Theodore Koby-Hercsky

April 22, 2022

DSC630-T301 Predictive Analytics (2225-1)

Dr. Brett Werner

**Course Project: Milestone 2 and 3– Data Selection and Project Proposal**

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 be able to run and evaluate multiple models and may even include other data sets if applicable throughout the course of this class. Which I am proposing that the findings will show that some values in variables from the data set will show a strong correlation between 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. But this data set will take our research further as some variables such as cholesterol and Chest pain types that include TA for typical angina, ATA for atypical angina, NAP for non-anginal pain, and last ASY for asymptomatic. 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.

**What types of model or models do you plan to use and why?**

A model that I plan on using is the Gradient Boosted Model as this model will create a prediction model that is made up of decision trees that are created one tree at a time. This model uses techniques that are considered machine learning that is for the most part boosted. Once a decision tree is created it will fix errors made from previous trees. Another model I plan on using is the Random Forest Classifier as this is one of the most common algorithms that is used. While also being a classifier that can handle large sets of data that is comprised of an assortment of decision trees. As these trees are dependent on the value random vector sample and are created to the largest degree achievable. When using Random Forest for prediction the classifier will use random subsets of the training data which is known as bagging that can have multiple samples to generate an overall average. The reasoning for using Random Forest is to generate a solid learner compared to the Gradient Boosted Model as these will be two great models to compare.

**How do you plan to evaluate your results?**

The results from the Gradient Boosted Model will be evaluated by creating a plot that shows the variables with the highest value for predicting heart failure. As for the most part I will most likely use the training set data when creating the Gradient Boosted Model to evaluate the variables and the connection, they display with heart failure. I will evaluate the Random Forest Classifier by reviewing the accuracy, area under the curve, and the F1-score to determine the performance of the model regarding the different metrics.

**What do you hope to learn?**

When creating this report, I hope to learn new techniques and different models and or classifiers. While also learning which variables have the highest likelihood of predicting heart failure for the patients when using the Gradient Boosted Model. As for the Random Forest Model I hope to learn that the accuracy along with the F1-score to be high with a low standard deviation. On another note, I would also like to expand my knowledge on creating visualizations from my model outputs.

**Assess any risks with your proposal.**

A risk that can be associated with the proposal of this project can be that the change in values within a variable could have no positive or negative effect on a patient’s chance for heart failure. While one might say another risk with this proposal could be that the prediction of a variable could be incorrect meaning that values, we thought had an effect could be completely opposite of what was thought. Another risk I might face could be that the data I chose might not be compatible with other data sets I might find and want to use as the class goes on. If that happens then I will need to determine what other data sets I could use and if I would even need additional data

**Identify a contingency plan if your original project plan does not work out.**

In contingency if my original project plan does not work out, I would like to create a Logistic Regression as this also is a very popular regression that is used in classification. The use of this regression can estimate the mean and median of the data which is great as it can attempt to prevent overfitting. On another note, the use of a Logistic Regression does have a disadvantage as some may say that the results can be complex and misunderstood to some viewers. While I would also like to create a Decision Tree in my contingency plan as this is a straightforward algorithm that is easy to interpret. As a Decision Tree will produce an accuracy score along with a F1 score that can be used to compare against the results from the Logistic regression. While also the use of a Decision Tree can help when creating a Random Forest if no issues arise in my original plan.

**Course Project: Milestone 3 – Preliminary Analysis**

**Will I be able to answer the questions I want to answer with the data I have?**

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 review the output for the description from my data set, I noticed I had some variables such as Resting 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

**What visualizations are especially useful for explaining my data?**

A visualization that I am always drawn towards when creating a report is a heat map of my complete data set as this allows me to see a high overview of each variable in my data set compared to each variable. 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

While another visualization that I especially 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 a majority of 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. As I would like to create additional visualization that could compare values such as non-angina pain to diagnosed heart disease patients or even male and female ratio to see the amount of heart disease patients for both male and female.

Chart, bar chart, box and whisker chart

Description automatically generated

**Do I need to adjust the data and/or driving questions?**

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. Graphical user interface, text

Description automatically generated

The next adjustment I have decided to make is the creation of dummy variables in my data frame. As I turned all my categorical variables such as gender, chest pain type, resting EKG types, exercise angina, and ST slope to my dummy variables by using get\_dummies on the data frame as seen below.

Table

Description automatically generated

As the weeks go on, I plan to dive deeper into my data frame and see what other adjustments I can make to my variables. While the driving question for my project will stay the same for now as I believe the models, I have chosen will be able to determine which is best to use when predicting if a patient will have heart disease or not.

**Do I need to adjust my model/evaluation choices?**

The models I have selected are the Gradient Boost and the Random Forest Classifier. These models will test the error for each number of trees that I have selected and find the best possible outcome for each model. When testing the Gradient boost, I found that the lowest error seen was at 25 trees coming in at 14.13% and rising back up till 90 trees and slowly starts to make its way back down.

Chart, line chart

Description automatically generated

I have decided I will probably adjust these models aa I am currently fitting different amounts of trees for each model as the Gradient Boost ranges from 3 to 100 trees fitted while my Random Forest Classifier is 25 to 300 fitted. I am not sure if I should fit the same number of trees for both models or should they be different ranges of fit? I will be researching into this further for the weeks to come and determine the best course of action for my models.

**Are my original expectations still reasonable?**

I am expecting to have two great models that can predict if a patient has heart disease or not while being above a 75% accuracy. I believe this is attainable goal, but I was expecting that my Gradient Boost would have a higher accuracy than my Random Forest Classifier. So, I will need to further evaluate my models and determine if I want to create a third model if I have time. As now I feel as if my data is clean for the most part, but I would like to dig deeper into my data and do some more research on the medical side as I know some information on heart disease, but I would like to learn more to be able to better manipulate my data frame.

**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/.