Can Prior Olympics Medal Outcomes Predict Future Instances?

An Analysis With Logistic Regression and Gradient Boosting Trees

Sam Edds: samantha.dawson9@gmail.com

February 17, 2019

1 INTRODUCTION

We examine 120 years of Olympic History and Events data to better understand whether prior medal outcomes can predict future medal outcomes for a given country and event. We specifically focus on data from the 2000, 2004, 2008, and 2012 Summer Olympics to predict whether that country will obtain a medal for a specified event in 2016.

We recommend using a Gradient Boosting trees model, because it correctly predicts some events in which teams obtain a medal, compared to the Logistic model which does not. Based on the known outcomes of those events, our Logistic regression model correctly predicted obtaining a medal in 77.0% of cases. We predict almost completely correctly when a team-event will not obtain medal, but almost always incorrectly classify events in which teams do medal. Our Gradient Boosting trees model correctly predicted 79.1%, more accurately predicting medaled events, but less accurately predicting non-medal events. We best predict obtaining a medal for the United States, a country which has a large share of overall medals. Unsurprisingly, we found the most important variables in prediction are whether a team obtained a medal for that event in 2012, the count of said medals, and the specific event.

Overall, our prediction results indicate it is difficult to separate and distinguish outcomes in which a team obtains a medal for a given event.

2 EXPLORATORY DATA ANALYSIS

Before answering our question we explore our data to better understand the nuances, including missing data, number of observations, and other descriptive statistics. Our data features individual level event data for each Olympics game from the late 1800s to 2016. While we examined missingness and outliers for all of our data, our exploratory analysis will focus on those data pertaining to our medal prediction question, and additional analysis is in the code appendix.

We aggregate our individual level data to the event level by team and find 1,980 events by team have data for all of the 5 most recent Olympics. This is made up of 114 teams across 259 different Olympic events. Additionally at least one athlete for each team-event year report age, height, weight, and sex. We found only 44 teams that had missing height or weight data for all athletes at the team-event level, which was not concentrated on a specific team, event, or year, so we drop these (all were one athlete events).

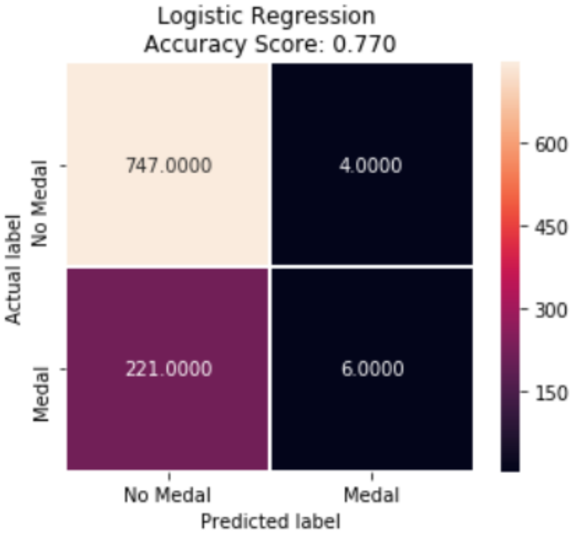
3 DATA DESCRIPTION

Our team-event level data measure the median, mean, and maximum age, height, weight of athletes for a given year. Additionally we record the number and sex of athletes that participate in the event, whether or not any medal was obtained, and a count of medals for that year. Finally, we include information about the event's sport, and where the team is from, including region and country name, for a total of 54 features.

When determining our holdout data we stratify by team, and hold out about half of any given team's data. This means we will predict 2016 outcomes for a total of 978 events, since many teams have a large number of qualifying events. We cross-validate, iteratively training and testing on medal outcomes for 1,002 events. Below we describe our specific methodology and detailed outcomes.

4 LOGISTIC REGRESSION AND OUTCOMES

We start with logistic regression to test whether 2016 medal outcomes can be predicted by 2000, 2004, 2008, and 2012 demographics and medal outcomes, for a given country and event. We use 5-fold cross validation, leaving out the intercept because we have many reference categories.



