FORECASTING THE OLYMPIC MEDAL DISTRIBUTION – A SOCIOECONOMIC MACHINE LEARNING MODEL

Christoph Schlembach	Sascha L. Schmidt	Dominik Schreyer	Linus Wunderlich

This is a summary of a published paper, that should be **referred** as follows:
Schlembach, C., Schmidt, S. L., Schreyer, D., & Wunderlich, L. (2022). Forecasting the Olympic medal distribution – A socioeconomic machine learning model. Technological Forecasting and Social Change, 175, 121314. DOI: 10.1016/j.techfore.2021.121314.

ABSTRACT

Research question: In this paper, we forecast the number of Olympic medals for each nation, which is highly relevant for different stakeholders. Ex ante, sports betting companies can determine the odds while sponsors and media companies can allocate their resources to promising teams. Ex post, sports politicians and managers can benchmark the performance of their teams and evaluate the drivers of success.

Research methods: We apply machine learning, more specifically a two-staged Random Forest, to a dataset containing socioeconomic variables of 206 countries (1991–2020).

Results and findings: We evaluate the forecast accuracy based on five different metrics: (1) Number of correct forecasts of total (60%), (2) non-zero (17%), and (3) zero (95%) medals, as well as (4) 95% confidence intervals +/- 2 medals (89%), and (5) absolute deviation for top-17 nations (122 medals). We also estimate that COVID-19 hardly impacts the number of medals among the top-20 nations.

Implications: For the first time, we outperform the more traditional naïve forecast for four consecutive Olympics between 2008 and 2020.

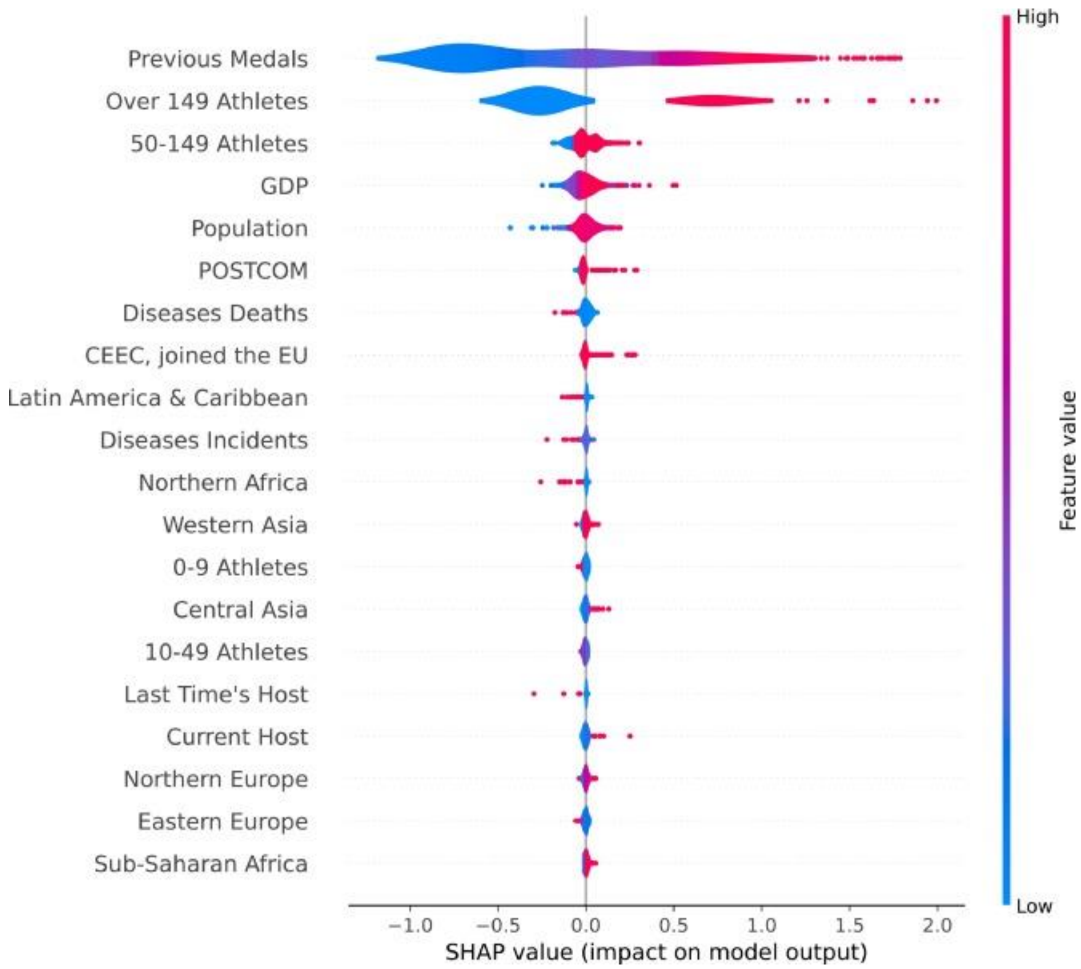
STUDY HIGHLIGHTS

- Accurately forecasting Olympic performances has gained considerable research interest over the last decades.
- Such forecasts, typically medal forecasts, are necessary to provide both a government and its citizens with a benchmark against which they can evaluate the nation's Olympic success ex-post.
- Despite constant methodological improvements, a naïve forecast still outperforms previous forecasting approaches regularly.
- In our study, we apply machine learning to forecast the Olympic medal distribution.
- To the best of our knowledge, our approach is the first to consistently beat the naïve forecast for the 2008, 2012, 2016, and 2020 Games.
- Besides the naïve forecast, we benchmark against seven other models from five different papers.

IMPLICATIONS

- Ex post, sports politicians and managers are facing the challenge to judge the performance of their teams. Our forecast allows them to detect over- or underperformance against what was to be expected ex ante more precisely.
- We believe that there are two ways to improve the model's performance further: First, by including additional socioeconomic features (e.g., investments in sports infrastructure, athlete-specific characteristics such as an athlete's age or disciplines); second, by considering COVID-specific features, such as the number of canceled national sports events.
- Besides working on model-specific adjustments, scholars can build upon our research within the scope of new applications in sports forecasting.
- Besides working on model-specific adjustments, future research could build upon our research within the scope of new applications in sports forecasting.

Figure 1. Feature importance of the two-staged Random Forest



To understand the main drivers behind the forecasts better, we use the explanatory Shapley Additive Explanations (SHAP) value. The SHAP value of one feature describes the change in the expected model prediction when conditioning on that feature; starting from the base value, i.e., the prediction without knowledge of any features, the combination of all SHAP values then leads to the full model forecast.

In Figure 1, we depict the most relevant features. Here, one dot represents one observation in the training data, i.e. one Olympia-nation-combination. To make the assessment of feature importance more intuitive, we ranked all variables in descending order according to their feature importance. Then, the horizontal location shows whether the effect of the value is associated with a higher or lower prediction. Further, the color then indicates whether that variable is high (red) or low (in blue) for each observation. For instance, a high medal count at the previous Olympics ("Previous medals") has a high and positive impact on the number of medals at the 2020 Tokyo Olympics held in 2021. The "high" comes from the red color, and the "positive" impact is shown on the X-axis.

More information – download more research papers
© www.SPLISS.net