Heart Disease Capstone Report

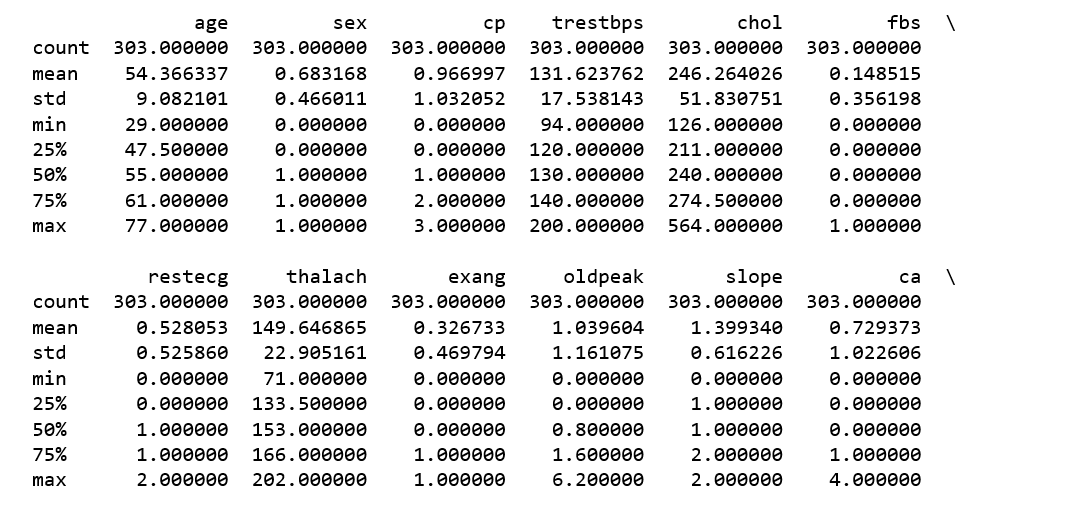
Executive Summary:

The purpose of this analysis was to determine which characteristics were most predictive in determining whether or not a patient had heart disease. A dataset of 303 records was analyzed, with each record consisting of 13 patient characteristics(features), and a label of whether each record referred to a patient with heart disease or not.

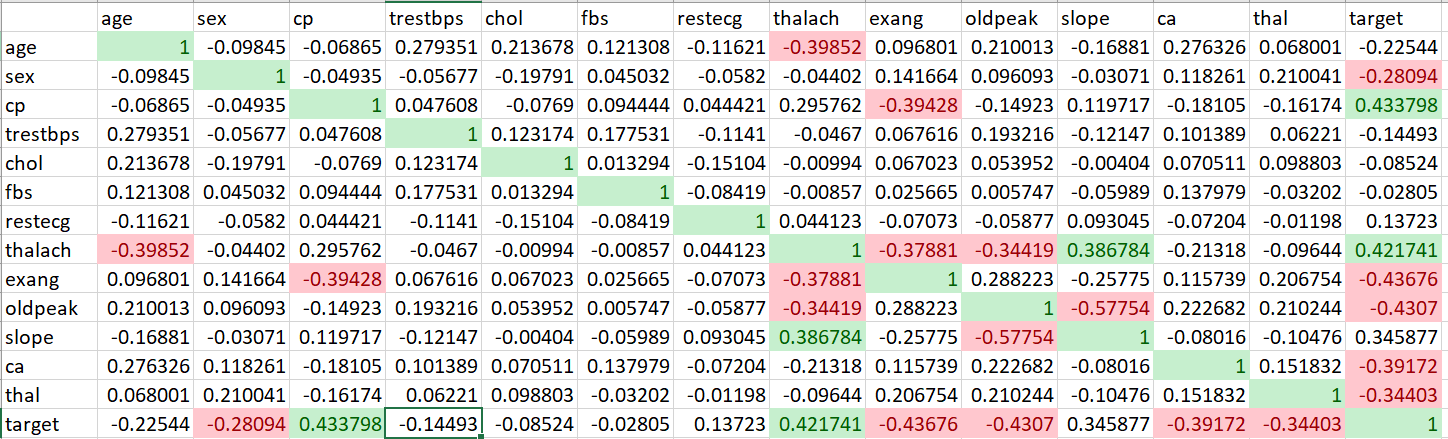
The first step of the process was to do initial data discovery and exploration. Normally in a predictive analysis like this, preprocessing must be done. However, this dataset was taken from Kaggle and already pre-processed, so things like replacing missing values, applying feature transformations, etc. were not needed. This data analysis and discovery allowed me to identify which characteristics had some of the strongest correlations with the presence of heart disease, which were then included in various machine learning models to try to achieve the highest accuracy.

I was able to determine the following characteristics were the most important in terms of accurately predicting the presence of heart disease: Age, OldPeak, Ca(number of major vessels covered by fluoroscopy), cp(chest pain type), and the Slope.

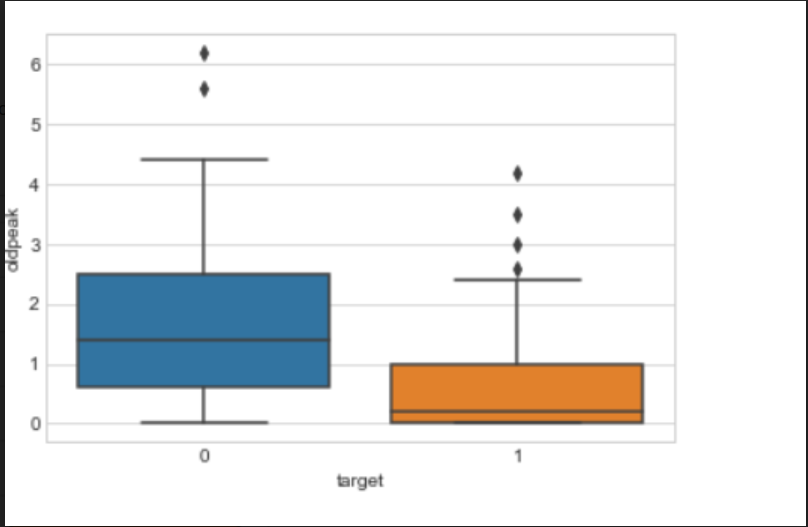
Data Exploration:



After this initial describe operation was performed to get acclimated to the distributions of the feature set, I did a correlation analysis to identify any strong correlations between either the features themselves or the features with the target. A table is shown below:



Furthermore, visualizations were created using python’s matplotlib:



This type of graph was created for each feature to get a since of how the distributions of each feature differed when grouped by whether the target has heart disease or not(1 or 0). In the graph above, one can see that the oldpeak feature can be valuable in predicting heart disease.

Classification Analysis:

Because we are attempting to predict whether a patient has heart disease or not, we use two class classification algorithms in order to try to maximize our accuracy. Prior to instantiating any models, we use a train test split in order to set aside some records in order to test on. Both random forest and logistic regression models were created, with several permutations of the initial 13 features included in different trials. The logistic regression models consistently outperformed the random forests. The highest accuracy achieved with random forest models was .75, whereas my final logistic regression model achieved an accuracy of .82. As mentioned above, this accuracy was achieved when the following five features were included: Age, OldPeak, Ca, cp, and the Slope. Furthermore, cross validation was performed to determine the best C parameter for my logistic regression model in order to maximize accuracy. It was determined the best C parameter was 1.

Conclusion:

In conclusion, we were able to determine that, using a logistic regression model with five of the 13 given characteristics, we could predict whether a patient would have heart disease with around an 82% accuracy.