Technological Forecasting & Social Change 175 (2022) 121314

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Forecasting the Olympic medal distribution – A socioeconomic machine learning model

Christoph Schlembach[a], Sascha L. Schmidt[a,c,d], Dominik Schreyer[a,*], Linus Wunderlich[b]

[a] Center for Sports and Management, WHU - Otto Beisheim School of Management, Campus Düsseldorf, Erkrather Str. 224a, 40233 Düsseldorf, Germany
[b] School of Mathematical Sciences, Queen Mary University London, Mile End Road, E1 4NS London, United Kingdom
[c] CREMA – Center for Research in Economics, Management and the Arts, Südstrasse 11, CH-8008 Zürich, Switzerland
[d] LISH – Lab of Innovation Science at Harvard, 175 N. Harvard Street, Suite 1350, Boston, MA 02134, United States of America

ARTICLE INFO

Keywords:
Olympic games
Medals
Sports
Forecasting
Machine learning
random forest

ABSTRACT

Forecasting the number of Olympic medals for each nation 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. We apply machine learning, more specifically a two-staged Random Forest, to a dataset containing socioeconomic variables of 206 countries (1991–2020). For the first time, we outperform the more traditional naïve forecast for four consecutive Olympics between 2008 and 2020.

1. Introduction

Forecasting based on socioeconomic indicators has a long tradition in academia, in particular in the social sciences. As Johnston (1970, p. 184) noted early, validating the associated "social projections [empirically] may serve to generate appropriate policies or programs whereby we can avoid the pitfalls which would otherwise reduce or eliminate our freedom of action." Consequently, ever since, there have been endeavours to predict the future, exemplary in the field of economics (e.g. Modis, 2013), public health (e.g. Puertas et al., 2020), civil engineering (e.g. Kankal et al., 2011), ecology (e.g. Behrang et al., 2011) or urban planning (e.g., Beigl et al., 2004).

In the economic literature, in particular, accurately forecasting Olympic performances has gained considerable research interest over the last decades (cf. Leeds, 2019), primarily because 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. For a government, often investing heavily in athlete training programs to enhance the probability of a nation's Olympic success (cf. Humphreys et al., 2018), such an assessment is pivotal because it allows them to understand better whether the application of funds, i.e. the taxpayers' money, to their National Olympic Committee (NOC) is productive. For instance, as major sporting events such as the Olympic Games are often associated with increasing both national pride among their citizens (cf. Ball, 1972; Grimes et al., 1974; Allison and Monnington, 2002; Hoffmann et al., 2002; Tcha and Pershin, 2003) and their willingness to begin engaging in sporting activities (cf. Girginov and Hills, 2013; Weed et al., 2015), thereby reducing long-term healthcare costs, a government might be willing to raise funds if their NOC meets (or even exceeds) the medal forecasts. In contrast, because Olympic success is a well-known antecedent of civic willingness to support funding a government's elite athlete training programs (Humphreys et al., 2018), falling behind the predictions might motivate a government to increase the pressure on the NOC, not least by reducing future funds.

Likewise, accurately forecasting the Olympic success is highly relevant for many different non-governmental stakeholders. For instance, sports betting companies rely on precise estimates to determine their odds, while both the media and Olympic sponsors must allocate their resources to promising teams and their athletes. Thus, analysing the Olympic Games empirically has become a relevant field of research, both, with a focus on forecasting (e.g. De Bosscher et al., 2006) and

[*] Corresponding author.
E-mail address: Dominik.Schreyer@whu.edu (D. Schreyer).

https://doi.org/10.1016/j.techfore.2021.121314
Received 21 June 2021; Received in revised form 16 September 2021; Accepted 23 October 2021
Available online 12 November 2021

beyond (e.g. Streicher et al., 2020).

Since the first contribution by Ball (1972), the quality of such Olympic forecasts has steadily improved for two reasons. First, those authors interested in predicting a nation's Olympic success have successively begun employing new estimation techniques. Second, over time, the predictive power of models has gradually increased as authors operating in the field have explored diverse, increasingly extensive data sets.

Since Ball (1972) pioneered with a correlation-based scoring model, forecasting models have continuously become more sophisticated. Initially, as we exemplary show in Table A1 in the appendix, most authors referred to the use of ordinary least squares regressions (OLS), as it delivered results that were easy to interpret (e.g. Baimbridge, 1998; Condon et al., 1999; Kuper and Sterken, 2001). However, a significant challenge when predicting Olympic medals is to reflect the large number of nations without any medal success properly. As the incorporated exponential function punishes small predicted numbers of medals, some authors (e.g. Lui and Suen, 2008; Leeds and Leeds, 2012; Blais-Morisset et al., 2017), then, moved to Poisson-based models (i.e., a Poisson model, negative binomial model), to tackle this methodological problem. However, until today, because the dependent variable, typically the number of medals, has zero as lower bound, most authors have employed Tobit regression to predict Olympic success (e.g. Tcha and Pershin, 2003; Forrest et al., 2015; Rewilak, 2021). Only recently, employing a two-step approach, estimating the probability of winning any medal before determining the exact number of medals in case of success, became more popular. In particular, both Scelles et al. (2020) and Rewilak (2021), employing a Mundlak transformation of the Tobit model, could, again, increase the prediction accuracy with their respective Hurdle models. In contrast, other authors (e.g. Hoffmann et al. (2002)), have circumvented the underlying methodological problems by splitting their sample into nations that did and did not win any medals in the past, while few authors have employed alternative methodological approaches.[1] However, despite all these methodological improvements, a naïve forecast still outperforms these previous forecasting approaches regularly.

Somewhat similarly, during the last years, authors have significantly increased the data sets used for medal forecasting in three ways. First, by increasing the level of granularity beyond country-specifics; second, by including more years; and third, by exploring additional independent variables.

As a common way to incorporate more granular data, and thus to increase the forecast accuracy, some authors considered predicting the Olympic success by focussing on different sports (e.g. Tcha and Pershin, 2003; Noland and Stahler, 2016a; Vagenas and Palaiothodorou, 2019), sometimes even exploring data on the level of the individual athlete (Condon et al., 1999; Johnson and Ali, 2004). Due to the increasing relevance of gender studies, other authors have begun differentiating their data sets by gender (Leeds and Leeds, 2012; Lowen et al., 2016; Noland and Stahler, 2016b). As Garcia-del-Barrio et al. (2020) report that gold medals generate more media attention than silver and bronze medals, also more granular forecasts including the medal type seem appealing. Noticeably, these more nuanced empirical approaches are certainly important to answer very specific questions. Yet, macro-level models, in contrast, have the "advantage of averaging the random component inherent in individual competition [leading to] more accurate predictions of national medal totals" (Bernard and Busse, 2004, p. 413). Thus, macro-level analysis remains a frequently used approach in Olympic medal forecasting.

A different approach to potentially increase the forecast accuracy is to expand the data set's temporal dimension. As such, some authors have incorporated one hundred years of Olympics and more in their models (e.g. Baimbridge, 1998; Kuper and Sterken, 2001; Trivedi and Zimmer, 2014). However, more recently, most authors seem to limit the number of events under investigation for three particular reasons. First, specific incidents such as "large-scale boycotts as occurred at the 1980 Moscow and 1984 Los Angeles Games" (Noland and Stahler, 2016b, p. 178) and "the East German doping program [, which] was responsible for 17 percent of the medals awarded to female athletes" (Noland and Stahler, 2016b, p. 178) in 1972 skewed the medal count in the past. Second, international borders shifted particularly in the course of two World Wars and the breakdown of the Soviet Union; only since the 1990s nations remained relatively stable (Forrest et al., 2017). Third, the significance of variables changed over time, such that, for instance, a potential host effect, might play a different role in times where international travel has become a part of our daily lives (Forrest et al., 2017).

Finally, another way to augment a data set and, thus, to improve the accuracy of a forecast is to incorporate additional independent variables. Early, Ball (1972) found that "[g]ame success is related to the possession of resources, both human and economic, and the centralized forms of political decision-making and authority which maximize their allocation" (p. 198). Particularly, the extraordinary Olympic performance of countries with a certain political system, at that time the Soviet Union, has been confirmed by research until today (e.g. Scelles et al. (2020)). Other authors (cf. Kuper and Sterken, 2001; Hoffmann et al., 2002; Bernard and Busse, 2004; Johnson and Ali, 2004) found that hosting the Games increases the expected number of medals, among others due to an increased number of fans and reduced stress due to international travel. Maennig and Wellbrock (2008), Forrest et al. (2010) and Vagenas and Vlachokyriakou (2012) extended this finding and concluded that such a host effect already starts four years before the Olympic Games and, surprisingly, lasts until the subsequent Games. On a similar note, authors found a continuous over-, respectively underperformance of nations, such that lagged medal shares significantly improve the prediction accuracy (cf. Bernard and Busse, 2004).[2]

In addition, it is important to mention that many scholars experimented with additional variables such as the climate (Hoffmann et al., 2002; Johnson and Ali, 2004), public spending on recreation (Forrest et al., 2010), health expenditure, growth rate, unemployment (Vagenas and Vlachokyriakou, 2012), and income (Kuper and Sterken, 2001). In general, there are mixed findings on most of these variables and only few are available in public as comprehensive data sets. In this regard, De Bosscher et al. (2006) conducted a meta-analysis of variables predicting sportive success, even beyond Olympic Summer Games, and found that both the Gross National Product and the country's population "consistently explain over 50% of the total variance of international sporting success" (p. 188). It is, therefore, not surprising that these two variables, in particular, have been used by most authors in the past 50 years. As such, also taking the potential issue of multicollinearity from exploring too many distinct, though potentially related, socioeconomic variables into consideration, forecasting models should not be augmented infinitely.

Given the high policy relevance and academic attention of Olympic forecasting, it is somewhat surprising that the potential of machine learning in detecting hidden patterns and, thus, improving forecasting accuracy has not yet been exhausted in this context. However, this methodology has recently received an increasing level of popularity in a sports context, e.g. in football (Baboota and Kaur, 2019). Particularly the Random Forest approach often delivers excellent results, for instance, in forecasting football scores (Groll et al., 2019) or horseracing outcomes (Lessmann et al., 2010). As acknowledged by Makridakis

[1] For instance, Condon et al. (1999) employed neural networks for one Olympiad. Other authors used (single-step) binary regression models (Probit, Logit) while using more granular data (Johnson and Ali, 2004; Andreff et al., 2008; Noland and Stahler, 2016b).

[2] This also explains why none of the published models repeatedly outperformed a naïve forecasting model assuming the number of medals from the preceding Games for the upcoming Games as well.

et al. (2020), statistical knowledge can be applied in the world of machine learning as well.

As such, in this study, we translate the proven concept of the Tobit model to machine learning by using a two-staged Random Forest model to predict Olympic performance. In that way, we identify the first model to consistently outperform a naïve forecasting model, in four consecutive Summer Games (2008, 2012, 2016, and 2020) by about 3 to 6 percentage points. On a side note, we thus also improve the forecasting accuracy presented in more recent work on the potential determinants of Olympic success (Scelles et al., 2020) by roughly 20 percent.[3]

The remainder of our manuscript is structured as follows: After motivating the variables used in the model and introducing the concept of a two-staged Random Forest, we evaluate the quality of the forecast, present an estimate for Tokyo 2020, and discuss the implications of COVID-19. We conclude with a summary, ex-ante and ex-post consequences of the prediction, and an outlook for further research.

