Forecasting Olympic medal distributions with Machine Learning - Discussion

Aidan T, Michael W, Omair K

Dec 7, 2022

ABSTRACT

While forecasting the future is applied to many fields, forecasting olympic medals has grown in popularity for multiple reasons. The Olympic games have been a global tradition for over a century. Modern Olympics provide an opportunity for elite athletes from all around the world to gather and take part under one competition to earn pride for their nations. Each year there will be winners and runner-ups, and many stakeholders need access to accurate prediction models to appropriately assess the multitude of factors that go into making an olympic medalist and make adjustments to their own benefits. Examples include sports betting companies, National Olympic Committees, or even governments. An accurate medal prediction could earn sports betting companies the best odds for their offerings. For National Olympic Committees, medal predictions could call for adjustments in training focuses to tailor for better overall team performance. Finally, governments can use medal predictions to optimize budget allocations on areas related to sports and training programs, which are especially relevant for nations that take greater pride in thriving in competitive sports. Having a good medal prediction is no simple task, however. Many researchers over the many past Olympic games have refined said practice using multiple techniques and data-analyzing strategies, with different degrees of success. First attempts of Olympic medal predictions can be dated to decades back, where researchers used various features of participating countries to draw conclusions based on their preferred models. These features range from GDP, previous medal scores, to even the genders of athletes. The data are then applied to forecasting models such as OLS regressions, Poisson, or negative binomial models. The focus of this discussion will be based on a two-step random forest model that uses machine learning and revises upon previous forecasting strategies. Both feature selection and prediction are improved by learning and improving from older models, and the results consistently outperform.

CHAPTER 1: Introduction to Forecasting Olympic Medals

In essence, the paper combines the previously developed tobit regression concepts with the random forest technique that showed quite successful results in predicting

football scores. The authors tackle the problem using a two-stage random forest model. First, by using a random forest classification to estimate if a country will win any medals at all then by estimating the exact number of medals they will win using a random forest regression. They do this by using similar variables to previous studies such as if the country is, has or will host the Olympics, GDP, previous number of medals, country population, region, and political system. The result of this model is the ability to predict non-medal and medal winning countries with greater accuracy than naive forecasting (where we just assume the same number of medals as last olympics). Similarly when compared to previous tobit and logit regressions, the two-staged random forest did significantly better at forecasting all medal outcomes over the back 4 olympics (2008, 2012, 2016, 2020) than any other previous model.

The future scope to improve prediction of the two-staged random forest may include additional features. Although GDP is accounted for, there are socioeconomic factors such as investment in sport infrastructure and athletic factors such as age that could greatly affect the model's performance. As well, this model can be applied to other leagues and sports to forecast their outcomes.

CHAPTER 2: Literature Review

The current model attempts to overcome issues in previous models while also taking learnings from previous models, and applying them to the current techniques. The factors being used in the model remained relatively similar with the exception of a new variable which takes into account the effects of COVID-19 (By factoring in the deaths and incidents by country). While the two-step method is not new and was applied in previous tobit regressions, the authors use only 1 year of recent socioeconomic data rather than the entire 4 years leading up to the olympics. This solved an incorrect assumption of previous authors that an athlete had 4 years to prepare for an Olympics when in fact, sports forecasting was proven to be more accurate with more recent data. As well, OLS, binary, poisson and tobit have been previously used but never a random forest.

The two-step random forest contained a classification and regression. This overcame the issue of factors not being linearly related and being averaged out across a

single regression coefficient. Utilizing a random forest assigned different probabilities to the same variable based on other factors and improved model accuracy.

The biggest issue facing the current model is overfitting. While the tobit regression has similar issues, the authors must still remain cautious not to overfit the model with too many decision trees. In an attempt to reduce overfitting, a last block cross-validation was completed which uses 1991 - 2004 as the training set, 2008 as the validation set, and 2012 & 2016 as the test set. This is to avoid the fact that data should not be used to forecast previous results, only future events and therefore the model is validated and tested using the more recent data and trained using older data.

CHAPTER 3:

The first step in the two stage random forest model is choosing the dependent and independent variables and the motivation behind choosing them. Because this is a two-stage random forest model the first stage involves identifying whether a nation will win a medal or not. This is a classification problem. The second-stage is if a country will win a medal what is the number of medals it will win. This is a regression problem. The dependent variable for this model is therefore the number of medals won. It is important to note there is no distinction between whether the medal won was gold, silver or bronze as it only looks at whether there was a medal won because previous research has shown that models that do not distinguish between medal types produce more accurate forecasts. The dependent variable is also logged and rescaled and then finally rounded to increase accuracy and interpretability. There are 11 independent variables in this model which can be seen in the table in Exhibit 1 & 2. The GDP relative to global GDP is chosen as a proxy for economic resources of a country as it is widely used in other Olympic studies. The reason is that countries with higher GDPs are likely to invest in better sporting facilities/ have more opportunities for sporting in general. Population is also used as a feature because a larger population is, in general, likely to produce more athletes. The number of participating athletes is a categorical variable which classifies the number of athletes sent by a country into different buckets. The model also includes COVID-19 deaths and incidents as well other deaths and incidents related to lower respiratory diseases as this could have impacted the Olympics performance as well as represent