We obtain an accuracy of 77.0% where almost all instances of not obtaining a medal are correctly predicted. While there is a 99% True Negative rate, for 97% of events where a team obtained medal, the model incorrectly predicts those as having not obtained a medal (False Negative). This means, the model predicts almost every team-event as not obtaining a medal. The accuracy comes from correctly predicting events in which teams do not medal, so the overall model fit is very poor.

If we predicted at a higher level, for example, sport, we would likely have higher accuracy. Event-level is very specific, such as the 'Women's 200m Breaststroke', while sport, 'Swimming' is much less granular. It may be easier to predict yes / no a team obtained any medal, and potentially medal counts within a range because those data are less granular. Additionally, for specific events athletes often turn over from one Olympics to the next, so there is not always consistency and we would expect it is difficult to predict.

Because logistic regression struggled with fitting our data, we try prediction using trees with gradient boosting to see if that provides a better fit.

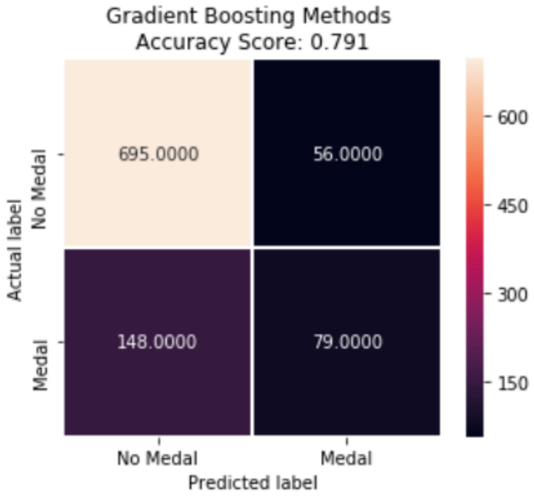
5 GRADIENT BOOSTING TREES AND OUTCOMES

We utilizing Gradient Boosting trees modeling because it builds each new tree, learning from itself. This method typically performs well for unbalanced data, and since we have many more non-medal events than medal events, we chose it.

We still utilize 5-fold cross validation, tuning our learning rate from 2.5% to 5.0% to 7.5%. We found, on average, the learning rate of 7.5% produced the best results on our training data. On our holdout 2016 data, we obtain an accuracy of 79.1%.

While there is a 7% False Positive rate, the True Negative rate is still 93%. This seems to be a better tradeoff for a 35% True Positive rate, and a 65% False Negative rate, compared to the Logistic model which has almost no correct predictions for team-events that obtain a medal. This is superior to the Logistic regression because medals are correctly predicted for at least some team-events, which could likely be augmented with more time.

We found the most important variables in predicting whether or not a team will obtain a medal for a given event to be whether that team medaled for the event in 2012, the count of medals for that event in 2012, and what is the specific event.



We examine our model in detail to understand patterns of correct prediction, compared to the holdout sample as a whole.We correctly predicted obtaining a medal in 69 different events. This means almost every event predicted correctly was a different event, because we correctly predicted 79 medal outcomes. We also found we correctly predicted obtaining a medal in 22 (of 32) sports.

Additionally, we found we correctly predicted obtaining a medal for 21 among 85 teams, so the correct medal predictions are more concentrated by team than sport or event. In particular, the United States had the most correctly predicted events (24), although they also had a large share of medals overall.

6 CONCLUSION

We recommend using the Gradient Boosting trees model because we found the trees method predicted some medal results correctly, while logistic regression did not. Gradient Boosting is more flexible and better with the unbalanced data than Logistic regression.

Overall we found there is trade off between correctly predicting some medals, which is incorrectly predicting some non-medal events. This indicates our data are similar to each other in a number of ways that makes it difficult to separate and distinguish medal and non-medal outcomes.

Augmentations to our modeling would include adding variables on whether any of the athletes are returning and their average rank, changing the number of years used to predict, adding more features from outside data (if possible), and examining additional methods beyond gradient boosting trees and logistic regression.