2. Material and methods

We forecast the number of medals in the Tokyo Olympic Games for each participating nation based on a two-staged Random Forest. It is, however, important to note that, as part of this exercise, we also quantify the impact of COVID-19 on the expected Olympic medal count based on the independent variables (i.e., features) national GDP, incidents of and deaths from lower respiratory diseases. In this section, we motivate the underlying variables, explain the concept of a two-staged Random Forest, and describe the forecasting process.

2.1. Variables

2.1.1. Dependent variable (output variable)

The number of Olympic medals represents economic and political strength and promotes national prestige (Allison and Monnington, 2002). De Bosscher et al. (2008, p. 19) acknowledge the figure as “the most self-evident and transparent measure of success in high performance sport”. As most scholars (e.g. Andreff et al., 2008; Scelles et al., 2020), we define the number of medals as dependent variable without distinguishing between gold, silver and bronze medals. Although Garcia-del-Barrio et al. (2020) report that gold medals generate more media attention than silver and bronze medals, Bernard and Busse (2004) acknowledge that models, that do not distinguish between different medal types, produce more accurate forecasts as they smoothen random effects. Following Choi et al. (2019), who note that a log-transformation can reduce the skewness and, thus, improve prediction accuracy in machine learning, we take the logarithm of the number of medals, which reduces the (right-) skewness from 3.2 to 0.4 (i.e., only among non-zero medals due to the definition of logarithm). As the independent variables do not change at the same rate as the Olympic medal totals, we cannot expect an exact match between forecast and actual medals at stake. Thus, we need to rescale the prediction to the number of scheduled events times three (assuming no double bronze). Further, rounding is necessary to get natural numbers.

2.1.2. Independent variables (features)

The predictive power of GDP for sportive success is widely accepted in Olympic medal forecasting (cf. Bernard and Busse, 2004) and robust across both geographies (cf. Manuel Luiz and Fadal, 2011) and sports (e.g. Klobučník et al., 2019).[4] Several reasonable explanations include that richer nations invest higher sums in sports, provide more extensive sports offerings and cater for a better overall fitness among the population (De Bosscher et al., 2008). Due to limited data availability and high allocation complexity on a more granular level, aggregate figures, such as GDP, have become a de facto standard in academia (De Bosscher et al., 2006; Manuel Luiz and Fadal, 2011). To account for the character of the Olympic Games as a competition, we normalize the GDP (purchasing power parity in international dollars) to reflect the share of a nation in the global GDP as feature.

Besides GDP, the population of a nation is a well-established predictor of Olympic medals (Bernard and Busse, 2004; De Bosscher et al., 2008); that is, larger countries have larger resources of potential medal winners (Bernard and Busse, 2004). As the number of world-class athletes in a country, however, is exhausted at some point and population alone does not lead to more medals anymore,[5] we take the logarithm of the population which grows slower than a linear function.

We reflect the number of participating athletes in the model: Scelles et al. (2020) suggest the use of categorical variables for the number of athletes. Here, the rationale is that the final number of competitors is generally not known at the time of forecasting. Furthermore, the categories suggested by them have rarely changed in the past. As an example, Afghanistan has always sent between zero and nine athletes since 1992. Before forming these groups, we count athletes that started in multiple disciplines multiple times as their chances to win a medal multiply.

While existing research confirms an impact of specific socioeconomic variables on the number of medals in the Olympic Games, the connection of public health crises and sportive performance was not to be presumed before the COVID-19 crisis. Yet, this pandemic did not only lead to the postponement of the Games to 2021 (International Olympic Committee, 2020) but also affected the athletes’ preparation (Mohr et al., 2020; Mon-López et al., 2020; Wong et al., 2020), as well as the funds available in the sports industry (Hammerschmidt et al., 2021; Horky, 2021; Parnell et al., 2021). We reflect the impact of COVID-19 via incidents of and deaths from lower respiratory diseases, as well as GDP. The rationale here is, that incidents and deaths serve as a proxy for the sanitary conditions in a country; increased figures represent a higher risk of infection for athletes and staff, respectively a higher risk of complications during a medical treatment. Beyond direct health risks, the features also reflect training restrictions due to political countermeasures against a disease (e.g. a lockdown, obligation to wear masks, etc.). In short, both features have an impact on the preparation for

[3] The significant increase in predictive accuracy is based on two effects: First, Scelles et al. (2020) build on a generalized linear model in form of a Hurdle, respectively Tobit, model. We, in contrast, apply a two-staged Random Forest algorithm taking into account more complex, non-linear interactions. Second, it is often argued that the time to prepare an Olympic team is four years (cf. Forrest et al., 2010; Scelles et al., 2020). This would imply that, ideally, only socio-economic data until 2016 should be used to predict the Tokyo 2020 results. However, Stekler et al. (2010) evaluate different sports forecasting methodologies and find that more recent data generate better results. Thus, we include data until 2020 in our model to overcome this issue, which is even amplified by the WHO’s decision to declare COVID-19 a pandemic (Cascella et al., 2020) and the subsequent postponement of the Games to 2021 (International Olympic Committee, 2020).

[4] Even though the extent research on predicting Olympic success typically suggests that the total GDP “is the best predictor of national Olympic performance” (Scelles et al., 2020, p. 698), we have further experimented with a number of both alternative and additional independent variables capturing a nation’s economic status. For instance, to proxy the degree of inequality and poverty in a certain country, we have also added information based on the Human Development Index (HDI) to our model, that is, a statistic composite index of life expectancy, education, and per capita income indicators, which are frequently employed to rank countries into four tiers of human development. However, we did not find any improvements in prediction accuracy. These additional results are available from the first author upon request.

[5] For instance, in India, the world’s second largest country based on population, the significant growth of population between 1992 and 2016 (CAGR: 1,58%) was hardly converted into Olympic medals; while India won zero medals in 1992, the number did not increase significantly until 2016 (two medals). As a comparison, China, a country with a similar population, won 70 medals in 2016. For an in-depth analysis of the Olympic performance of India, please refer to Krishna and Haglund (2008).

Olympia and the subsequent sportive success. The change in GDP addresses available funds in the sports industry in general and devoted to the National Olympic Committees in particular. Thereby, adjusted investments in talent development have a mid-term effect, while the direct preparation for the Olympic Games (e.g. availability of training material, travel comfort, etc.) takes effect in the short term. We categorize incidents and deaths in quintiles to limit the effect of potential outliers. The broad availability of data allows us to create a synthetic "no COVID-19″ scenario by eliminating COVID-19 incidents and deaths, and by leveraging GDP forecasts made before the beginning of the pandemic; hence, we can quantify the impact of COVID-19 on Olympic medals.

Already Ball (1972, p. 191), in his seminal contribution, mentioned that "hosts [of Olympic Games] are more successful, at least in part because of their ability to enter larger than usual teams at relatively low financial expenditure". However, in the broader home advantage literature authors typically tend to link a host's home advantage to three different, alternative mechanisms (e.g. Singleton et al., 2021). First, reduced travel fatigue (e.g. Courneya and Carron, 1991). Second, venue familiarity (e.g. Pollard, 2002). Third, the influence of supportive home spectators on either athletes (e.g. Ferraresi and Gucciardi, 2021) or referees (e.g. Bryson et al., 2021). Because all of these potential explanations could apply to the Olympic Games (c.f., Balmer et al., 2003), we include a categorical variable for past, current, and future host countries.[6]

Bernard and Busse (2004) detected, that Soviet countries outperformed their expected medal success on a regular basis due to the essential role of sports in the communist regimes. Starting early on and combining competitive sports and education was an essential component in their strategy (Metsä-Tokila, 2002). Reflecting such peculiarities in political systems, we use the trichotomy in capitalist market economies, (post-) communist economies and Central Eastern European countries, that joined the EU, as refined by Scelles et al. (2020).

Further, geographic characteristics determine the capabilities of succeeding in a given sport because of culture, tradition and climate (Hoffmann et al., 2002). Subsequently, we use 21 regions as defined by the United Nations, Department of Economic and Social Affairs (2020) as categorical independent variable.

Finally, as recommended by Scelles et al. (2020) and Celik and Gius (2014), the number of medals in the preceding Olympics (non-logarithmic, as the value zero can occur) is added to the model as it significantly improves the predictive power. This suggests that there are some unconsidered country-specific factors, which may "include a nation's athletic tradition, the health of the populace, and geographic or weather conditions that allow for greater participation in certain athletic events" (Celik and Gius, 2014, p. 40).

We display the descriptive statistics of the numerical variables used in the model in Table 1 and list ordinal and categorical variables in Table 2. As a rule of thumb, Stekler et al. (2010) find that more recent data generate better results in sports forecasting. Thus, we leverage data from one year prior to the Olympics to make a forecast. Thus, being able to retrieve data points of 206 countries between 1991 and 2020, we can feed our models with 1379 country-year observations.

2.2. Data pre-processing

Data pre-processing is a vital step to ensure accurate forecasts (Wang et al., 2018; Chen et al., 2019). Here, we perform three steps: First, mapping of nations; second, inter-/extrapolation; and third, regional benchmarking.

2.2.1. Mapping of nations

As Olympic teams according to the definition of the International Olympic Committee do not necessarily match the country list in other data sources, we need to (dis-) aggregate socioeconomic data to adequately represent Olympic teams, e.g. by adding the population of Anguilla, a part of Great Britain, to the British population as accounted in the data source. In 2016, we attribute nine International Olympic Athletes (IOA) winning two medals to Kuwait based on their nationality. The Unified Team (EUN) represents Russia being banned from the Games because of doping (Hermann, 2019). A "neutral" team called "Russian Olympic Committee (ROC)" participated in 2021 again; our prediction applies to athletes from Russia regardless the name of their team. We split the athletes (269) and medals (7) of the former Czechoslovakia into Czech Republic (178 / 5) and Slovakia (91 / 2) based on their respective population; this allows adequate forecasts for the two nations that emerged from Czechoslovakia. The Refugee Olympic Team (12 athletes in 2016) has not won a medal yet. Hence, we assume a constant forecast meaning that there will be no medals in 2021 either.

2.2.2. Inter- / extrapolation

We obtain missing data points in a specific year by inter- / extrapolation, which is a common approach when pre-processing data (e.g. Christodoulos et al., 2010; Chen et al., 2019). This concerns the four features *Diseases Deaths, Diseases Incidents, GDP,* and *Population.* While we always interpolate linearly, we extrapolate tailored to the respective features, in a way that we estimate the actual values in a senseful way: For the features *Diseases Deaths* and *Diseases Incidents*, we keep the extrapolation constant to not mis-interpret local events (Armstrong and Collopy, 1993); a sudden rise of incidents would unrealistically be amplified over the course of a few years. As we do not observe such a stark movement for *GDP* and *Population*, which typically inhibit a consistent growth, linear extrapolation is sensible here;[7] this approach stabilizes the respective curves.

2.2.3. Regional benchmarking

If there are no data points for one country available at all, we leverage the average of the respective region (United Nations, Department of Economic and Social Affairs, 2020) as a benchmark. The rationale here is that countries within one region also share socioeconomic characteristics, such as economic strength. Yet, this approach is only necessary for some features and nations: Out of 206 nations participating in 2020, 14 require regional benchmarking for at least one feature. This corresponds to 1.18% of Olympic athletes and 0.72% of Olympic medals across all Olympic Games in the dataset.

