Olympic Games Analysis

Tansel Simsek, Alessandro Quattrociocchi, Leonardo Masci, Juan Mata Naranjo

Abstract
This report proposes several type of regression models (both linear and logistic) and aims to answer different questions related to 120 years of Olympic history. The proposed models will try to predict what sport the authors of this paper will most likely succeed in, check the existence of the home advantage, the effect of macro-economic factors in winning medals and the optimal features of the athlete in the upcoming Olympics. The first model aims to identify which sport we are more likely to succeed, in particular, the best results were obtained using the random forest model with f1-score significantly higher to a dummy classification model, while the probability of winning a medal has been performed applying a Logistic Regression over several feature combinations. The second model uses a Logit Regression and Logistic Regression to identify any links that demonstrate home-advantage. Despite other papers and works, performance measurements are not found satisfactory to prove the home advantage with our data. Some additional macro-economic features are considered for the third model, proving that a greater gross domestic product (GDP) correlates with winning more medals, while the human development index (HDI) and the population size do not have as much of an impact. The fourth model shows the trend over the years of some features such as Height, Weight and Age of the Olympic medal winners and allowed us, within the limit of the linear model, to approximate the ideal features of future athletes.

Keywords
Logistic Regression — Linear Regression

1. Introduction

This report proposes several type of regression models and aims to answer different questions related to 120 years of Olympic history. The proposed models will try to predict what sport the authors of this paper will most likely succeed in, check the existence of the home advantage, the effect of macro-economic factors in winning medals and the optimal features of the athlete in the upcoming Olympics.

2. Proposed method explained

Given we have 4 different questions we want to answer we will tackle each separately, naming each as model1, ..., model4 respectively:

Model1: This model will try to predict the ideal sport an athlete should compete in based on features such as age, height, weight, medals won by their country. This first task is performed by inspecting 5 different models, specifically a Decision Tree Classifier, Random Forest, Multinomial Logistic Regression (one-vs-rest), K-Means and Support Vector Machine [1]. We have decided not too look into the last of these models in much detail since the computational time is very high and the first results were not very promising. These models have been evaluated using a 5-fold cross validation approach over the training data to obtain the best hyper-parameter for each model. Once the ideal hyper-parameter was identified the models were tested against some unseen data (test data) providing the champion model evaluated over the F1-Score. Once the ideal sport for a given athlete was identified, given the same attributes, the probability of winning a medal was computed using a simple Binary Logistic Regression model. Similarly to the previous case, a set of feature combinations (first, second and third order degree polynomials) were tested over the training data and the features combination with highest F1-Score were then evaluated over the test data. Each sport was assigned a different "ideal" feature combination.

Model2: This model attempts to understand if the host effect can be considered as advantage or not. Compared to the other models, data pre-processing was not as relevant since the model considers team, sport and host variables which does not contain any missing data. In addition to that, selection of both independent and dependent variables behaved differently. Two different approaches are tested to confirm if home-advantage exists in Olympic games. Firstly, logistic regression applied by taking independent variables as both sport and team, only sport and only team. In this approach, the average medals won by the team in the specific sport while host and in overall are calculated. If the average medals when host is greater than average medal in general, the data is labeled as advantage (not advantage otherwise). In the second approach, logit-regression, the response variable was calculated as the number of medals won each year for the specific team over the total overall medals. Explanatory variables are created as pre-host, host and post-host as Nevill, Balmer and Edward proposed [2].

Model3: This model checks if there is any correlation

between the wealth and size of a country and the amount of medals won in the Olympic Games. This is done through a linear regression model which required merging a number of different data sets, one for each of the features taken into account (GDP, HDI and population size). A linear regression model is then created on the basis of the data obtained. This does not yield satisfactory results ($R^2 < 0.3$ for all three features), while considering one Olympic Game at a time does.

Model4: The final model describes the evolution of quantities such as weight, height and age over time. In particular, the Winter and Summer Olympic Games were separated and the sports with the most medals were taken into account. For each edition of the games since 1960, the mean and variance of each size was calculated for that year, which allowed the representation through a scatterplot and, from there, the regression on the plots. For each sport and each season, the four most important events were analysed and a linear regression model was applied on them. Each regression on the physical features of the athletes was evaluated using the R^2 coefficient.