overall health and sanitary conditions in a country. Home advantage is also important in Olympics or any sport and a categorical variable for the current, past and future host is also used in the model. There is also a categorical variable to take into account the political system because Soviet countries outperformed other countries due to emphasis on sport. Therefore, there is a variable which determines whether a country was previously a part of the Soviet Union, was a part of the Soviet Union and is now an EU member, or is a capitalist market economy. To take into account culture, tradition and climate there is also a categorical variable for the region of the country as defined by the United Nations. The number of medals won in previous Olympics events is also included in our model as this is an important determinant recommended in previous studies.

The next step is to carry out the necessary data preprocessing. Our data contain the Olympics data from 1991 to 2020 for 206 countries creating a total of 1379 country year observations. Firstly, the researchers map the data to their respective nations, for example the population of Anguilla, which is a part of Great Britain, is added to the population of Great Britain. For missing data interpolation and extrapolation is used. In interpolation the missing values are forecasted linearly however in extrapolation the missing values are forecast a method that would be appropriate to the feature itself. For countries with no data point regional benchmarking is used which takes the average of the region.

CHAPTER 4:

The Tobit regression model is the underlying statistical model used in this particular Random Forest Model. The Tobit model sets an upper and lower bound for the dependent variable instead of having no upper or lower bounds as is the case with standard linear models. In this particular case the lower bound is 0 as this is the minimum number of medals a country can win. There is no information on whether an upper bound is used in this model. The Tobit model assumes an underlying latent (non-observable) dependent variable for the model and maps any values less than for the dependent variable as 0 and values greater than 0 as is. The researchers then use a Random Forest model which determines each step of the two-step Random Forest model. A random Forest model is a type of an ensemble model which means that it combines several tree based models to

generate the output. For classification models we use the Random Forest Classifier model, it takes the majority vote of the individual models. For example if the majority of the models predict that a country will not win a medal then the output is that the country will not win a medal. For the second step or Random Forest Regressor model the model takes the average of the individual models. The reason the researchers use the Random Forest Model is due to computational time, small number of parameters, and a statistical point of view.

Next the researchers experimented using different models. For the first step of the two-step model the researcher experimented using a SVM model, a simple tree based model and a Random Forest Model with 10, 100, 1000 trees. The Random Forest Model with 10 trees performs best and results in a ROC curve with an AUC of 0.95 which indicates that there was a low trade off between sensitivity(True Positive rate) and specificity (True Negative Rate). For the second step the researchers used classical regression models, Boosting methods which take into account several decision trees such as AdaBoost and XGBoost, and neural networks, and Random Forest model with 1000 trees. The Random Forest with 1000 trees outperformed all the other models so the researcher continued with the Random Forest with 10 trees and 1000 trees for the first and second step respectively. Cross Validation is used to ensure that the model avoids overfitting to the training dataset. This is a method that randomly splits the data into training and validation sets. Since the data is a time-series data the researchers use a unique cross validation method called last block cross-validation which only uses the most recent data points as the validation set.

The researcher also benchmarked the performance of their model against a naive model which predicted the number of medals won by a country as the amount won the previous year. When calculating the accuracy of their models the researchers used the 5 measures in Exhibit 1.

The results obtained from the models are that this is the first paper to consistently beat the naive model forecasts in predicting the number of medals for each country. The model also managed to predict that the United States would continue to dominate in the Olympics however it would lose some of its lead over China. The model also predicted the performance of Spain and South Korea exactly as they did in the 2020 Olympics. The

researchers also used the SHAP value of a feature to determine feature importance and identified that the number of medals won at the previous event was the most important feature followed by team size and GDP. The reason the naive model had outperformed previous models was because it drew all its predictive power from the number of previous medals won. Despite this the researchers thought the additional variables still benefited the overall accuracy of their model.

Conclusion and recommendations:

In summary, the two-step random forest model outperforms older models through better feature selection and predicting power. By applying a two-step process, noise from medal numbers is reduced and notable relevant features such as GDP become more impactful in yielding accurate results. Further recommendations upon the existing model may include expanding its capabilities, both vertically and horizontally. The current model is powerful enough to forecast results for individual athletes as well as other global sporting events such as the world cup, with only requiring minor adjustments to feature selection and data cleaning. The two-step process can also evolve further into more steps to gain detailed predictions for more specific information, such as performance during specific dates during the Olympic games. The model on hand has made several meaningful improvements and takes great advantage of random forest in machine learning to gain high forecasting performance.

References:

Tobit regression in Stan and R. skeptric. (2021, August 22). Retrieved December 8, 2022, from https://skeptric.com/stan-tobit/

Appendix:

Exhibit 1

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	

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		

Exhibit 2

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.