[6] As one reviewer has rightfully argued, it is an interesting question whether such home advantage was still apparent in the Tokyo 2020 Games as spectators were effectively banned from attending most competitions. While we believe that the first two arguments for such an advantage – reduced travel fatigue and venue familiarity – are likely to hold under the exceptional circumstances in Japan, the expected negative effect of banned (home) audiences is negligible because the athletes competed in numerous different sports in Tokyo, many of which involve no critical subjective judgments (cf. Singleton et al., 2021). Interestingly, we find some noticeable support for this latter argument in the emerging literature on the causal effects of an absent crowd on performances and refereeing decisions during COVID-19 in professional football, which seem to be moderate at best (e.g., Bryson et al., 2021). Eventually, by earning 58 medals, the Japanese athletes exceeded our forecast (51), which might point to another argument in favour of an intact host effect – a host can suggest new sports to the IOC in advance of the Olympic Games, and Japan earned 14 of their 58 medals in these new sports (baseball/softball: 2; karate: 3; skateboarding: 5; sport climbing: 2; surfing: 2).

[7] From a mathematical point of view, we proceed as follows: For nations, where not more than five consecutive points are missing and there are more data points available than missing, we extrapolate linearly to take the implicit trend into account by using a constrained least-squares approach: With $n<6$ missing values, we use the $n + 1$ nearest available values to estimate the slope of the line. The intercept is given by the nearest available value.

Table 1
Descriptive statistics of numerical variables used in the model including data sources.

Variable	Type	Mini-mum	Maxi-mum	Mean	Std. deviation	Skew-ness	Data Source
Number of medals	Numerical	0	121	4.639	13.190	5.024	Griffin (2018)
Share of global GDP	Numerical	<0.001	0.200	0.005	0.017	7.773	International Monetary Fund, 2019, 2020; The World Bank, 2020
Population ($E + 8$)	Numerical	2.287	14.157	8.398	2.275	−0.512	(Nations, 2019)

Abbreviations and notes. We display all values from 1991 to 2016 as 2020 medals were not known at the time of forecasting.

Table 2
List of ordinal and categorical variables used in the model including data sources.

Variable	Type	Number (ones)	Data Source
Number of athletes	Ordinal		Griffin, 2018; Scelles et al., 2020
0–9 Athletes		589	
10–49 Athletes		388	
50–149 Athletes		230	
Over 149 Athletes		172	
Diseases Deaths (deaths due to lower respiratory diseases)	Ordinal (quintiles)		(Global Burden of Disease Collaborative Network 2018)
Diseases Incidents (people affected by lower respiratory diseases)	Ordinal (quintiles)		(Global Burden of Disease Collaborative Network 2018)
Deaths due to COVID-19	Ordinal (added to Diseases Deaths)		Institute for Health Metrics and Evaluation 2020; World Health Organization, 2020
COVID-19 incidents	Ordinal (added to Diseases Incidents)		Institute for Health Metrics and Evaluation 2020; World Health Organization, 2020
Host country	Categorical		(Wikipedia 2020)
Current Host		7	
Last Time's Host		7	
Next Host		7	
Political regime	Categorical		Scelles et al. (2020)
CAPME (capitalist market economies)		1161	
POSTCOM ((post-) communist economies)		141	
CEEC, joined the EU (Central Eastern European countries)		77	
Region	Categorical		United Nations, Department of Economic and Social Affairs (2020)
Sub-Saharan Africa		314	
Latin America & Caribbean		263	
Western Asia		122	
Southern Europe		95	
South-eastern Asia		72	
Northern Europe		70	
Eastern Europe		67	
Western Europe		63	
Southern Asia		61	
Eastern Asia		49	
Northern Africa		42	
Polynesia		31	
Central Asia		30	
Micronesia		30	
Melanesia		28	
Northern America		21	
Australia and New Zealand		14	
Western Africa		7	

2.3. Conceptual development

The Tobit model marked a milestone in Olympic medal forecasting accounting for the large number of nations winning zero medals (Bernard and Busse, 2004). The concept traces back to Tobin (1958), who argues by the example of household expenditures for luxurious goods that for variables with an upper or lower bound linear regression models do not deliver suitable results. In the context of Olympic medals, a significant share of nations stays without any medals; thus, in this case zero marks the lower bound of the dependent variable. A Tobit model assumes a latent, i.e. non-observable, dependent variable for the number of medals. The observable dependent variable is defined to be equal to the latent one if the latter is greater than zero and zero in all other cases. The resulting graph of the probability distribution resembles a non-linear hockey stick rather than a straight line and describes the observations better than an OLS. To apply this statistical concept in machine learning, we develop a two-staged algorithm: First we determine whether a nation should win any medal at all; then we estimate the exact number of medals.

In both steps, we employ a Random Forest algorithm (cf. Breiman, 2001; Lee, 2021), an *ensemble learner* which has been proven advantageous in various disciplines of sports forecasting (cf. Lessmann et al., 2010; Groll et al., 2019). The ensemble combines the predictive power of multiple decision trees. Each tree resembles an individual flowchart, in which nodes represent decisions based on specific features (cf. Fig. 1). There are two types of decision trees: classifiers and regressors. The former estimate (discrete) labels for each observation. The latter, in contrast, assign a specific number to the variable. In both cases, several independent and identically distributed, randomized trees form a Random Forest.[8] To derive one estimate from n individual decision trees, classifiers apply a majority vote: In an example with $n = 100$ trees and $y_1 = 80$ votes for *"the dependent variable is true"* and $y_2 = 20$ votes for *"the dependent variable is false"*, the Random Forest would vote for *"the dependent variable is true"*. As opposed to this, regressors average the estimates of all trees to generate an aggregate numerical figure. In our model, we train a binary classifier to determine whether a nation should win any medals or not, as a first step. Then, we train a regression model to forecast the exact number of medals for countries with predicted medal success.

Cutler et al. (2012) explains why Random Forests are appealing from both, a computational (among others due to training and prediction time, small number of parameters, and direct use for high dimensional problems) and a statistical (among others due to measures of variable importance, differential class weighting and outlier detection) point of view. The main shortcoming of (individual) decision trees is that they are prone to overfitting (Kirasich et al., 2018). Even though Random Forests partly account for this issue "by using a combination or 'ensemble' of decision trees where the values in the tree are a random, independent, sample" (Kirasich et al., 2018, p. 7), a diligent setup of the forecasting process is essential.

2.4. Forecasting process

The forecasting process comprises three steps: training of different models, model benchmarking, and forecasting (cf. Fig. 2).

[8] The random component of a decision tree stems from training it on a random subset of the data.

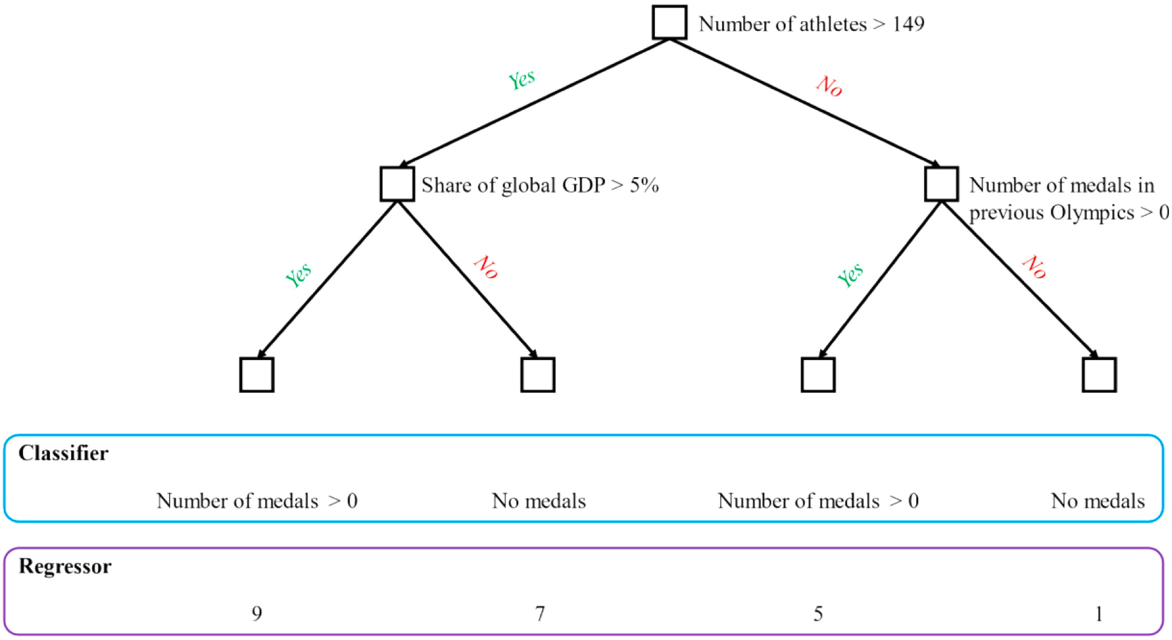

Fig. 1. Exemplary (hypothetical) decision tree.
Abbreviations and notes. The shown decision tree is binary and has depth 2. The difference between a classifier and a regressor is that the classifier computes the label “Number of medals > 0″ or “No medals”, where the regressor computes a numerical value.

In the first step of the forecasting process (*training of different models*), we experiment with promising machine learning models:[9] For the first part of the model, as classifier, we consider a support vector machine (SVM), a decision tree, and Random Forests with 10, 100, respectively 1.000 trees.[10] We find that a Random Forest with 10 decision trees outperforms the other models based on the number of correctly predicted medals. This is also reflected by an area under the receiver operating characteristic (ROC) curve (AUC) of 0.95 (cf. Fig. 3).

For the second part of the model, the regression, we benchmark a Random Forest with 1.000 trees against a range of classical regressions, boosting methods, and neural networks: As classical regressions, we consider a linear regression, a SVM taking into account non-linear transformations of features (Chang and Lin, 2011), and a decision tree regression (Breiman et al., 1984). Boosting methods (Bühlmann and Hothorn, 2007) perform several instances of decision trees (in this case). Each tree compares the output variable to the forecast from the previous step and adapts the setup for the next step based on the error. We include AdaBoost (Freund and Schapire, 1997), which directly takes into account the error, and XGBoost (Chen and Guestrin, 2016), which first transforms the error, as benchmark. Neural networks (Hastie et al., 2001) were motivated by the structure of a biological brain (Hopfield, 1988); they use a computational network, where each node performs a simple transformation and hands over the result to subsequent nodes.[11] Also as regressor, the Random Forest outperforms the described algorithms. As a consequence, we define our final model as a two-staged Random Forest featuring a classifier with 10 trees followed by a regressor with 1.000 trees.

We use *cross-validation* (Kerbaa et al., 2019; Li et al., 2020) to avoid cases of overfitting (Dwork et al., 2015; Roelofs et al., 2019). Cross-validation splits the dataset in training and validation data; while the training data determines the model, the validation data ensures generalization beyond a fixed data sample. There are different forms of cross-validation for cross-sectional data, i.e. samples of a stationary process, and for time series. In case of cross-sectional data, the split is typically random; this approach, however, is inappropriate for time-series as is does not consider the temporal evolution and dependencies in the data. In the context of the Olympic Games, a random selection of datapoints would be prohibitive as, for instance, 2016 data should not be used to predict the results of the 2012 Olympics. Therefore, Bergmeir and Benítez (2012) suggest *last block cross-validation*, a special case of cross-validation using the most recent datapoints as testing data. We use data collected from the years 1991 to 2004 as the *training set*, and data from the 2008 Olympic Games as the *validation set*, to evaluate and compare the performance of distinct models. Only then, we evaluate the final model on the *test set*, which includes data of the 2012 and 2016 Olympic Games and benchmark against other models.

In the second step of the forecasting process (*model benchmarking*), we benchmark the performance of the selected model, the two-staged Random Forest, against a naïve forecast and other forecasts presented in academic literature. The 2012, respectively 2016 Olympic Games serve as *test set*. When computing our estimates, we use the same

[9] We implement all models in Python 3.8.5 (Oliphant, 2007) using the packages pandas 1.1.2 (McKinney, 2010), scikit-learn 0.23.2 (Pedregosa et al., 2011), XGboost 1.2.0 (Chen and Guestrin, 2016), NumPy 1.18.5 (van der Walt et al., 2011), and Shap 0.36.0 (Lundberg and Lee, 2017).

