Olympics Data Exploration and analysis

Analyze trends in data and make predictions for future Olympics.

Chinmayi P S (01FB15ECS078) Data Analytics, Dept. of Computer Science and Engineering , PES University

Bangalore ,India
chinmayi2398@gmail.com

Devika Mishra (01FB15ECS093) Data Analytics, Dept. of Computer Science and Engineering , PES University

Bangalore, India
mishra.devika23@gmail.com

***Abstract*— This project examines the history and current state of analytics in sports specifically for the summer olympics held every 4 years.**

Keywords—GDP , Olympics , Predictive analytics.

I . INTRODUCTION

Sports analytics is defined as “the management of structured historical data, the application of predictive analytic models that utilize that data, and the use of information systems to inform decision makers and enable them to help their athletes , coaches etc. in gaining a competitive advantage on the field of play.”

Sports analytics was brought to the public eye by the movie Moneyball, a 2011 sports-drama film that portrayed how a baseball coach, Billy Beane rebuilt his team against all odds using empirical data and statistical analyses on players’ performance. His trial with sabermetrics changed the way the game is played forever and made analytics a dream for many.

There are two key aspects of sports analytics - on-field and off-field analytics. On-field analytics deals with improving the on-field performance of teams and players. It digs deep into aspects such as game tactics and player fitness. Off-field analytics deals with the business side of sports. Off-field analytics focuses on helping a sport organization or body surface patterns and insights through data that would help increase ticket and merchandise sales, improve fan engagement, etc. Off-field analytics essentially uses data to help right holders take better decisions that would lead to higher growth and increased profitability.

By using big data in sports to collect every probable data out of an athlete, bringing all the information together and building upon the data driven analytics to strategize the most suitable approach for a win.

II . *PRE-PROCESSING*

We have used multiple data sets like the summer Olympics dataset available on kaggle, GDP dataset available on gapminder etc and preprocessed this data and merged them to create a single dataset in order to build models.
The data we downloaded had details of each medal won by each athlete of the country in individual as well as team events. As the problem statement required the totals medal count for each country for each year , the data had to be sliced and diced accordingly.

The available data was organized using the IOC codes and did not have details about the GDP per capita. Another dataset was downloaded from gapminder containing the details of per capita GDP for each country from 1950 to 2008. The missing values in the GDP data (i.e. 1896 to 1948) were filled using the first simple repetition method of signal extension. This part of the preprocessing was done manually.

In order to map the IOC codes to the countries names another data set having the country names and the corresponding IOC codes was used to merge to the two data sets.
Since no ready made testing data was available (for 2012 and 2016) , Testing_data was created manually .

III. Most relevant predecessor work

WHO WINS THE OLYMPIC GAMES: ECONOMIC RESOURCES AND MEDAL TOTALS

(Andrew B. Bemard and Meghan R. Busse)

This paper examines determinants of Olympic success at
the country level by considering the role of population and economic resources.
The conclusion drawn in this paper was Per capita income and population have identical effects at the margin, suggesting that total GDP is the best predictor of national Olympic performance.

A. LIMITATIONS

1. The number of Olympics considered were very limited i.e. 9 (from 1960 to 1996).
2. One of the questions answered in this paper was: Why does China win only 6% of the medals even though it has one-fifth of the world's population? - But the fact is that china started participating in Olympics only in 1980's and have reached the top of the table within a very short time period.
3. This paper also did not consider the previous year performance of a country for the prediction of future medals.

B. ASSUMPTIONS

1. For team events, the data considers one medal per athlete and not one medal per team which is a norm.

IV. MODELS BUILT

1)SVM :

Support vector machines is a supervised learning models with associated learning algorithms that analyze data used for regression analysis and classification .

2)Multiple regression model:

Multiple regression is an extension of simple linear **regression**. It is used when we want to predict the value of a variable based on the value of two or more other variables. Here we have used the model to predict the value of medal count using the variables i) Year ii) GDP per capita iii) previous year performance .
The t values obtained for the variables are :
i) Year - 3.07251
ii) GDP per capita - 5.04522
iii) previous year performance - 52.15761
Residual standard error: 18.58 , and error for the testing data is 12.899

3)Neural Network :

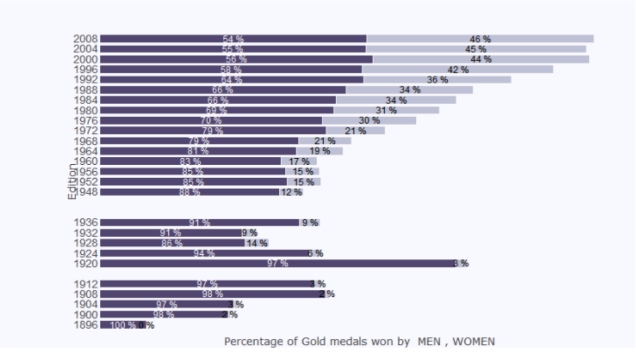
Artificial **neural networks** are forecasting methods that are based on simple mathematical **models** of the brain. They allow complex nonlinear relationships between the response variable(medal count) and its predictors.
(i)Year ii) GDP per capita iii) previous year performance) .
The models has a very high error , this is because the number of input variables is very low and the training data may also not be sufficient. The neural networks model works best for prediction using a large number of input variables , with multiple layers .The structure of the network has to be carefully decided to improve the performance . It is not very suitable for this case.

4)Decision tree - The decision tree model we tried building failed as the number of attributes were not enough for it to work properly.

5) Naive Bayes - Naive Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. The output of naive Bayes model is a set of conditional probabilities, which is not appropriate for this case .

