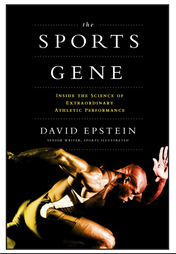
Classifying Olympic Athletes By Sport

Matt Dickenson
mcd31@duke.edu

Duke UNIVERSITY



Motivation

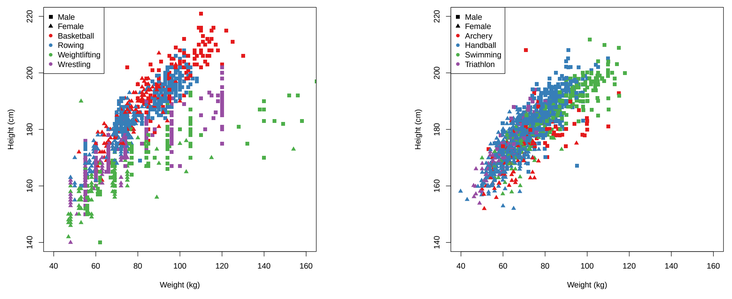
To what extent do environmental or biological traits determine sporting success? At the highest level of amateur sports–the Olympic games–we can notice differences in the physical characteristics of participating athletes across sports. Can these differences be exploited to classify individuals by sport or event given their physical attributes?

This project was inspired by a claim made by David Epstein, author of *The Sports Gene*. This claim is expressed in an interview with Russ Roberts:

> **Roberts**: [You argue that] if you simply had the height and weight of an Olympic roster, you could do a pretty good job of guessing what their events are. Is that correct?
> **Epstein**: That's definitely correct. I don't think you would get every person accurately, but... *I think you would get the vast majority of them correctly.* And frankly, you could definitely do it easily if you had them charted on a height-and-weight graph, and I think you could do it for most positions in something like football as well.

Data

Data was obtained for participants in the 2012 London Olympics from *The Guardian*. The original data consisted of 10,383 participants, which was reduced to 6,956 observations after data processing. The processed data was randomly split into training (n=3,520) and test (n=3,436). Athletes' height, weight, and age were used as features to predict their Olympic sport. Some sports exhibit relatively well-clustered features, whereas others are more scattered.

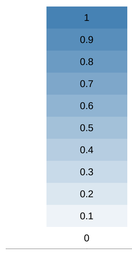


Methodology

Several machine learning methods were applied to this classification problem. Hierarchical clustering using Gaussian marginal likelihoods was performed using the `mclust` package.

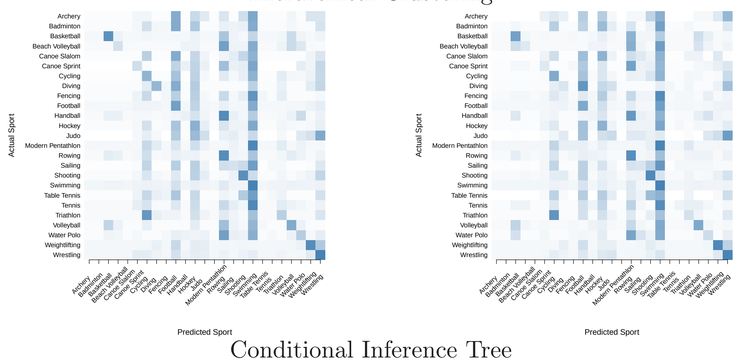
	Training Accuracy	Test Accuracy	Ratio
Hierarchical Clustering	.272	.271	.998
Conditional Inference Tree	.279	.219	.784
Evolutionary Tree	.292	.236	.807
Random Forest	.923	.244	.265
Neural Network	.280	.265	.949

The next column presents these results visually, with row-normalized observed frequencies. The figure at right serves as a legend for all heatmaps.