[10] For all models, we use the implementation and standard configuration of Scikit-learn. Two hyperparameters of Random Forests with a particular high impact on the performance are the depth and the number of trees. To determine the optimal depth, we conduct a grid search algorithm, i.e. we compare results for varying (maximum) depths of the trees. We find that a depth of eight performed best. When setting up the number of trees in an ensemble, Oshiro et al. (2012) find that larger trees do not necessarily improve the performance. Based on the number of correctly predicted medals, we use ten trees in the first step of the model. If we were to evaluate the performance based on the area under the receiver operating characteristic (ROC) curve (AUC), we could also opt for a Random Forest with 1.000 trees; however, this does not significantly improve the quality of the forecast (AUC=0.96 for n = 1.000 trees rather than AUC=0.95 for n = 10 trees). In the second step, as regressor, we aim to provide meaningful confidence intervals for the final estimates (based on ensembles of ten trees) and, thus, use 1.000 trees.

[11] We use a dense neural network with three hidden layers of 100 nodes each, with a rectified linear unit (ReLU) as the activation function; furthermore, we use the Adam optimiser (Kingma and Ba, 2014) and run up to 500,000 optimisation steps.

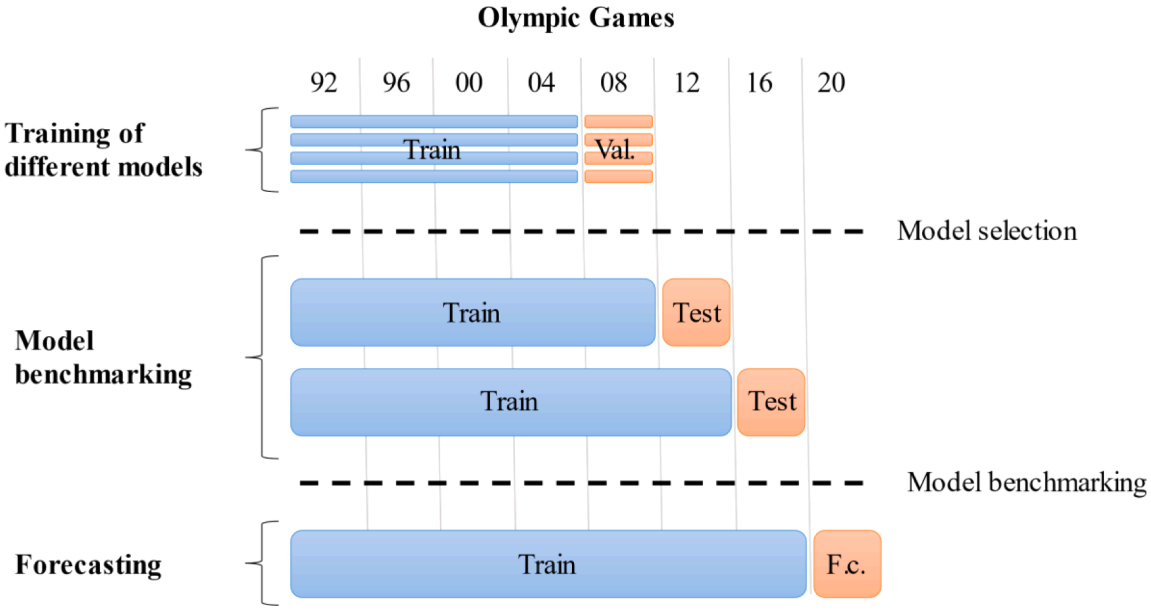

Fig. 2. Illustration of forecasting process.
Abbreviations and notes. Val. = Validation Set, F.c. = Forecast. Following Hastie et al. (2001), we use the training set to fit the parameters of (potential) models. Next, we compare the performance of these models on the validation set; this includes tuning hyperparameters, e.g. determining the number of trees in a Random Forest. Finally, we benchmark the best performing model against a naïve forecast and models published by other researchers based on the predicted labels for the training set.

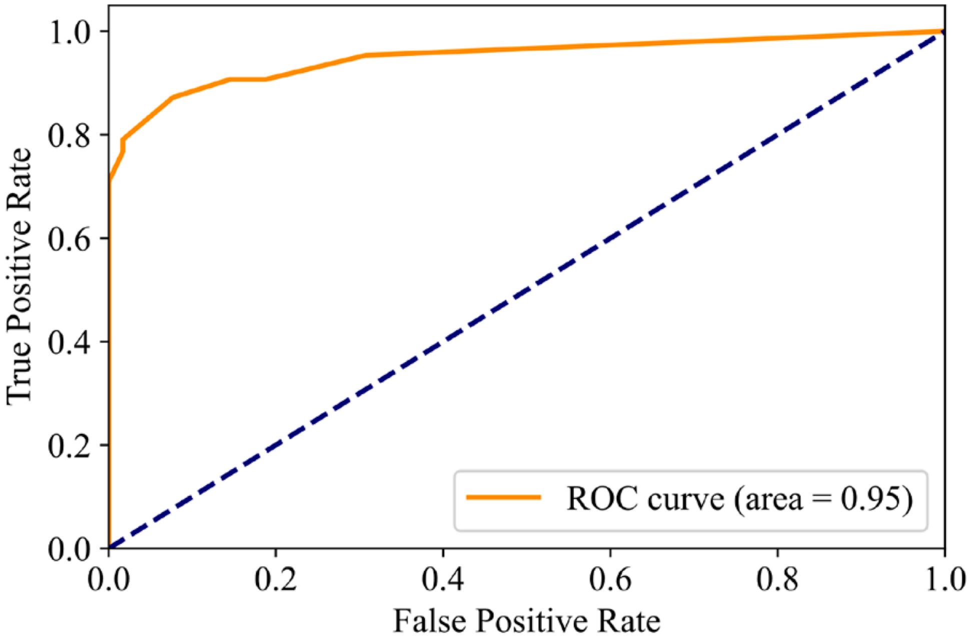

Fig. 3. ROC for the Random Forest classifier with 10 decision trees.
Abbreviations and notes. The curve illustrates the goodness of fit. A value of 1 for the AUC would mean perfect classification.

datapoints as *training set*, that had already been available when the respective papers were developed (cf. Fig. 2). We evaluate the forecast accuracy based on five different metrics M_1-M_5: number of correct forecasts of total (M_1), non-zero (M_2), and zero (M_3) medals, as well as 95% confidence intervals +/- 2 medals (M_4), and absolute deviation for top-17 nations (*M*_5). The number of correct forecasts M_1 is calculated as follows:

$$M_1 = \frac{correct\ medal\ forecasts\ without\ deviation\ to\ actual\ results}{number\ of\ reported\ nations}$$

The metrics M_2 and M_3 follow the same calculation; numerator and denominator are, however, restricted to nations, that won any, respectively no, medals. Furthermore, we check whether the forecast lies in a 95% confidence interval augmented by two medals (M_4). Note that not all papers report this metric. For the two-staged Random Forest, we compute the confidence intervals as follows: We group the 1000 decision trees of the Random Forest in 100 groups of 10 decision trees. These groups serve as smaller Random Forests and we compute their mean values. Then we reduce the 100 obtained data points by eliminating the five data points with the greatest deviation from the mean. The remaining 95 values determine the 95% confidence interval. We only compute confidence intervals for nations with predicted medal success and assume zero otherwise. Then, we augment these confidence intervals by two medals at the lower and upper bound to get M_4. As final metric M_5, we sum the absolute deviation of forecast and actual medals for the top-17 nations:

$$M_5 = \sum_{i=1}^{17} \left| medal\ forecast_i - actual\ number\ of\ medals_i \right|,$$

where i denotes the top-17 nations and |.| the absolute value. Contrary to M_1-M_4, a good forecast is characterized by a small number M_5. We opt for the top-17 nations to ensure comparability with other papers, e.g. Scelles et al. (2020).

In the third step of the forecasting process (forecasting), we make predictions for the 2020 Olympic Games based on training data containing the 2016 Olympics as well. In the following chapter, we show detailed figures for the steps model benchmarking and forecasting and explain the main drivers behind our forecast.

3. Results

While Scelles et al. (2020) improved the prior forecast quality, the presented models still fail to outperform a naïve forecast, i.e. assuming that each country wins exactly the same number of medals as in the previous Olympics. The approach presented in this paper is, to the best of our knowledge, the first to consistently beat the naïve forecast for the 2008, 2012, 2016, and 2020 Games (cf. Table 3). Besides the naïve forecast, we benchmark against seven other models from five different papers.

Applying the algorithm in the context of Tokyo 2020, we forecasted that the United States would defend their top position; however, we also expected that the lead over the principal pursuer China would diminish (cf. Table 4). Indeed, our model correctly predicted China’s strong performance, rooted in a strongly growing economy and a relatively good COVID-19 management. South Korea and Spain performed exactly as expected. On the other hand, we find the greatest deviation between the actual results and our forecast for Australia and the Netherlands. Australia is characterized by a very isolated and remote location; this allowed its government to shield the citizens quite well from the COVID-19 virus by strictly limiting unnecessary international travel. Consequently, the Australian Olympic team could prepare for the Olympics almost as before the pandemic. Due to the novelty of this situation, none of the forecasting models in our benchmark could anticipate this. The Netherlands certainly overperformed the expectations winning 36 medals in Tokyo. A strict focus on specific sports granted the team twelve medals in cycling disciplines and five in rowing. Such a specialization is difficult to factor into a macro-model. Another important remark when evaluating the results addresses the five new types of sports that were introduced during Tokyo 2020 (softball/baseball, karate, skateboarding, sport climbing, and surfing). Based on historic data, it is very difficult to anticipate the performance in these sports; likewise discontinued sports can perturbate models. In 2021, Japan collected 14 out of their 58 medals in newly introduced sports. Finally, boycotts are not predictable; in Tokyo, North Korea did not send any athletes to protect them from COVID-19.

Table 3
Forecasting accuracy of selected models.

