Project 2 Write Up

Mihir Gadil, Elie Tetteh-Wayoe, Jessica Fogerty, Aluya Omofuma, Pierre Bamba

11/17/2018

Contents

Introduction

The Olympic games is international sporting events featuring winter and summer sporting competitions for men and women. The Olympics games is arguably the most prestigious sporting event in the world and its popularity is on the increase. In the 2016 edition held in Rio-Brazil, there were 306 events compared to the very first Olympic games in 1896 held in Athens-Greece which only had 43 events. The growth of the Olympics has been a trend from the early years of the event. In the 1950s, after the second world war and the cusp of the cold war the number of events had tripled to 150 from the very first Olympics. The growth continued till present time although it slowed down a bit. There might be an increase in the number of unique events over the next few years, but it would most likely be minute. The focus of our study is to build on our previous work and create a model that will beable to predict the proportion of medals won at future Olympic Games. Our SMART question is: What proportion of medals will Japan (the next Olympic host country) win during the 2020 Summer Olympic games. Previously we were able to answer the question: Does the host country have an advantage in the olympics? The answer was yes, that the host team preforms better than their performances when not a host team. Drawing on previous Olympic model building research we will be taking the following variable into consideration: GDP per Capita, Population, historical medal counts, planned economy (communist economy), soviet (or was once a soviet) nation, host country (is or isn't).

Exploratory Data Analysis

Olympic Data Set We first started our EDA process by importing the Olympic data from: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results. The first thig we did was add a Host_Country Column. We then noticed that each athlete was awarded a medal including each team member, we were able to subset the data so that each team was awarded one medal so that there was no discrepency. We changed the structure of some of the columns such as medal, changed to ordered factor. We dropped Age, Height, Weight, City and Games. Next we insert host country data into the data frame. Change some country codes to simplify analysis and select only the Summer Olympics data. We Examined the structure of the data frame and then examined the five statistic summary. *Structure: Summary:*

This summary shows us that the United States, Great Britain, and France are the teams that participated the most frequently. The sports that had the highest number of participants were: Athletics, Gymnastics, Swimming. It is important to note that the three types of medals are not equal in numbers. This is because of ties.

GDP and Population Datasets We then imported GDP and population data from: . We selected to keep only data from the year 1896 and earlier, this is when the first Olympic games was. We then calculated the per capita GDP to use as a predictor variables. **Discuss these calulcations further** We the subset the medal counts of host nations and selected data from the year 1988 and later. Based of prior research and out own

prior findings we wanted to include the soviet (or peviously soviet) countried in our model. So we researched which countried are or were soviet and subset those. We did the same for planned (communist) economies. We then found the previous proportion of medals won for each nation.

Model Building

Selecting the Correct Model We started of by splitting our data into training and test groups. The test group was the Olympics in 2016. The training was data previous to the 2016 Olympics. We first began by using all the predictor variables that we discussed earlier (host nation, soviet nation, planned economy, GDP per capita, previous medal proportion) and created a linear model. After trial and error we were able to determine the best predictor variables are: Previous proportion (of medals won), GDP Share, and if the country is a host country. The next, better model that we created used the following predictor variables: host country, planned economy, soviet nation, log(GDP Per capita), log(Population), GDP , GDPShare, previous medal proportion, Total Medals won.

Reliability of results Accuracy rates, the VIF, exc. Anything that proves that this is the correct model.

Predictions we can make By using this model we can predict the proportion of medals that each country will win at future Summer Olympic games with _____ accuracy.

Additional Information or Analysis that might improve the Model Results Including all the olympic years could help improve results. (Need to brainstorm on this).

Conclusion

Can add this after we are done with paper.

Citations

Cite the research paper we based out model off.