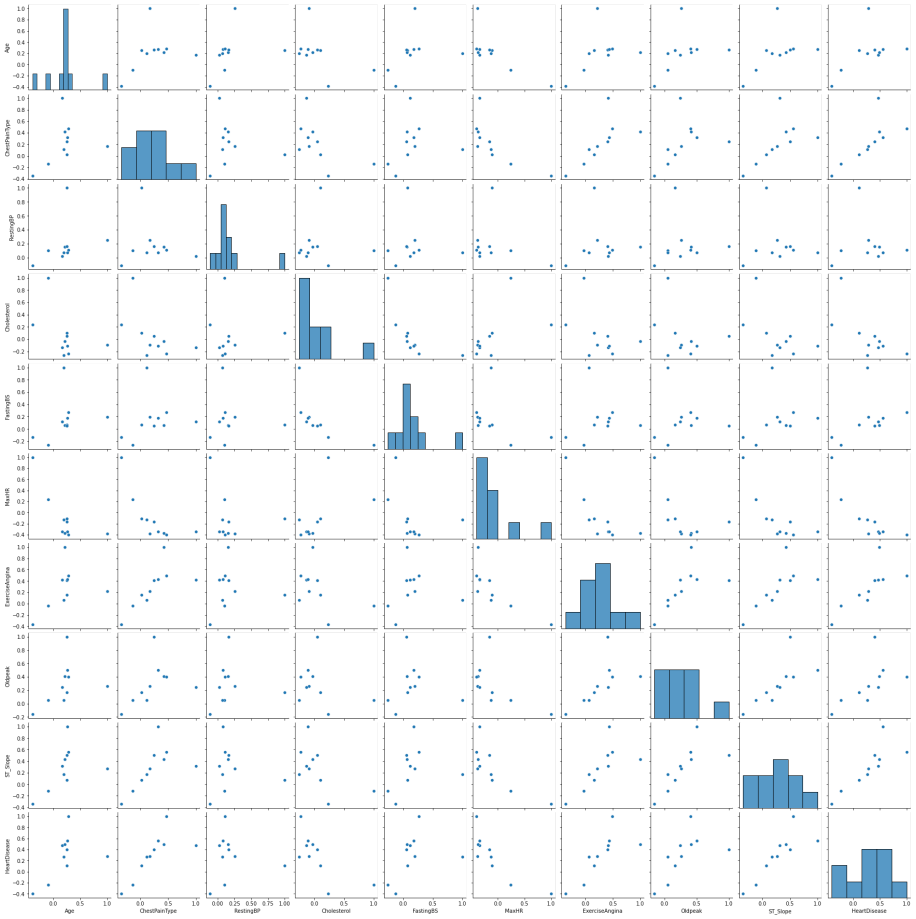
I first looked to see how the data was distributed in each column and then looked at the Correlation Coefficients. I was looking for two things; to see if there were any categories that were more correlated with Heart Disease than others, and to see if there were any collinearity among categories so that I could perform feature reduction. I found that the strongest correlations among categories within individuals that developed Heart Disease was ST_Slope, ChestPainType, and ExerciseAngina.

From there I performed Chi-Statistic, and t-test evaluations to check for a relationship between the various features and the indicator of Heart Disease. All of the features were shown to have some sort of relationship with Heart Disease.

Heatmap showing Correlation amongst features



Pairplot showing Correlation amongst features



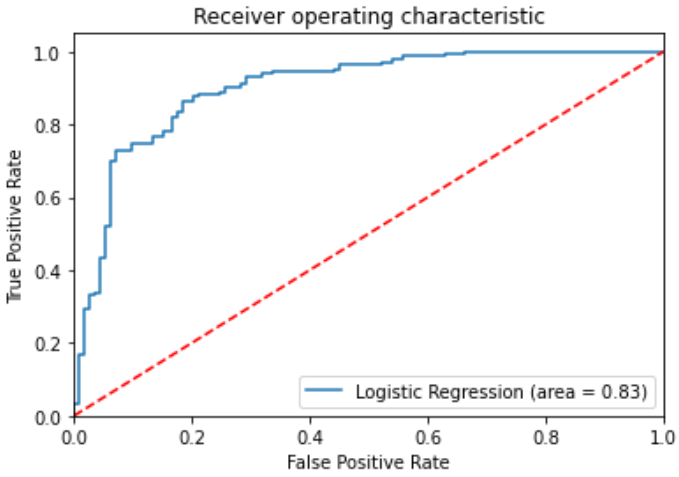
Pre-Processing

For the pre-processing step I split the data into numerical and categorical features, and scaled the numerical features. I then split the data into Test and Training splits to prepare for model testing.