	2008	2012	2016	2020
M_1: Correct forecast				
Two-staged Random Forest (this article)	63%	59%	64%	60%
Naïve forecast	59%	56%	60%	54%
Tobit model (Forrest et al., 2010)	47%			
Tobit model (Andreff et al., 2008)	5%			
Logit model (Andreff et al., 2008)	0%			
Hurdle model (Scelles et al., 2020)			22%	4%
Tobit model (Scelles et al., 2020)			43%	0%
Tobit model (Maennig and Wellbrock, 2008)	41%			
OLS (Celik and Gius, 2014)		10%		
M_2: Correct forecast (non-zero medals)				
Two-staged Random Forest (this article)	14%	11%	17%	17%
Naïve forecast	9%	11%	16%	9%
Tobit model (Forrest et al., 2010)	17%			
Tobit model (Andreff et al., 2008)				
Logit model (Andreff et al., 2008)				
Hurdle model (Scelles et al., 2020)			22%	
Tobit model (Scelles et al., 2020)			11%	
Tobit model (Maennig and Wellbrock, 2008)	11%			
OLS (Celik and Gius, 2014)		10%		
M_3: Correct forecast (zero medals)				
Two-staged Random Forest (this article)	98%	93%	97%	95%
Naïve forecast	96%	88%	92%	92%
Tobit model (Forrest et al., 2010)	94%			
Tobit model (Andreff et al., 2008)				
Logit model (Andreff et al., 2008)				
Hurdle model (Scelles et al., 2020)			22%	
Tobit model (Scelles et al., 2020)			69%	
Tobit model (Maennig and Wellbrock, 2008)	83%			
OLS (Celik and Gius, 2014)				
M_4: 95% confidence intervals +/- 2 medals				
Two-staged Random Forest (this article)	92%	96%	93%	89%
Naïve forecast				
Tobit model (Forrest et al., 2010)				
Tobit model (Andreff et al., 2008)	60%			
Logit model (Andreff et al., 2008)	45%			
Hurdle model (Scelles et al., 2020)			93%	58%
Tobit model (Scelles et al., 2020)			91%	58%
Tobit model (Maennig and Wellbrock, 2008)				
OLS (Celik and Gius, 2014)				
M_5: Absolute deviation top-17 nations				
Two-staged Random Forest (this article)	152	91	128	122
Naïve forecast	154	115	114	140
Tobit model (Forrest et al., 2010)	92			
Tobit model (Andreff et al., 2008)	135			
Logit model (Andreff et al., 2008)	204			
Hurdle model (Scelles et al., 2020)			139	175
Tobit model (Scelles et al., 2020)			138	131
Tobit model (Maennig and Wellbrock, 2008)	153			
OLS (Celik and Gius, 2014)		104		

Abbreviations and notes. Note that percentages are based on the number of nations n for which forecasts were published: two-staged Random Forest (This Paper): $n = 203$ in 2008, $n = 205$ in 2012, and $n = 206$ in 2016–2020*;* Naïve Forecast: $n = 203$ in 2008, $n = 205$ in 2012, and $n = 206$ in 2016–2020*;* Tobit Model (Forrest et al., 2010): $n = 127$*;* Tobit Model (Andreff et al., 2008): $n = 20$*;* Logit Model (Andreff et al., 2008): $n = 20$*;* Hurdle Model (Scelles et al., 2020): $n = 192$ in 2008–2016 and $n = 26$ in 2020; Tobit Model (Scelles et al., 2020): $n = 192$ in 2008–2016 and $n = 26$ in 2020; Tobit Model (Maennig and Wellbrock, 2008): $n = 168$*;* OLS (Celik and Gius, 2014): $n = 82$.

To understand the main drivers behind the forecasts better, we use

Table 4
Comparison between actual results and forecasts in the Tokyo 2020 Olympic Games (scenarios with and without COVID-19) and the results of the Rio De Janeiro 2016 Olympic Games.

Rank	Nation	Results 2016	Results 2020	Forecast 2020	Forecast 2020 (No COVID-19)	Forecast Impact of COVID-19	Devia-tion (Forecast 2020 - Results 2020)
1	United States	121	113	120	120	0	+7
2	China	70	88	87	85	+2	−1
3	Russia	56	71	63	62	+1	−8
4	Great Britain	67	65	74	71	+3	+9
5	Japan	41	58	51	50	+1	−7
6	Australia	29	46	29	30	−1	−17
7	Italy	28	40	32	32	0	−8
8	Germany	42	37	45	44	+1	+8
9	Netherlands	19	36	19	19	0	−17
10	France	42	33	44	42	+2	+11
11	Canada	22	24	20	19	+1	−4
12	Brazil	19	21	13	15	−2	−8
13	South Korea	21	20	20	19	+1	0
14	Hungary	15	20	16	16	0	−4
15	New Zealand	18	20	16	16	0	−4
16	Ukraine	11	19	10	11	−1	−9
17	Spain	17	17	17	19	−2	0
18	Cuba	11	15	12	11	+1	−3
19	Poland	11	14	11	11	0	−3
20	Switzerland	7	13	7	8	−1	−6

Abbreviations and notes. For a comprehensive overview of our forecast please refer to Table A2 in the appendix.

the explanatory SHAP value (Lundberg and Lee, 2017). SHAP stands for "Shapley Additive Explanations" and quantifies the importance of features for a forecast. The SHAP value of one feature describes "the change in the expected model prediction when conditioning on that feature" (Lundberg and Lee, 2017, p. 5); starting from the base value, i.e. the prediction without the knowledge of any features, the combination of all SHAP values then leads to the full model forecast (Lundberg and Lee, 2017). The game-theory-based algorithm dates back to Shapley (1953) and runs in polynomial time (Lundberg et al., 2020). The most important features in our model are the number of medals won at the previous Olympic Games, the categorical variable representing the team size (more than 149 athletes), and the normalized GDP (cf. Fig. 4). All of them, generally, have a positive impact on the number of medals won. Note that the naïve forecast draws all its predictive power from the number of medals won at the previous Olympic Games; the high feature importance explains the strong performance of the naïve model.[12] Nevertheless, we find that the remaining features in our model still improve the overall forecasting accuracy.

Three of the features are directly impacted by COVID-19: GDP, incidents of and deaths from lower respiratory diseases.[13] This allows us to create a theoretical scenario without the presence of the pandemic, such that we can clearly quantify its impact (cf. Table 4). Although all three features significantly impact the number of medals, we find that there is little movement caused by COVID-19 amongst the top-20 nations. Notable, however, is the severe impact of the pandemic on the American economy and health system causing a further reduction of the advance of the United States, although absolute medal figures remain largely constant.

Among the top-20 nations, we expected the highest gains in Great Britain (+3 medals), China (+2), and France (+2). Although the pandemic originated in China, the country was hit less severely in global comparison. Liu et al. (2021) mention a high media coverage and an efficient contact tracing as success factors of the Chinese government in fighting COVID-19. We illustrate the main drivers of this development (cf. Fig. 5). Compared to the no COVID-19 scenario, China could slightly increase its share in the world's GDP versus weaker economies around the globe. On a practical level, this means that training measures and competitions as preparation for Tokyo 2020, do not have to be canceled unexpectedly to cover funds in other areas. Both incidents of and deaths from lower respiratory diseases remained on a low level. Altogether, we experience no change in the medal forecast for the Chinese team. However, as the total number of medals forecast by the model declines, scaling moves China up in the medal count. On the opposite, Spain experiences a loss of two medals. Both, a relatively smaller GDP and an increased number of incidents of lower respiratory diseases are responsible for this development (cf. Fig. 6).

4. Discussion and conclusion

Applying a two-staged Random Forest, we significantly improved the forecasting accuracy regarding Olympic medals outperforming a naïve model and currently existing statistical approaches applied in academic literature. Our forecast of the Tokyo 2020 Olympic Games hosted in 2021 suggested that the United States lead the medal count followed by China. Particularly China, that has largely invested in sports development could exhibit a rise in medals. These findings are highly relevant for several stakeholders, not only ex ante, i.e. before the Olympics, but also ex post. Ex ante, media companies and sport sponsors could allocate their resources to promising nations, that are likely to increase their performance compared to the previous Olympics (e.g. China). While spectators demand stories about Olympic heroes rather than group-stage knockouts, the right focus when planning documentaries or interviews is essential for the media to reach a high audience share. The same concept holds for sponsors who profit from signing Olympic teams, that are at the center of high media attention. As discussed at the beginning of this article, we use nation-specific, rather than sports- or athlete-specific

[12] We thank an anonymous reviewer for pointing this out.

[13] We find that our model performance is largely robust with regards to the in-/exclusion of the three COVID-specific variables, i.e. forecasts of both models do not deviate significantly from each other. However, also modelling the impact of COVID-19 improves four out of five metrics measuring the forecasting accuracy: M_1 improves from 57% to 60%, M_2 from 13% to 17%, M_3 from 94% to 95%, and M_4 from 87% to 89%. Only for the absolute medal deviation for top-17 nations M_5 is slightly better in the no-COVID-19 scenario (120 vs. 122).

Table A1
Underlying models of published Olympic forecasts.

Authors	Data sample	Summer / Winter	OLS	Binary	Poisson	Tobit	Two-step	Other model
Ball (1972)	1964	S						X
Grimes et al. (1974)	1936, 1972	S				X		
Baimbridge (1998)	1896–1996	S	X					
Condon et al. (1999)	1996	S	X					X
Kuper and Sterken (2001)	1896–2000	S	X					
Hoffmann et al. (2002)	2000	S	X					
Tcha and Pershin (2003)	1988–1996	S				X		
Johnson and Ali (2004)	1952–2000	S, W	X	X				
Bernard and Busse (2004)	1960–1996	S				X		
Lui and Suen (2008)	1952–2004	S			X	X		
Andreff et al. (2008)	1976–2004	S		X		X		
Maennig and Wellbrock (2008)	1960–2004	S				X		
Forrest et al. (2010)	1996–2004	S				X		
Leeds and Leeds (2012)	1996–2008	S			X			
Vagenas and Vlachokyriakou (2012)	2004	S	X					
Emrich et al. (2012)	1996–2010	S, W	X					
Celik and Gius (2014)	1996–2008	S	X					
Trivedi and Zimmer (2014)	1988–2012	S					X	
Forrest et al. (2015)	1960–2008	S				X		
Lowen et al. (2016)	1996–2012	S				X		
Noland and Stahler (2016b)	1960–2012	S	X			X		X
Noland and Stahler (2016a)	1960–2012	S, W				X		
Noland and Stahler (2017)	1960–2012	S, W				X		
Blais-Morisset et al. (2017)	1992–2012	S			X			
Forrest et al. (2017)	1992–2012	S				X		
Vagenas and Palaiothodorou (2019)	1996–2016	S				X		
Scelles et al. (2020)	1992–2016	S				X	X	
Rewilak (2021)	1996–2016	S				X	X	

Abbreviations and notes. Exemplary selection of studies. Ordinary least squares (OLS); Binary Probit / Logit regression (Binary); Poisson-based model (Poisson); Tobit model (Tobit); Two-step / Hurdle model (Two-step); Summer Games (S); Winter Games (W).

Table A2
Complete forecast medal count of the Olympic Games Tokyo 2020 including 95% confidence intervals (scenarios with and without COVID-19).