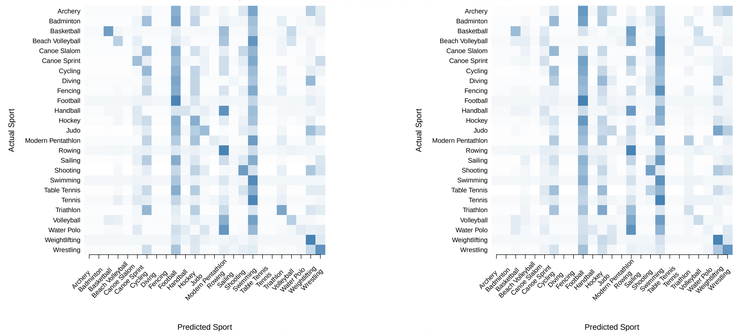


Results

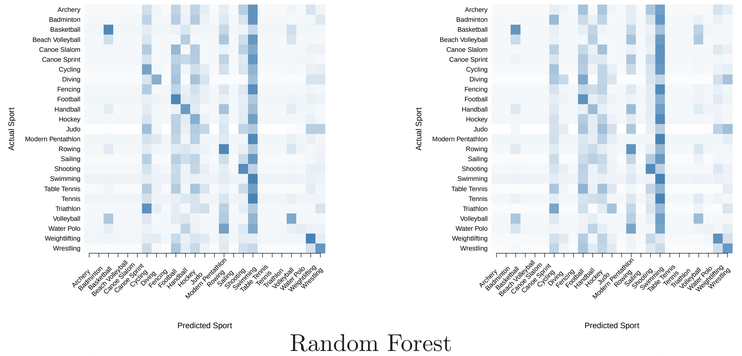
Hierarchical Clustering



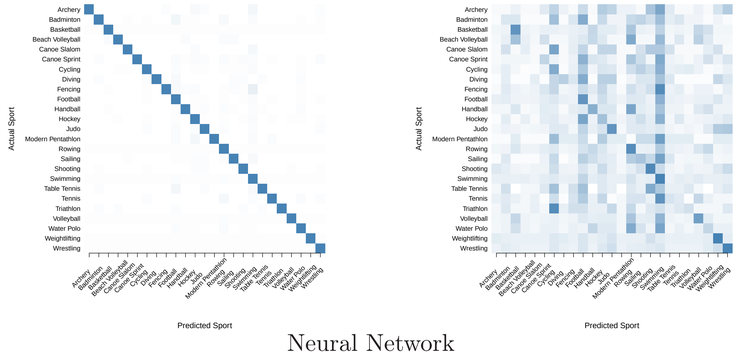
Conditional Inference Tree



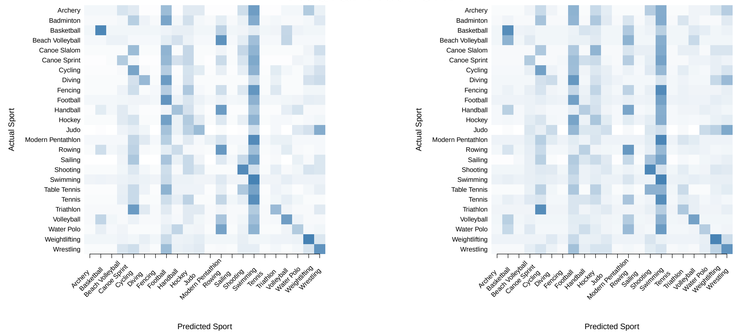
Evolutionary Tree



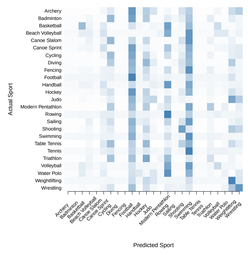
Random Forest



Neural Network



Results



Discussion

- Classifying athletes by sport can be achieved with moderate accuracy using only a few features
- Classification with a large number of categories is difficult
- Traits of athletes in some sports exhibit noticeable clustering, while other clusters are less distinct (multi-modal)
- Athletes in some sports have a well-defined body type, but athletes exhibit a wide range of physical features