**Final Project – Milestone 5**

**Presentation Link:** [**https://youtu.be/wTrhfby78sA**](https://youtu.be/wTrhfby78sA)

Theodore Koby-Hercsky

Bellevue University

DSC 630: Predictive Analytics

Professor Werner

31 May 2022

**Final Project – Milestone 5**

**Introduction**

The Data set I have selected for my project is an Excel CSV file that is on heart failure that is broken up into twelve variables. Which consist of the patients age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG (electrocardiogram), maximum heart rate, exercise induced angina, old peak, ST slope, and heart disease. I would like to use this data set to predict the likelihood of patients experiencing heart failure with the variables given. As I have a congenital heart condition and have had multiple open-heart surgeries including a pacemaker which makes this topic furthermore important to me.

Regarding my project, I am proposing that I will run and evaluate multiple models such as the Gradient Boosted Model and the Random Forest Classifier. Which will show that some variables from the data set will have a strong correlation with heart failure while others will show a perfectly healthy patient. As for example some may say a male is more likely to suffer from heart failure compared to a female, but we will let the data decide and determine the true probability. While a patient that is having non-anginal pain and has high cholesterol can be proposed to have a higher likelihood of having heart failure compared to an individual with low cholesterol and no pain.

As I believe this data set has the needed variables and amount of data needed to run models such as a Random Forest and Gradient Boosted Models. Creating these two models will allow me to evaluate two different approaches and test different numbers of trees for each model to determine the best fit for each model. As this will allow me to further evaluate both models and see which one has the best accuracy, F1-score, and recall determining which model would be best to predict heart failure for the patients. While in the end I want to be able to determine which model will be best for predicting patients with heart failure.

**Data Review**

When I first reviewed the output for the description from my data set, I noticed I had some variables such as Resting Blood Pressure, Fasting Blood Pressure, and Cholesterol at a minimum of 0.00. Seen below from my output these findings will need to be adjusted further in my project to create an accurate model to predict the correct outcome for the patients.

Table

Description automatically generated

Above we see that the output of the data set has a mean age of 53 years old. With a mean max heart rate of 136.80. While an issue I am seeing is Resting blood pressure as we see the min is zero which is not possible as you need a pulse to be alive so these will need to be removed.

A visualization that I found helpful in understanding this data set is a heat map of the complete data set as this allows shows a high overview of each variable in the data set that is compared to each other. Below the screenshot of my code and heat map show that the highest correlation with the variable heart disease stands with old peak and max heart rate. The use of seaborn has allowed me to determine that the old peak and max heart rate had 40% correlation between them and the variable heart disease which will make both these variables’ ones to look out for in the future when creating and evaluating the models. Chart

Description automatically generated

Above we see that the highest correlation with heart disease stands with Oldpeak and MaxHR coming in at only 40% which could be two variables to watch when building the models. While another visualization that I especially find useful for this project is another seaborn plot but this time it is a box plot. As the focus of this code was to create a loop that created several plots for all my categorical variables such as sex, resting EKG, ST slope, chest pain type, and exercise angina. Below we can see that these plots have indicate that our data set is for the most part male patients as we have 700 males and a little less than 200 female patients. Which is interesting to see as is this just because the hospital or facility that collected the data had most male patients or was the data filtered and picked through before posted. It is also interesting to see that most patients have asymptomatic chest pain coming in at 500 while the second highest is NAP at a little over 200 patients. Chart, bar chart

Description automatically generated

**Summary of Methods Used and Results**

When I viewed my data set and used describe on it, I found that the minimum value for resting blood pressure and cholesterol was zero. Which is wrong as a patient would need to have a resting blood pressure and cholesterol level to be alive. Due to this I decided to use replace to change any zeros in these variables to the median rate for each variable. By removing the zeros, I got rid of outliers that were created by error that would have skewed my data. Table

Description automatically generated

As seen above we see the min for Cholesterol is 85.00 and Resting Blood Pressure is 80.00. Which has gotten rid of outliers that would be considered outliers in error as these two variables need a numeric value greater than zero. Next, I will use get\_dummies to create dummy variables for all categorical variables within the data set. Table

Description automatically generated

Above I have turned all my categorical variables such as gender, chest pain type, resting EKG types, exercise angina, and ST slope to my dummy variables. I will split my data and into training and test data for my updated\_heart\_data set. Once my data was split, I started with my Gradient Boost Model that fit trees that ranges from three to one hundred. Graphical user interface, text, application, email

Description automatically generated

As seen below the lowest error we see is the model with 25 trees and levels off and heads back down at around 80 trees.