Rank	Nation	Medal Fore-cast	Min Confi-dence	Max Confi-dence	Medal Fore-cast (No COVID-19)	Min Confi-dence (No COVID-19)	Max Confi-dence (No COVID-19)	Delta COVID-19 vs. no COVID-19
1	United States	120	111.0	131.8	120	116.5	128.1	0
2	China	87	79.5	94.9	85	78.6	95.9	+2
3	Great Britain	74	68.6	80.8	71	67.0	77.2	+3
4	Russia	63	55.6	70.8	62	56.5	70.6	+1
5	Japan	51	43.6	58.7	50	43.6	58.7	+1
6	Germany	45	42.4	47.6	44	42.7	47.2	+1
7	France	44	38.9	48.6	42	40.1	46.7	+2
8	Italy	32	28.7	37.0	32	29.5	35.9	0
9	Australia	29	25.8	32.5	30	28.3	33.9	−1
10	Canada	20	17.3	22.7	19	17.9	21.2	+1
11	South Korea	20	18.5	21.1	19	18.5	21.1	+1
12	Netherlands	19	15.5	24.3	19	17.7	22.3	0
13	Spain	17	12.7	23.6	19	17.4	20.6	−2
14	Hungary	16	15.1	17.4	16	14.7	17.4	0
15	New Zealand	16	14.2	18.3	16	14.2	18.3	0
16	Kazakhstan	16	13.2	19.6	16	13.7	19.7	0
17	Azerbaijan	14	12.7	16.3	15	13.2	16.9	−1
18	Uzbekistan	14	12.8	16.6	14	12.8	16.3	0
19	Brazil	13	8.4	20.3	15	11.6	20.8	−2
20	Kenya	13	11.5	15.1	13	12.0	15.2	0
21	Denmark	13	10.8	14.9	12	10.9	14.6	+1
22	Cuba	12	9.9	13.8	11	9.9	13.8	+1
23	Poland	11	9.8	12.7	11	10.0	12.9	0
24	Jamaica	11	9.0	12.6	10	9.0	12.6	+1
25	Serbia	10	8.8	11.8	10	8.1	12.0	0
26	Belarus	10	8.6	11.4	10	8.3	11.6	0
27	Ukraine	10	7.3	13.3	11	9.9	13.7	−1
28	Czech Republic	9	7.9	11.0	10	8.2	11.6	−1
29	Ethiopia	9	8.0	10.5	9	8.0	10.0	0
30	Croatia	9	7.8	10.8	9	7.8	10.8	0
31	Sweden	9	6.6	12.1	10	9.1	11.5	−1
32	Georgia	8	6.9	10.3	8	7.5	9.8	0
33	South Africa	8	5.3	13.3	11	8.9	13.7	−3
34	Switzerland	7	5.8	9.6	8	6.5	9.9	−1
35	Turkey	7	5.3	8.9	7	5.5	8.7	0
36	Colombia	7	4.4	10.5	10	8.5	12.2	−3
37	North Korea	6	5.5	7.3	6	5.5	7.3	0

(continued on next page)

Table A2 (*continued*)

Rank	Nation	Medal Forecast	Min Confidence	Max Confidence	Medal Fore-cast (No COVID-19)	Min Confi-dence (No COVID-19)	Max Confi-dence (No COVID-19)	Delta COVID-19 vs. no COVID-19
38	Iran	6	5.1	7.3	6	5.2	7.1	0
39	Thailand	5	4.6	5.8	5	4.6	5.8	0
40	Greece	5	4.3	6.4	5	4.1	6.3	0
41	Belgium	5	3.5	6.6	6	4.7	6.6	−1
42	Chinese Taipei	5	3.9	5.3	4	3.9	5.3	+1
43	Slovakia	4	3.8	5.2	4	3.5	5.0	0
44	Malaysia	4	3.3	5.8	4	3.3	5.8	0
45	Lithuania	4	3.7	5.2	4	3.7	5.1	0
46	Venezuela	4	3.2	5.8	4	3.2	5.8	0
47	Armenia	4	3.5	5.2	5	4.2	5.6	−1
48	Romania	4	3.5	5.3	4	3.8	5.5	0
49	Slovenia	4	3.6	4.9	4	3.7	4.8	0
50	Norway	4	3.7	4.8	4	3.7	4.8	0
51	Bulgaria	4	3.5	4.8	4	3.1	4.5	0
52	Indonesia	4	3.2	5.1	4	3.5	5.1	0
53	Mexico	4	2.9	5.1	5	4.3	6.0	−1
54	Tunisia	4	3.1	4.3	4	3.1	4.3	0
55	India	4	2.1	5.8	4	2.1	6.1	0
56	Argentina	4	2.4	5.2	6	4.6	7.7	−2
57	Algeria	3	2.5	4.1	3	2.1	3.7	0
58	Vietnam	3	2.3	4.1	2	1.7	3.0	+1
59	Egypt	3	1.9	4.2	3	2.0	4.4	0
60	Ireland	3	2.3	3.6	3	2.4	3.3	0
61	Mongolia	3	2.1	3.5	3	2.1	3.5	0
62	Philippines	3	2.0	3.4	3	1.9	3.4	0
63	Nigeria	2	1.9	3.2	2	1.8	3.2	0
64	Latvia	2	2.1	2.7	2	2.1	2.7	0
65	Israel	2	1.8	3.1	2	1.7	2.8	0
66	Finland	2	1.9	2.9	2	1.9	3.0	0
67	Estonia	2	1.9	2.8	2	1.9	2.5	0
68	Morocco	2	2	3	2	1.6	2.1	0
69	Trinidad and Tobago	2	1.8	2.4	2	1.8	2.4	0
70	Bahrain	2	1.7	2.3	2	1.6	2.2	0
71	Portugal	2	1.6	2.2	2	1.7	2.3	0
72	Austria	2	1.6	2.1	2	1.6	2.2	0
73	Ivory Coast	2	1.5	2.1	2	1.4	2.0	0
74	Fiji	2	1.5	2.1	2	1.5	2.1	0
75	Kyrgyzstan	2	1.5	1.8	0			+2
76	Tajikistan	2	1	2	2	1	2	0
77	Singapore	2	1.4	1.8	2	1	2	0
78	Bahamas	2	1.3	1.8	2	1	2	0
79	Moldova	2	1.3	1.8	2	1	2	0
80	Kosovo	1	1	2	2	1	2	−1
81	Grenada	1	1	1	1	1	1	0
82	Guatemala	0			1	1	1	−1

Abbreviations and notes. Confidence intervals are computed by grouping the 1000 decision trees of the Random Forest in 100 groups of 10 decision trees. Within these groups serving as smaller Random Forests mean values are computed. The 100 obtained data points are reduced by eliminating the five data points with the greatest deviation from the mean. The remaining 95 values determine the 95% confidence interval. Confidence intervals are only computed for nations with predicted medal success. All nations from rank 83 on have zero values.

data as this level of granularity generally produces more robust forecasts (Bernard and Busse, 2004). Nevertheless, a more granular model might be useful to identify particularly promising individuals in addition to promising teams. As a third major stakeholder, sports betting companies offer bets on the Olympic medal count. While they generally apply different datasets than the one used in this paper to determine the odds, the strong performance of the two-staged Random Forest suggests that a detailed comparison of both models and potential re-calibration might be beneficial.

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. Such an evaluation helps to assess the impact of specific investments or training concepts. Subsequently, funds for preparing the team for the next Olympics can be allocated. The forecast shows that COVID-19 hardly impacts the number of medals among the top-20 nations. This is mainly due to the fact that decision trees generally (and hence the two-staged Random Forest) exhibit weaknesses regarding extrapolation, in this case caused by the surge in incidents of and deaths from COVID-19 (e.g. Zhao et al., 2020). Training our model with data of the Tokyo 2020 Olympic Games will allow us to quantify the impact of a pandemic like COVID-19 even better. Only then, policy leaders will get a reliable picture on the connection between the management of a pandemic and national sportive success.

Two ways to further improve the performance of the model are the inclusion of additional features and a novel approach for missing data points: First, socioeconomic features, e.g. investments in sports infrastructure, athlete-specific features, e.g. age or disciplines of athletes, and COVID-specific features, e.g. number of canceled national sports events, deliver additional insights and, thus, might improve the forecasting accuracy. Brown et al. (2018) use social media data to forecast football matches, which is an approach that could be applied to Olympic Games as well. However, as machine learning methodologies are prone to overfitting, adding new features is only possible to some extent. Second, while we use inter- and extrapolation to handle missing data points, Hassan et al. (2009) generate the missing values using their probability distribution function. This approach outperforms the conventional mean-substitution approach, however, superiority to inter- and

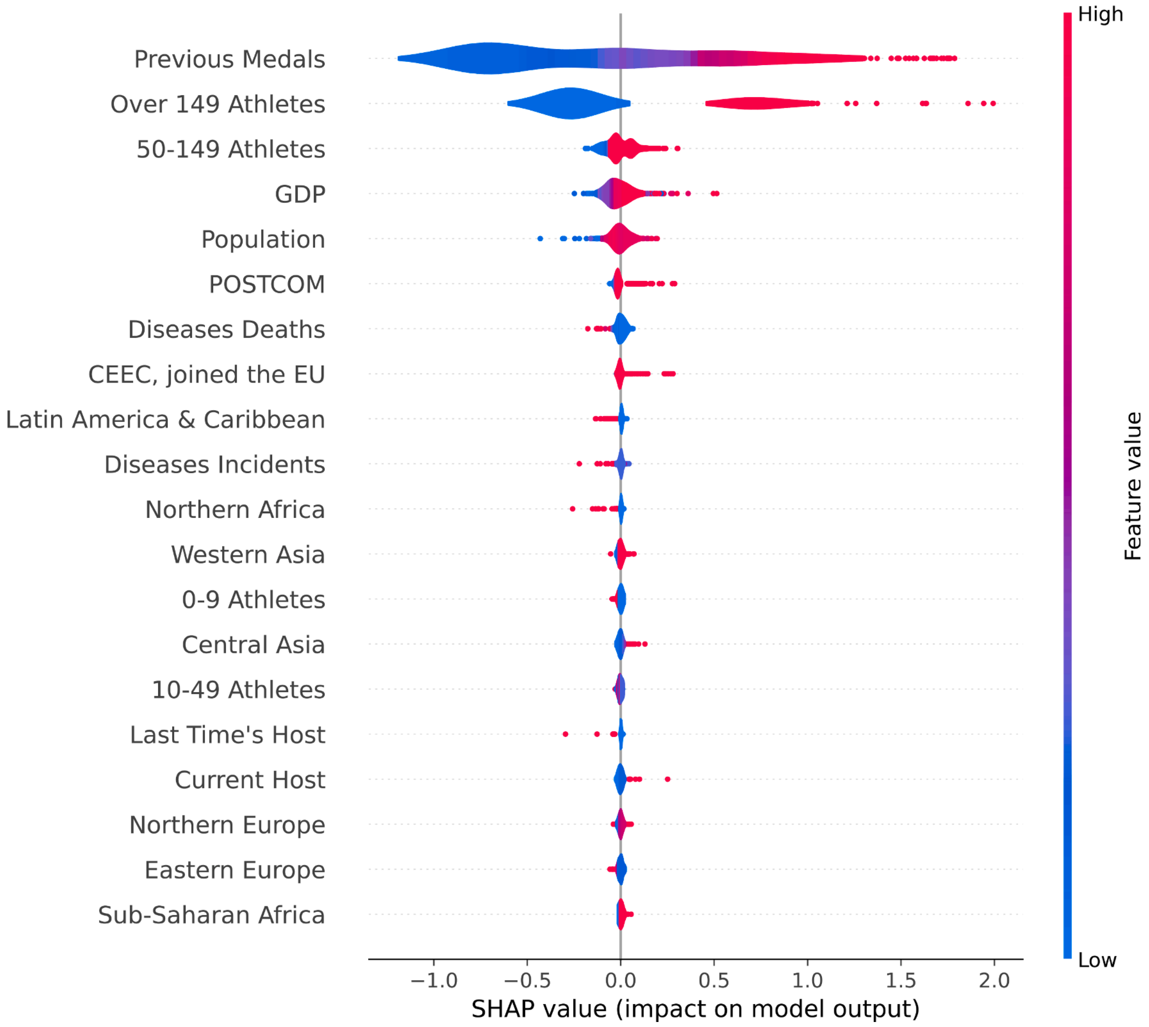

Fig. 4. Feature importance of the two-staged Random Forest.
Abbreviations and notes. Only the 20 most relevant features are depicted. One dot represents one observation in the training data, i.e. one Olympia-nation-combination. Variables are ranked in descending order according to their feature importance. The horizontal location shows whether the effect of the value is associated with a higher or lower prediction. Color shows whether that variable is high (in red) or low (in blue) for each observation. A high "Number of Medals at previous Olympics" has a high and positive impact on the number of medals at the current Olympics. The "high" comes from the red color, and the "positive" impact is shown on the X-axis. Similarly, the "Diseases Deaths" is negatively correlated with the dependent variable.

