**Road to success: Predicting the medals for the Tokyo 2020 Olympics**

Team: Nikhanj Savla, Manoranjan Kanakaraj, Ratheesh Kumar, Tridib Pal

At the outset, we would like to thank the Great Learning team for hosting this exciting contest. It was an opportunity for us to gain hands-on experience, applying what we have learnt over the last six months to a real-life scenario. Read on to understand how we combined economic indicators with past Olympics performance, to build the solution model.

**Approach:**

* Dataset: Results of last 3 Olympics (2008, 2012, 2016). The intent was to base our analysis on the recent trends, as it would be closely reflective of the evolving demographics of athletes/countries.
* Prioritized the analysis for developed countries, given that they account for the bulk of the medal tallies.
* In addition to their past Olympic performance, identified key economic indicators such as ‘health expenditure per capita’ and ‘GDP per capita’, which showed better predictive power in the medals analysis of these developed countries.

**Model Design:**

* Built and tested a simple sample linear regression model using OLS method on the basis of the predictions made for the 2016 games, where we used data from only 2 prior Olympics – the 2012 and 2008 games.
* As the sample model was trained using data from only 2 editions of the games, we were confident that when the final model for the prediction of the 2020 games would be trained using data from 3 editions, the results were more or less only going to better if not similar, and it happened to be so.
* The attributes were finally zeroed in after much of trial and error (the data description pertains to data collected for the 2016 games):
  + medals\_won – medals won by a country in the 2016 games – also the **target variable**
  + total\_medals – total medals won by a country in the history of the games prior to the 2016 games
  + medals\_per\_edition – mean medals won by a country before the 2016 games (total\_medals divided by the number of times the country participated in the games)
  + prev\_medals\_1 – medals won by a country in the previous edition, i.e., 2012 games
  + prev\_medals\_2 – medals won by a country in the edition prior to the previous edition, i.e., 2008
  + athletes – number of athletes that represented/will represent a country in the said edition
  + health – health expenditure per capita of a country for the corresponding year
  + gdp\_pc – per capita GDP (or the Gross Domestic Product) of a country for the corresponding year
* The above data was collected in similar fashion for the other games as well, i.e., 2012 and 2008 – for example, the prev\_medals\_1 and prev\_medals\_2 was were collected for 2008 and 2004 games for 2012 and for 2008 games it was 2004 and 2000.
* The training data for the final model included the above data for the 2016, 2012 and 2008 games with their respective target columns and the test data included the data for the 2020 games without the target column.

**Challenges:**

* Data collection and cleaning – This seemed to be like a never-ending process and took most of the time, nearly 90% of our total effort.
  + The data that we had gathered was from 4 different sources, so it had to be reconciled with a lot of attention to detail – more than a few countries’ names were referred upon differently in different sources. For e.g., ROC or the Russian Olympic Committee, was referred as Russia, Russian Empire or Russian Federation, and so on.
  + The ‘total\_medals’ which indicated the historic count of medals won by a country, needed to be re-calculated for a few countries as they were re-organized. For e.g., East Germany, West Germany, Saar, United Team of Germany and finally Germany.
* The other major challenge was to get the right prediction for one country, China. None of the indices we tried helped getting close to their tally; if the prediction for China turned out well (like when we used the population and GDP indices separately), then the prediction for rest of the countries would go for a toss. Hence, we had to stick with the original format of the model. This meant going in with the risk of predicting the nearest possible tally for most of the top countries, at the cost of China’s predictions.

All in all, this was a great experience for us being the first time in such a contest, and we’re looking forward to more such events. A quick word of appreciation for our program manager Mr. Arun Sharma and our mentor Mr. Mandeep Singh, for their continued support and encouragement.