Table

Description automatically generated

While the create of a plot shows the number of errors for each set of trees that the model has created as we have determined that the model with 25 trees has the lowest error but starts to head back down towards 80 trees Chart, line chart

Description automatically generated

While below we see that the accuracy of the f1-score for the Gradient Boost model is 0.86 with support of 368. while the precision for the macro avg and weighted avg is 0.86

Table

Description automatically generated

The next model I have decided to go with is the Random Forest Classifier that tested random trees that ranged from 25 to 400. Graphical user interface, text, application, email

Description automatically generated

Chart, line chart

Description automatically generated

Above we see that the random forest model with the minimum error is at 400 that comes in at a little above 0.136. As the random forests with less trees, we see a higher error rate. which shows us as the number of random trees rises in our random forest model we see less of an error. While below we see that the accuracy of the f1-score for the Random Forest model is 0.87 which is .01 higher than the Gradient Boost but has the same support of 368. while the precision for the macro avg and weighted avg is 0.87 which is also higher than the Gradient Boost.

Table

Description automatically generated

Next, I have decided to create several bar plots that compare the F1-score, Accuracy, and Recall of both the Random Forest Model and the Gradient Boost Model. Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated

As seen above the Gradient Boost Model shows a Recall of 0.8461 while the F1-Score is seen as 0.8569 and last the Accuracy is 0.8559. While the Recall for the Random Forest model is 0.8687 while the F1-Score is seen as 0.8674 and last the Accuracy is 0.8674. Which clearly shows us that the Random Forest model will be the best option when predicting patients with heart disease. While I found it very interesting that my Random Forest Model showed the minimum error is at 400 trees while the Gradient Boost Model shows the lowest error, we see at 25 trees.

But I decided to dive a little deeper into these models and pick out the two features that have the highest correlation with heart disease which stands with Oldpeak and MaxHR coming in at 40%. Graphical user interface, text, application

Description automatically generatedAbove you can see that I have created a new data set that only has Max heart rate, oldpeak, an heart disease. Graphical user interface, text, application

Description automatically generatedWhile I have also decided to use SMOTE which is an oversampling technique that will even out my heart disease and non-heart disease patients. Text

Description automatically generated

As I have decided to create an updated gradient boost model that fit trees that ranged from 3 to 100. Graphical user interface, text

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

As seen above the lowest error we see is 19.78% at 90-100 trees which is higher than the original gradient boost model that was 14.13% at 25 trees. Which is higher than our previous model since we zoned in on the most important variables and the use of SMOTE that works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point.

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

Above we see that for the updated random forest model the minimum error is at 725 trees coming in at 25.59% while my original random forest model minimum error was at 400 that comes 13.60%. As we can see that the updated random forests that uses less variables and with SMOTE we see a higher error rate.

**Conclusion**

In the original models we saw that the Gradient Boost Model shows a Recall of 0.8461 while the F1-Score is seen as 0.8569 and last the Accuracy is 0.8559. While the Recall for the Random Forest model is 0.8687 while the F1-Score is seen as 0.8674 and last the Accuracy is 0.8674.

While in the new models directly above we can see that the minimum error for the updated random forest model is at 725 trees coming in at 25.59% while my original random forest model minimum error was at 400 that comes 13.60%. While the lowest error we see for the updated Gradient Boost Model is 19.78% at 90-100 trees which is higher than the original gradient boost model that was 14.13% at 25 trees. As we can see even though the new minimum errors are higher, we have a change in the minimum error as the original random forest model was the lowest while now the lowest minimum error is coming from the updated Gradient Boost Model. These new models took a big spike in error, but this can be because the new models are evened out in terms of heart disease patients and healthy ones while I also only focused on the features that have the highest correlation with heart disease which stands with old peak and Max Heart Rate that came in at 40%.

Overall, both original and updated models shared interesting output and we saw that the updated models had higher minimum errors and the updated random forest model required me to expand my random trees to be 900 as the SMOTE brought my values up to over a 1,000 due to oversampling. Which brought me to a conclusion that out of both models I believe that the original models had a better minimum error rate which would be better for predicting even if it may be slightly skewed. While the original models had a clear winner as we saw that the original Random Forest model will be the best option when predicting patients with heart disease. While I found it very interesting that my Random Forest Model showed the minimum error is at 400 trees while the Gradient Boost Model shows the lowest error, we see at 25 trees.

**Reference**

PARTHASARATHY, SRIRAM. “Top 5 Predictive Analytics Models and Algorithms: Logi Analytics Blog.” Logi Analytics, 12 Feb. 2021, <https://www.logianalytics.com/predictive-analytics/predictive-algorithms-and-models/>.

Shah, R. (2020, July 24). Imbalanced classification: Handling imbalanced data using Python. Analytics Vidhya. Retrieved May 11, 2022, from <https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/>

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from <https://www.kaggle.com/fedesoriano/heart-failure-prediction>.