3. Dataset and Benchmark

We started working on a kaggle data set comprising data related to athletes competing in the modern Olympic Games and further extended it with data sets relating to the extra features we wanted to analyse (GDP, HDI and population size). Several data pre-processing steps were performed to improve the accuracy of our models. For example, the data on athletes before 1960 was missing a lot of information and the HDI was only introduced on 1990. In addition, at times we only took into account information regarding athletes who had won at least a medal or only considered the number of medals won by a country. Since all the models had different goals and examined different features, each one had to start with some data-cleaning of its own. These are all illustrated in the corresponding notebooks.

We did not compare our analysis with any kaggle competitors, since most of them did not have the same goals as us. Therefore, our benchmark mainly consisted in splitting the dataset in a training and testing set. The former was used to build the models, while the latter to check the accuracy of most models.

4. Experimental results

Similarly to the section 2:

Model1: The best model to predict what sport an athlete belongs to is by far the Random Forest Classifier with an F1-score of approx. 0.35, for both male and female athletes (accuracy of approx 0.35 as well, and a higher precision than recall). This is much better compared to a dummy classifier with an accuracy of 0.03. The Logistic Regression used to predict medal winners had better performance in general but it's performance depended a lot on the sport (sports such as Basketball, Volleyball had a very high F1-score, approx. 0.7, while sports such as Athletics, Boxing, had F1-scores of approx 0.1).

Model2: In the first approach where logistic regression is used, none of the models performed good at predicting of the test data (.23, .81, .09, .82, .17, .83 f1-score for predicting 0 and 1 accordingly for model 2.1.1 to model 2.1.3). Models mostly estimated observations as advantage. Since unbalanced support rate is inspected while data labeling, the results are aligned with this problem. In the logit-regression model, despite p-values' showing significance, the pseudo R-squared value was found to be infinity, proving that the chosen model does not follow the trend of the data. Having investigated 2 different approaches, it can be stated that there is no enough evidence to say host effect brings any advantage to the team.

Model3: The strongest correlation between number of medals won by a country and its macro-economic features, was the GDP. In this case, for the summer games, R^2 was consistently over 0.6, while population and HDI both had a R^2 of around 0.1. On the other hand, for the winter games R^2 dropped at around 0.3 for the GDP and 0.02 for population size, while HDI remained consistent. This is in line with what we were expecting, since some very small northern countries win a lot of medals in those competitions.

Model4: The application of linear regression on this data, after careful pre-processing, yielded encouraging results. There are many sports for both winter and summer, male or female, that reported a score of R^2 close to 1. Among the best is Ice Hockey Men's with 0.921, 0.874, 0.508 (height, weight and age), or Swimming Women's 4x400 Freestyle Relay with a score of 0.767, 0.688, 0.552. More simulations and predictions are presented in the model notebook.

5. Conclusions and Future work

Based on the previously presented results we can confidently affirm that despite all the hard work and training of the Olympic competitors, the Olympic Games have a set of events which can be predicted and estimated. We have seen that athletes are most likely to win in countries where the living conditions are best, future athletes will most likely have a predictable set of attributes, home advantage cannot be properly justified with these data sets and that we can somewhat predict what sport an athlete should compete in and how high his/her chances are of actually winning a medal. Despite all of these insights we consider that this analysis can be further extended and improved:

1. Much better predictions could be obtained if we extended the number of features (e.g. years since athlete practices sport, age when athlete started to practice sport, etc.). The features we have worked with are not sufficiently discriminant to do very strong predictions.

2. Inspect additional models to see how these would compare to the already inspected models (e.g. higher order regression models, etc.).

3. Due to time and knowledge constraint we have not included as many model visualizations as possible.

References

[1] Kaggle Competitor. *What Sport Will You Compete In?* Kaggle, 0th edition, 2016.

[2] Alan M Nevill, Nigel J Balmer, and Edward M Winter. Congratulations to team gb, but why should we be so surprised? olympic medal count can be predicted using logit regression models that include 'home advantage', 2012.