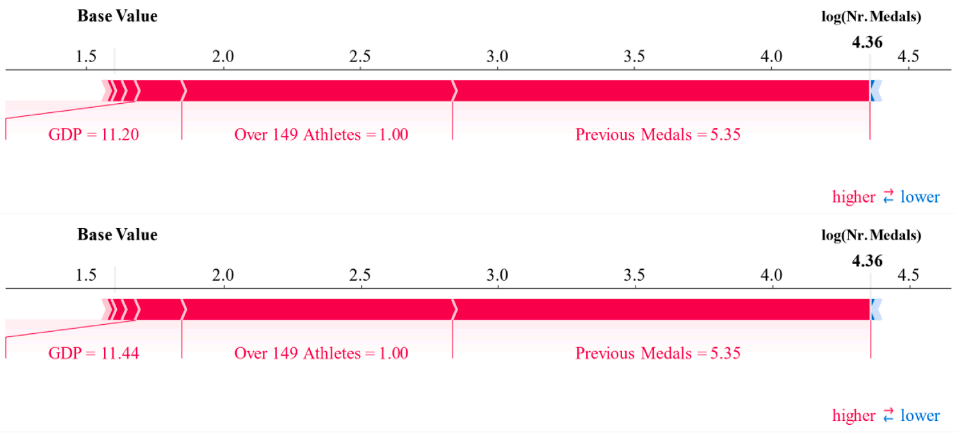

Fig. 5. Individual feature importance for China in the current scenario (top) and in the no COVID-19 scenario (bottom).
Abbreviations and notes. The value for log(Nr. Medals) describes the respective forecast. The base value would be predicted without any knowledge for the current output. Features that push the prediction higher, i.e. to the right are shown in red, while those pushing the prediction lower are illustrated in blue.

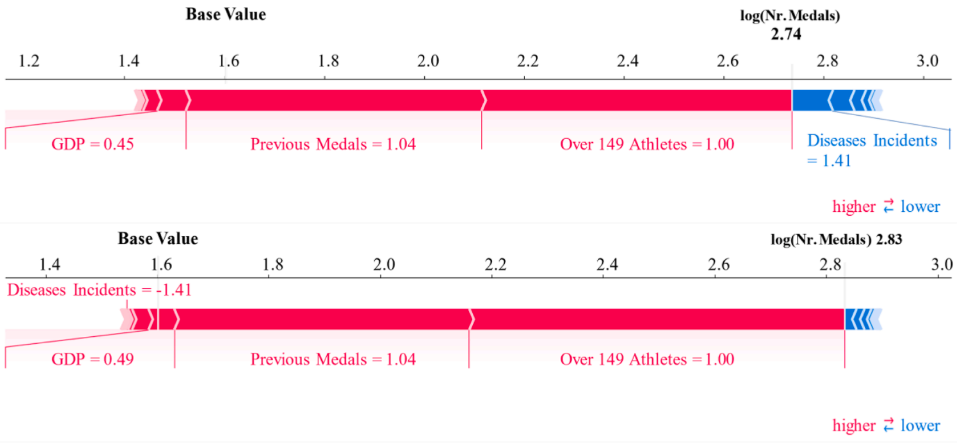

Fig. 6. Individual feature importance for Spain in the current scenario (top) and in the no COVID-19 scenario (bottom).
Abbreviations and notes. The value for log(Nr. Medals) describes the respective forecast. The base value would be predicted without any knowledge for the current output. Features that push the prediction higher, i.e. to the right are shown in red, while those pushing the prediction lower are illustrated in blue.

extrapolation, as applied in this paper, still needs to be proven.

Besides working on model-specific adjustments, scholars can build upon our research within the scope of new applications in sports forecasting. As the Olympic Games are not the only important global sports event, both the comprehensive data set and concept of the two-staged Random Forest presented in this paper, can be leveraged in the context of other competitions, e.g. the Football World Cup, as well.

CRediT authorship contribution statement

Christoph Schlembach: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Sascha L. Schmidt:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Dominik Schreyer:** Conceptualization, Project administration, Writing – original draft, Writing – review & editing. **Linus Wunderlich:** Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

There is no conflict of interest.

Acknowledgments

An earlier working paper was published before the Tokyo 2020 Olympic Summer Games in 2021 (cf., Schlembach et al., 2021). Comments from the Editor and six anonymous reviewers are gratefully acknowledged. Journal policies have been reviewed. The usual disclaimer applies.

Funding

No funding was received for this manuscript.

Data Availability Statement

The data that supports the findings of this study are available upon request.

References

Allison, L., Monnington, T., 2002. Sport, prestige and international relations. Gov. Oppos. 37, 106–134.

Andreff, M., Andreff, W., Poupaux, S., 2008. Les déterminants économiques de la performance olympiques: prévision des médailles qui seront gagnées aux Jeux de Pékin. Revue d'économie politique 118, 135–169.

Armstrong, J.S., Collopy, F., 1993. Causal forces: structuring knowledge for time-series extrapolation. J. Forecast. 12, 103–115.

Baboota, R., Kaur, H., 2019. Predictive analysis and modelling football results using machine learning approach for english premier league. Int. J. Forecast. 35, 741–755.

Baimbridge, M., 1998. Outcome uncertainty in sporting competition: the Olympic games 1896–1996. Appl. Econ. Lett. 5, 161–164.

Ball, D.W., 1972. Olympic games competition: structural correlates of national success. Int. J. Comp. Sociol. 13, 186–200.

Behrang, M.A., Assareh, E., Assari Ghanbarzadeh, A., 2011. Using bees algorithm and artificial neural network to forecast world carbon dioxide emission. Energy Sources Part A 33, 1747–1759.

Beigl, P., Wassermann, G., Schneider, F., Salhofer, S., 2004. Forecasting municipal solid waste generation in major European cities.

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inf. Sci. (Ny) 191, 192–213.

Bernard, A.B., Busse, M.R., 2004. Who wins the Olympic games: economic resources and medal totals. Rev. Econ. Stat. 86, 413–417.

Blais-Morisset, P., Boucher, V., Fortin, B., 2017. The impact of public investment in sports on the Olympic medals. Revue economique 68, 623–642.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Brown, A., Rambaccussing, D., Reade, J.J., Rossi, G., 2018. Forecasting with social media: evidence from tweets on soccer matches. Econ. Inq. 56, 1748–1763.

Bryson, A., Dolton, P., Reade, J.J., Schreyer, D., Singleton, C., 2021. Causal effects of an absent crowd on performances and refereeing decisions during Covid-19. Econ. Lett. 198, 109664.

Bühlmann, P., Hothorn, T., 2007. Boosting algorithms: regularization, prediction and model fitting. Stat. Sci. 22, 477–505.

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S.C., Di Napoli, R., 2020. Features, Evaluation and Treatment Coronavirus (COVID-19). Statpearls [internet].

Celik, O.B., Gius, M., 2014. Estimating the determinants of summer Olympic game performance. J. Appl. Econ. 11, 39–47.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 1–27.

Chen, S., Wang, J., Zhang, H., 2019. A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting. Technol. Forecast. Soc. Change 146, 41–54.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Choi, J.-.H., Kim, J., Won, J., Min, O., 2019. Modelling chlorophyll-a concentration using deep neural networks considering extreme data imbalance and skewness. 21st International Conference on Advanced Communication Technology (ICACT), 631–634.

Christodoulos, C., Michalakelis, C., Varoutas, D., 2010. Forecasting with limited data: combining ARIMA and diffusion models. Technol. Forecast. Soc. Change 77, 558–565.

Condon, E.M., Golden, B.L., Wasil, E.A., 1999. Predicting the success of nations at the summer Olympics using neural networks. Comput. Oper. Res. 26, 1243–1265.

Courneya, K.S., Carron, A.V., 1991. Effects of travel and length of home stand/road trip on tie home advantage. J. Sport Exerc. Psychol. 13, 42–49.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random forests. Ensemble Machine Learning. Springer, pp. 157–175.

De Bosscher, V., Heyndels, B., De Knop, P., van Bottenburg, M., Shibli, S., 2008. The paradox of measuring success of nations in elite sport. Belgeo 217–234.

De Bosscher, V., de Knop, P., van Bottenburg, M., Shibli, S., 2006. A conceptual framework for analysing sports policy factors leading to international sporting success. Eur. Sport Manag. Q. 6, 185–215.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A.L., 2015. Preserving Statistical Validity in Adaptive Data Analysis, STOC '15: Proceedings of the forty-seventh annual ACM symposium on Theory of Computing.

Emrich, E., Klein, M., Pitsch, W., Pierdzioch, C., 2012. On the determinants of sporting success–a note on the Olympic games. Economics Bulletin 32, 1890–1901.

Ferraresi, M., Gucciardi, G., 2021. Who chokes on a penalty kick? Social environment and individual performance during Covid-19 times. Econ. Lett. 203, 109868.

Forrest, D., McHale, I.G., Sanz, I., Tena, J.D., 2015. Determinants of national medals totals at the summer Olympic games: an analysis disaggregated by sport. The Economics of Competitive Sports. Edward Elgar Publishing.

Forrest, D., McHale, I.G., Sanz, I., Tena, J.D., 2017. An analysis of country medal shares in individual sports at the Olympics. Eur. Sport Manag. Q 17, 117–131.

Forrest, D., Sanz, I., Tena, J.D., 2010. Forecasting national team medal totals at the summer Olympic games. Int. J. Forecast. 26, 576–588.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Garcia-del-Barrio, P., Gomez-Gonzalez, C., Sánchez-Santos, J.M., 2020. Popularity and visibility appraisals for computing Olympic medal rankings. Soc. Sci. Q. 101, 2137–2157.

Girginov, V., Hills, L., 2013. A sustainable sports legacy: creating a link between the London Olympics and sports participation. Olympic legacies: Intended and Unintended. Routledge, pp. 240–265.

Global Burden of Disease Collaborative Network, 2018. Global Burden of Disease Study 2017 (GBD 2017) Results. http://ghdx.healthdata.org/gbd-results-tool. Accessed 4 August 2020.

Griffin, R.H., 2018. *120 years of Olympic history: athletes and results*. https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results. Accessed 4 August 2020.

Grimes, A.R., Kelly, W.J., Rubin, P.H., 1974. A socioeconomic model of national Olympic performance. Soc. Sci. Q. 777–783.

Groll, A., Ley, C., Schauberger, G., van Eetvelde, H., 2019. A hybrid random forest to predict soccer matches in international tournaments. J. Quant. Anal. Sports 15, 271–287.

Hammerschmidt, J., Durst, S., Kraus, S., Puumalainen, K., 2021. Professional football clubs and empirical evidence from the COVID-19 crisis: time for sport entrepreneurship? Technol. Forecast. Soc. Change 165, 120572.

Hassan, M., Atiya, A., Gayar, N., El Fouly, R., 2009. Novel ensemble techniques for regression with missing data. New Math. Nat. Comput. 05, 635–652.

Hastie, T., Friedman, J., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer, New York, New York, NY.

Hermann, A., 2019. The tip of the iceberg: the Russian doping scandal reveals a widespread doping problem. Diagoras: International Academic Journal on Olympic Studies 3, 45–71.

Hoffmann, R., Ging, L.C., Ramasamy, B., 2002. Public policy and olympic success. Appl. Econ. Lett. 9, 545–548.

Hopfield, J.J., 1988. Artificial neural networks. IEEE Circuits Syst. Mag. 4, 3–10.

Horky, T., 2021. No sports, no spectators – no media, no money? the importance of spectators and broadcasting for professional sports during COVID-19. Soccer Soc. 22, 96–102.

Humphreys, B.R., Johnson, B.K., Mason, D.S., Whitehead, J.C., 2018. Estimating the value of medal success in the Olympic Games. J. Sports Econom. 19, 398–416.

Institute for Health Metrics and Evaluation, 2020. *COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections*. http://www.healthdata.org/covid. Accessed 4 August 2020.

