CSC 2730 Final Project Report

Taylor Smith

The problem I chose for my final project was whether or not it was possible to predict if olympic weightlifters physical atributes (age, weight, height, ect) contribute to whether or not they place in the olympics (Getting a gold, silver, or bronze medal). The end goal would be to use the model to predict future olympic weightlifting events just by the contestants physical qualities.

To analyze this problem, I located a dataset that includes all of the olympians that have ever competed, the event they competed in, their physical qualityies, and if they recieved a medal or not.

Models

Decision Tree

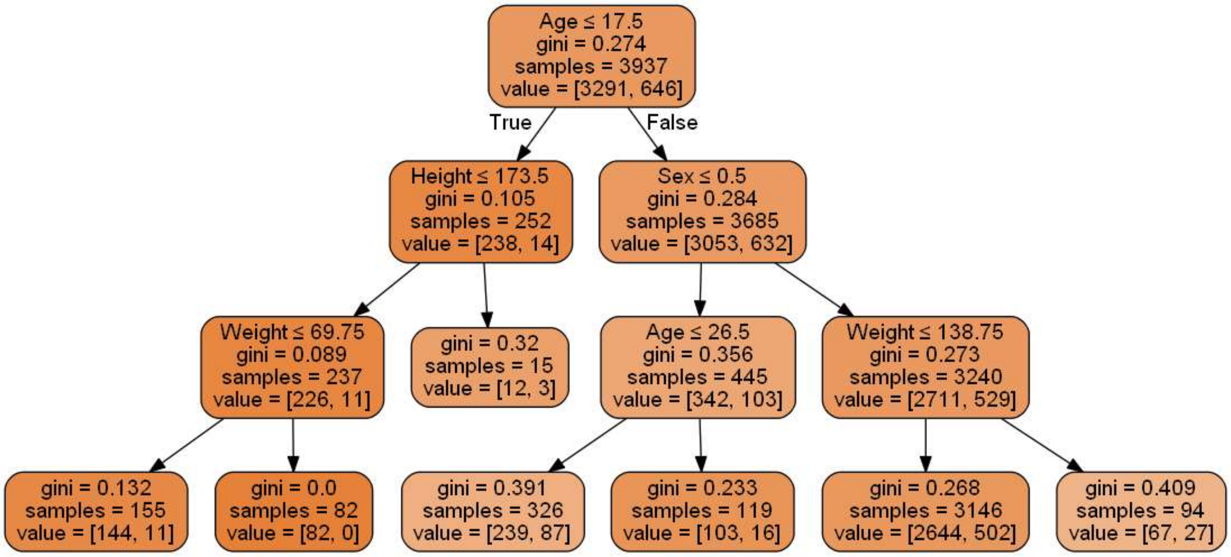
My first idea was to use a simple decision tree classifier. My reasoning behind this was the attrivutes for each athlete such as weight and height would contribute greatly to each to their overall weightlifting performance. And I could not use any sort of regression analysis because I need to classify each athlete into a group.

I ran a 5-fold cross validation to find the best hyper-parameters for the default decision tree, with my results below

In [1]:

```
from IPython.display import Image
Image(filename='download.png')
```

Out[1]:



- The best Hyper Parameters are {'max_depth': 3, 'min_samples_leaf': 10}
- The Best Decisions Tree Model's Accuracy Score is 0.8359156718313436

So the Decision Tree was able to get a pretty good predictor for whether or not the weightlifter placed, and judging by the tree it created, all it needed to get its prediction was age, weight, sex and height.

Random Forest

Next, I decided to test if the Random Forest classifier could improve upon the Decision Tree, my intuition being that it is just a more complex Decision Tree, thus it should be able to get a better prediction. I ran a 5 fold cross validation for the Random Forest, and tested for the following hyper-parameters

- min_samples_leaf [1,2,3,4,5,10]
- n_estimators [10,20,30,40,50,100]
- max_depth [3,4,5,6,7,10,15]

Results from running this cross validation is as follows

- The best Hyper Parameters are {'max_depth': 3, 'min_samples_leaf': 4, 'n_estimators': 10}
- The Best Random Forest Classifier has a accuarcy score of 0.8361696723393447

After looking at these results, it seems at first glace that Random forest was able to sligtly improve upon the Decision Tree.

Comparing the Random Forest With Just the Decision Tree

Table to Compare Models

	Decision Tree	Random Forest
Min Samples Leaf	10	4
Max Depth	3	3
Number of Estimators	1 (Decision Tree)	10
Best Accuracy Score	0.8359156718313436	0.8361696723393447

After Running a grid search on each model to find the best hyperparameters for both, it turns out the the Random Forest took different hyperparameters than the Decision Tree. The difference in number of estimators has to do with a Random Forest being a collection of Decision Tree's thus the Decision Tree is capped at 1. The only other difference was the Min Samples Leaf, with the Decision Tree using 10, instead of the Random Forest's 4. This probably means that the Random forest was able to make smaller leafs that containted a specific competitor (One person who competed multiple times) who was probably better or worst than the average person. In my opinion this smaller value of 4 would hurt my models prediction of new Olympians because I don't want the model to learn specific people, as new olympians might not compete the same way as the person that was effectivly 'Hardcoded' into the tree.

Even tho the Random forest had a higher accuarcy it was only on the magnitue of 0.01, which seems within the margin of error, thus being negligable. We can conclude that the most effective way of predicting whether or not a weightlifter will place (get a medal) based on

- Age
- Weight
- Height
- Sex

is the Decision Tree classifier. My predicition for why the Random Forest did not improve the performance of the Decision Tree is that anything more complex than the Decision Tree (More Trees) Began to hardcode the weightlifters based on which person was generally a better weightlifter, instead the classifier I have made here seems to do better by keeping the Gini values low, and not hardcoding the specific weightlifters who were the best or were the worst.

In conclusion, I am confident that this Decision Tree classifer would do a good job at predicting future olympians on whether or not they will recieve a medal by their age weight and height, due to the fact that the Gini values are low, but not completely homogenous. With Weightlifting being heavily imfluenced by a persons weight and height, the assumption that many weightlifters that are closer in those respects would perform around the same in the competition. This is why I believe a Decision Tree Classifer model that I made here was the best at predicting medal recieving olympians