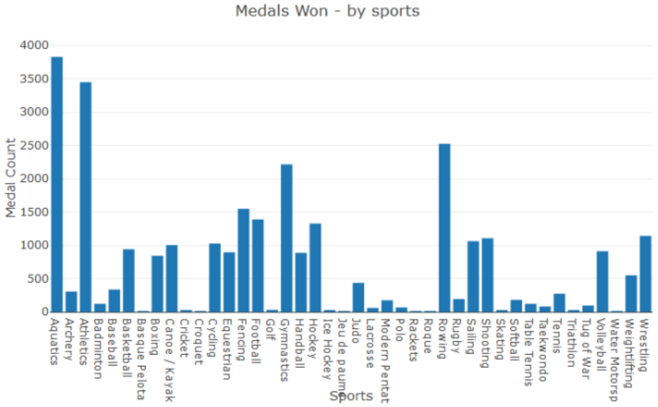
V. INSIGHTS GAINED

The number of women participating in organized sports is continuing to grow around the world, as evidenced by the shrinking gap between the numbers of female athletes and male athletes at the recent Olympics.



After more than 100 years, gender equality is still more goal than reality in the 2016 Rio Olympic Games where there are more men's events (161) than women's and mixed events (145).

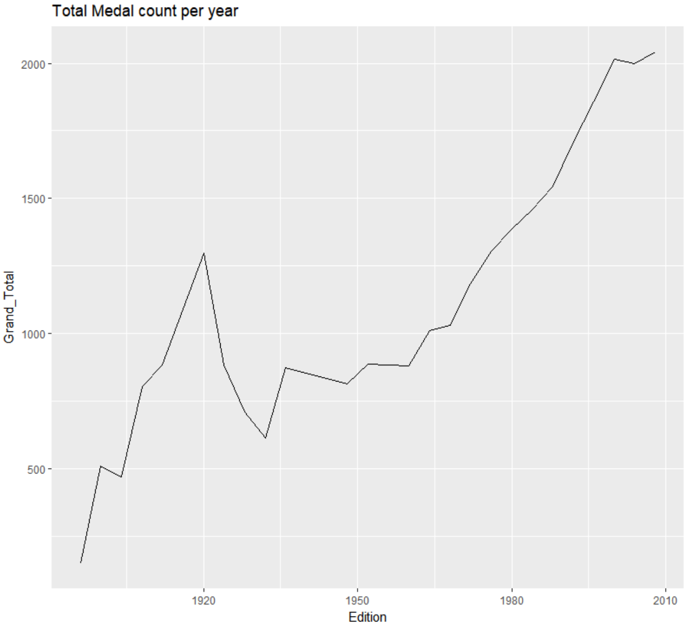
Below is a bar plot of medal count distributed by sports .



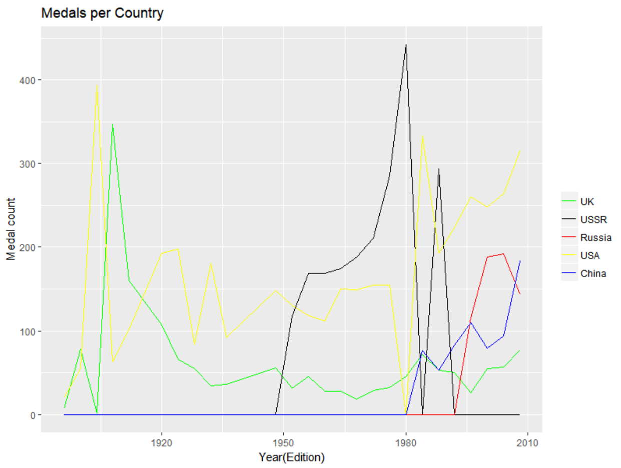
Along with track & field athletics and gymnastics, Aquatics is one of the most popular spectator sports at the Games , which is also clearly evident by the number of events under the category.

Below is a graph of total medals awarded each year at the Olympics .
At every edition of the Olympic games , changes to the competition format or the events themselves have been made .

- But the peak at 1920 , can be because the Belgium Olympics in 1920 were held in hopes of bringing a spirit of renewal to Belgium, which had been devastated during World war I.
- But later there is a dip which may be attributed to the widespread fear of Adolf Hitler and political unrest before world war II.



Below is a line plot of the medals won by the 5 highest grossing countries .
China started participating around 1980’s and since then it has consistently been at the top of the medal table.
The performance of the USSR countries has been fluctuating , mostly because of the predominant political unrest .



VI. CONCLUSION

Different models work differently for a given data. For this particular problem statement multiple regression worked the best with an error of only 12%.

The GDP attribute along with the previous year performance have a significant impact on the performance of a country in the current Olympics. This is probably because the team tactics and strategies are altered based on the performance of each player in the team. Also for individual events the player can improve his performance based on his as well as the performance of his competitors in the previous Olympics.

GDP per capita may not appear to affect the performance of athletes of a country by just the looking at data. But GDP of a country definitely does have an effect on the infrastructure, facilities and sports equipment provided to the athletes which do get affected by the GDP. Also the electricity supply of a country which is again affected by the GDP is a factor which can be considered, as the broadcasting of events and awareness of the sports can be spread across a country using the electronic media and help discover new talents which can prove to be game changers for a country in a particular sport.

The medal count depends on various other factors or attributes (climate , literacy rate , government support - national funds and promotional activities etc.) , this project could be easily extended by adding the most important of these factors . Due to the unavailability of such a data , we have limited ourselves to using only 2 such factors GDP and previous years performance.