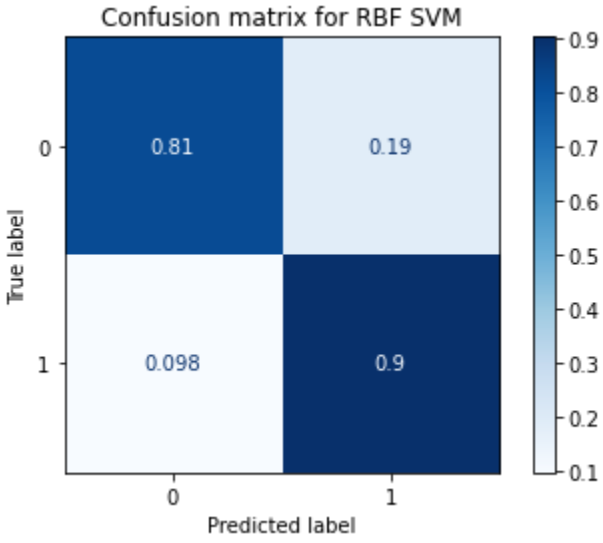
Model Selection

I developed several models in an attempt to determine which would be most successful. I created Logistic Regression, SVM, Entropy Tree with No Max Depth, Gini Tree with No Max Depth, Entropy Tree with Max Depth of 5, Gini Tree with Max Depth of 3, Random Forest Model, and KNN. I was able to obtain reasonably good results with most models right out the box. For example, the Logistic Regression had a total accuracy of 84%, and a recall of 87%. Ultimately I found the best results with a hyperparameter tuned KNN model.

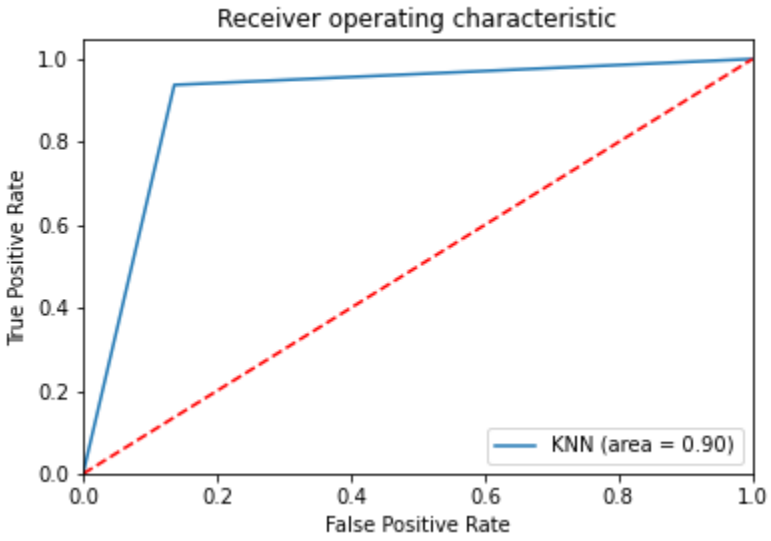
Logistic Regression ROC



SVM Confusion Matrix



Hyperparameter Tuned KNN Model Info



Accuracy: 0.9021739130434783
Balanced accuracy: 0.9005681818181819
Precision score: 0.8823529411764706
Recall score: 0.9375

Takeaways

The KNN model was able to accurately predict whether or not an individual would develop Heart Failure with 90% accuracy. Since this is a medical issue, however, Recall is much more important of a score (# of False Negatives). This model had a recall score of 94%. Having a lot of false negatives in this case would result in having more individuals with Heart Failure test negative, thus, not getting the treatment they need.

The KNN model beat out the Random Forests and Logistic Regression models in both total Accuracy and Recall.

Future Research

In the future I would love to look more into the features themselves to see if any further engineering can be done. I would also love to look into other diseases to see if there is any correlation with any other diagnoses and Heart Failure (might run into issues with HIPPA).

Also, I would like to be able to take more time to fine tune different models and hyperparameters to help ensure that each model is performing at its peak level. Any increases in performance for this type of model (medical diagnoses) must be prioritized.