Date: 23/02/18

Mark Friel

Task: Classify a patient on whether they will have heart disease and the degree of which they have it, based on a data set of medical records of patients who have varying degrees of heart disease (HD) from zero indicating no presence of HD to 4 indicating a very strong presence of HD. The performance will be the classification accuracy, the number correctly predicted degree of HD considered as a percentage.

Data[: Heart Disease UCI Dataset](http://archive.ics.uci.edu/ml/datasets/heart+disease) (link to original dataset)

Data dictionary: Here(link to GitHub)

Data Description: The data set that I am using is the Cleveland processed dataset, in the original collection of the data there were 72 variables collected but in the processing of the data it was determined that only 14 were relevant to the detection of heart disease. In the future I may revisit this dataset and start with the 72 variables and perform my own analysis to determine the relevant variables but as a first-time self-project I am happy to work with the cleaned version for now.

Classification Models

Ordinal Regression: Initially the first model that I was going to use for the classification model was going to be a simple logistic regression, to give myself a baseline accuracy metric, and them to move onto more complex algorithms in hope of better metrics. While researching logistic regression I happen up on ordinal regression and realised it would be a better fit for my data. As it is a type of regression analysis used for predicting a variable whose value exists on an arbitrary scale where only the relative ordering between different values is significant. This matched the scale of the degree of heart disease 0-4. The overall Accuracy I was getting for this model was 59%. I wasn’t too disappointed in this as I didn’t expect great metrics given it is quite a simple model.

Neural Network: I moved onto neural network as I was hoping to use a powerful algorithm to get an improvement from the ordinal regression however the results I received from this was poorer than the ordinal regression, so much so that I deemed it pointless getting the actual metrics. As it was predicting every instance as not having heart disease. I am still unsure if this is a fault in my data preparation for the model or just a feature of the data, as 55% of the cases have no degree of heart disease and the rest of the 45% of the data is split between the other 4 cases (11.25%) each. I feel this warrants further investigation at the end of the process as I feel that the neural network should be able to predict this. But as this is meant to be a quick analysis I will move onto another model.

Random Forest: The next model I moved onto that I considered powerful was random forest. It showed a small improvement on the ordinal regression in that its accuracy was 61%. I thought there would have been a more significant increase in the metrics maybe up to 70% considering Random forest is considered quite a powerful algorithm

Next Steps

The current metric that I am receiving from the models are not at an acceptable level, further analysis of the data is needed as I believe the problem Is that there are too many cases of 0 degree of heart disease and for the rest of the degrees the models are struggling to distinguish between the different degrees. I see two options for moving forward. One, in the UCI dataset there are two more datasets that have approximately 300 rows in each I feel this will produce an improvement in the predictions however I am reluctant to use this method as in real situations it is likely you will just have to work with the data you are given. The second method that I am considering is to narrow the scope of the task, in that instead of trying to predict the degree of heart disease I will instead just try to predict the presence of heart disease by making a Bin of the degrees 1-4 to signify the presence of heart disease, effectively turning the task from a multinomial classification into a binary classification.

Model Metrics After binning

Random Forest: The over all accuracy metric for this model was able to increase to 80% with the precision being 77% and recall being 85%

Logistic Regression: with there now only being two levels in the dataset I reverted to logistic regression, this could increase the Accuracy metric to 84% with precision being 83% and recall 84%.

Possible Next Steps.

To improve the performance of the models the next steps that I would take if I was to continue with the analysis was to get the two further datasets from the UCI data folder and train the model on all the data sets combined. Principal component analysis could be a step in the right direction however I am unsure on the performance improvement this would give. I think that the possible best step forward would be to figure out the Neural network model as even though no error message is being shown this should be giving better results than what is being demonstrated.