International Monetary Fund, 2019. *World Economic Outlook Database*October 2019. https://www.imf.org/external/pubs/ft/weo/2019/02/weodata/index.aspx. Accessed 4 August 2020.

International Monetary Fund, 2020. World Economic Outlook Database April 2020. https://www.imf.org/external/pubs/ft/weo/2020/01/weodata/index.aspx. Accessed 4 August 2020.

International Olympic Committee, 2020. ***Press statement on*** March 30th, 2020. https://www.olympic.org/news/ioc-ipc-tokyo-2020-organising-committee-and-tokyo-metropolitan-government-announce-new-dates-for-the-olympic-and-paralympic-games-tokyo-2020.

Johnson, D.K.N., Ali, A., 2004. A tale of two seasons: participation and medal counts at the summer and winter Olympic games. Soc. Sci. Q. 85, 974–993.

Johnston, D.F., 1970. Forecasting methods in the social sciences. Technol. Forecast. Soc. Change 2, 173–187.

Kankal, M., Akpınar, A., Kömürcü, M.İ., Özşahin, T.Ş., 2011. Modeling and forecasting of Turkey's energy consumption using socio-economic and demographic variables. Appl. Energy 88, 1927–1939.

Kerbaa, T.H., Mezache, A., Oudira, H., 2019. Model selection of sea clutter using cross validation method. Procedia. Comput. Sci. 158, 394–400.

Kingma, D.P., Ba, J., 2014. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980 .

Kirasich, K., Smith, T., Sadler, B., 2018. Random Forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review 1, 1–9.

Klobučník, M., Plešivčák, M., Vrábeľ, M., 2019. Football clubs' sports performance in the context of their market value and GDP in the European Union regions. Bull. Geogr. Phys. Geogr. Ser. 45, 59–74.

Krishna, A., Haglund, E., 2008. Why do some countries win more Olympic medals? lessons for social mobility and poverty reduction. Econ. Polit. Wkly. 143–151.

Kuper, G.H., Sterken, E., 2001. *Olympic participation and performance since 1896*. SSRN Electronic Journal 274295.

Lee, C., 2021. A review of data analytics in technological forecasting. Technol. Forecast. Soc. Change 166, 120646.

Leeds, E.M., 2019. Olympic performance. The SAGE Handbook of Sports Economics 377.

Leeds, E.M., Leeds, M.A., 2012. Gold, silver, and bronze: determining national success in men's and women's Summer Olympic events. Jahrbücher für Nationalökonomie und Statistik 232, 279–292.

Lessmann, S., Sung, M.-C., Johnson, J.E., 2010. Alternative methods of predicting competitive events: an application in horserace betting markets. Int. J. Forecast. 26, 518–536.

Li, T., Levina, E., Zhu, J., 2020. Network cross-validation by edge sampling. Biometrika 107, 257–276.

Liu, N., Chen, Z., Bao, G., 2021. Role of media coverage in mitigating COVID-19 transmission: evidence from China. Technol. Forecast. Soc. Change 163, 120435.

Lowen, A., Deaner, R.O., Schmitt, E., 2016. Guys and gals going for gold: the role of women's empowerment in Olympic success. J. Sports Econom. 17, 260–285.

Lui, H.-K., Suen, W., 2008. Men, money, and medals: an econometric analysis of the Olympic games. Pac. Econ. Rev. 13, 1–16.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, 30th ed. Curran Associates, Inc, pp. 4765–4774.

Maennig, W., Wellbrock, C., 2008. Sozioökonomische Schätzungen olympischer Medaillengewinne. Sportwissenschaft 38, 131–148.

Makridakis, S., Hyndman, R.J., Petropoulos, F., 2020. Forecasting in social settings: the state of the art. Int. J. Forecast. 36, 15–28.

Manuel Luiz, J., Fadal, R., 2011. An economic analysis of sports performance in Africa. Int. J. Soc. Econ. 38, 869–883.

McKinney, W., 2010. Data Structures for Statistical Computing in Python, Python in Science Conference.

Metsä-Tokila, T., 2002. Combining competitive sports and education: how top-level sport became part of the school system in the soviet union, sweden and finland. Eur. Phy. Educ. Rev. 8, 196–206.

Modis, T., 2013. Long-term GDP forecasts and the prospects for growth. Technol. Forecast. Soc. Change 80, 1557–1562.

Mohr, M., Nassis, G.P., Brito, J., Randers, M.B., Castagna, C., Parnell, D., Krustrup, P., 2020. Return to elite football after the COVID-19 lockdown. Manag. Sport Leis. 1–9.

Mon-López, D., García-Aliaga, A., Ginés Bartolomé, A., Muriarte Solana, D., 2020. How has COVID-19 modified training and mood in professional and non-professional football players? Physiol. Behav. 227, 113148.

Noland, M., Stahler, K., 2016a. Asian participation and performance at the Olympic games. Asian Econ. Policy Rev. 11, 70–90.

Noland, M., Stahler, K., 2016b. What goes into a medal: women's inclusion and success at the Olympic Games. Soc. Sci. Q. 97, 177–196.

Noland, M., Stahler, K., 2017. An old boys club no more: pluralism in participation and performance at the Olympic Games. J. Sports Econom. 18, 506–536.

Oliphant, T.E., 2007. Python for scientific computing. Comput. Sci. Eng. 9, 10–20.

Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest? In: Perner, P. (Ed.), Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science, pp. 154–168 vol. 7376.

Parnell, D., Bond, A.J., Widdop, P., Cockayne, D., 2021. Football worlds: business and networks during COVID-19. Soccer Soc. 22, 19–26.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 85, 2825–2830.

Pollard, R., 2002. Evidence of a reduced home advantage when a team moves to a new stadium. J. Sports Sci. 20, 969–973.

Puertas, R., Marti, L., Guaita-Martinez, J.M., 2020. Innovation, lifestyle, policy and socioeconomic factors: an analysis of European quality of life. Technol. Forecast. Soc. Change 160, 120209.

Rewilak, J., 2021. The (non) determinants of Olympic success. J. Sports Econom., 152700252199283

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., Schmidt, L., 2019. A meta-analysis of overfitting in machine learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, 32nd ed. Curran Associates, Inc, pp. 9179–9189.

Scelles, N., Andreff, W., Bonnal, L., Andreff, M., Favard, P., 2020. Forecasting national medal totals at the summer Olympic games reconsidered. Soc. Sci. Q. 101, 697–711.

Schlembach, C., Schmidt, S.L., Schreyer, D., Wunderlich, L., 2021. Forecasting the Olympic Medal Distribution during a Pandemic:A Socio-Economic Machine Learning Model. SSRN Electronic Journal 3745595. https://doi.org/10.2139/ssrn.3745595.

Shapley, L.S., 1953. A value for n-person games. Contributions to the Theory of Games 2, 307–317.

Singleton, C., Reade, J., Rewilak, J., Schreyer, D., 2021. How big is home advantage at the Olympic games? SSRN Electronic Journal (3888639). https://doi.org/10.2139/ssrn.3888639 https://doi.org/.

Stekler, H.O., Sendor, D., Verlander, R., 2010. Issues in sports forecasting. Int. J. Forecast. 26, 606–621.

Streicher, T., Schmidt, S.L., Schreyer, D., Torgler, B., 2020. Anticipated feelings and support for public mega projects: hosting the Olympic Games. Technol. Forecast. Soc. Change 158, 120158.

Tcha, M., Pershin, V., 2003. Reconsidering performance at the summer Olympics and revealed comparative advantage. J. Sports Econom. 4, 216–239.

The World Bank, 2020. Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators. ID: NY.GDP.PCAP.CD. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD. Accessed 4 August 2020.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24.

Trivedi, P.K., Zimmer, D.M., 2014. Success at the summer Olympics: how much do economic factors explain? Econometrics 2, 169–202.

United Nations, Department of economic and social affairs, 2019. World Population Prospects 2019. Online Edition. Rev. 1. https://population.un.org/wpp/. Accessed 4 August 2020.

United Nations, Department of Economic and Social Affairs, 2020. *Standard country or area codes for statistical use (M49)*. https://unstats.un.org/unsd/methodology/m49/overview/. Accessed 4 August 2020.

Vagenas, G., Palaiothodorou, D., 2019. Climatic origin is unrelated to national Olympic success and specialization: an analysis of six successive games (1996–2016) using 12 dissimilar sports categories. Sport Soc. 22, 1961–1974.

Vagenas, G., Vlachokyriakou, E., 2012. Olympic medals and demo-economic factors: novel predictors, the ex-host effect, the exact role of team size, and the “population-GDP” model revisited. Sport Manage. Rev. 15, 211–217.

van der Walt, S., Colbert, S.C., Varoquaux, G., 2011. The NumPy Array: a structure for efficient numerical computation. Comput. Sci. Eng. 13 (2), 22–30. Comput. Sci. Eng. 13, 22–30.

Wang, Y., Kung, L., Byrd, T.A., 2018. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast. Soc. Change 126, 3–13.

Weed, M., Coren, E., Fiore, J., Wellard, I., Chatziefstathiou, D., Mansfield, L., Dowse, S., 2015. The Olympic Games and raising sport participation: a systematic review of evidence and an interrogation of policy for a demonstration effect. Eur. Sport Manag. Q. 15, 195–226.

Wikipedia, 2020. *List of Olympic Games host cities*. https://en.wikipedia.org/wiki/List_of_Olympic_Games_host_cities. Accessed 4 August 2020.

Wong, A.Y.-Y., Ling, S.K.-K., Louie, L.H.-T., Law, G.Y.-K., So, R.C.-H., Lee, D.C.-W., Yau, F.C.-F., Yung, P.S.-H., 2020. Impact of the COVID-19 pandemic on sports and exercise. Asia-Pac. J. Sports Med. Arthrosc. Rehabil. Technol. 22, 39–44.

World Health Organization, 2020. *WHO Coronavirus Disease (COVID-19) Dashboard*. https://covid19.who.int/. Accessed 4 August 2020.

Zhao, X., Yan, X., Yu, A., van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav. Soc. 20, 22–35.

Christoph Schlembach is a doctoral researcher at WHU – Otto Beisheim School of Management, Center for Sports and Management, in Germany. He studied finance and information management at the Technical University of Munich and Augsburg University, as well as mathematics at the Technical University of Munich.

Sascha L. Schmidt is a professor, chair holder and director of the Center for Sports and Management (CSM) at WHU – Otto Beisheim School of Management in Dusseldorf, Germany. At the same time he is the academic director of “SPOAC - Sports Business Academy by WHU” and affiliate professor at the Laboratory for Innovation Science (LISH) at Harvard University in Boston, USA. The “Future of Sports” is one of his key research areas. He is author of the book *21st Century Sports: How Technologies Will Change Sports in the Digital Age* and published his work in several books and peer-reviewed journals, including Technological Forecasting & Social Change, Journal of Business Research, and Applied Psychology.

Dominik Schreyer is an assistant professor of Sports Economics at WHU – Otto Beisheim School of Management in Düsseldorf, Germany. In his research, he explores the role of sociopsychological factors in individual economic behavior and decision-making through the lenses of professional sports. Further, he takes a keen interest in analyzing sports demand (e.g. football spectator no-show behavior). He has published 25+ articles in international peer-reviewed journals, including in Games and Economic Behavior, the Journal of Economic Psychology, and Technological Forecasting & Social Change.

Linus Wunderlich is a postdoctoral researcher at the School of Mathematical Sciences of the Queen Mary University of London. In his work on computational finance, he combines his background in modern finite element methods with his interest in machine learning.