Forecasting Medal Counts for the 2024 Paris Olympic Games

Forecasting & Predictive Analytics Project Report

Enrico BARCHIESI
Anatole HOMANN
Elliot MASSEY
Alix VERMEULEN

Table of Contents

Introduction

The aim of this project was to forecast the total number of medals each country is predicted to win at the Paris 2024 Olympic Games. Our goal was to see if we could use what we have learnt about forecasting to create a model that would succeed in predicting a country's performance in the Olympic Games from independent variables.

Dataset & Variables

We found a rich dataset on Kaggle which contained data on over 200,000 athletes that competed in an Olympic event and medal results from Athens 1896 to Rio 2016. After extensive analysis of this dataset, we grouped the athletes by country and decided to include only the Olympic Games that had details about the number of athletes for every country. Some countries didn't have any athletes present in certain Olympic Games as they didn't begin participating until a later year or certain countries changed (Yugoslavia split into Croatia, Serbia and Slovenia for example).

In order to check the accuracy of our model we wanted to compare it with the most recent Olympic Games (Tokyo 2020), however this data was not present in our dataset. We therefore scraped Wikipedia for details about the Tokyo 2020 Olympic Games.

The medal count contained in the Kaggle dataset was ambiguous as it contained the total number of medals from both the Olympics and Special Olympic Games. This dataset also included each individual athlete's medal (even if they were a part of a team). For example the United States would gain 4 medals to their total medal count if they won a 4 x 100 metre relay race. These total medal count figures were also hard to verify as there were no other sources online that confirmed the same results, nor any precise description of how the results were calculated. In order to resolve this we decided to scrape Wikipedia for each country and just include the total medals each team won at the respective Olympic Games only. For example France would gain only 1 medal to their total medal count if the France 7's won.

In order to effectively compare the results between Olympic years we decided to express the variables as relative percentages. We did this in order to account for (1) the rising number of total medals issued as more and more sporting events were introduced into the Olympics year after year and (2) the increase of athletes per country to compete in these events. The Paris 2024 Olympics will be introducing 4 new events: break dancing, climbing, surfing and skateboarding.

The next amendment we made to one of the variables came about after we performed an analysis between the number of athletes and the number of medals won. From figure 1 we can see that the initial R^2 was 0.639 however the data points seemed to be following a more non-linear relationship. To better model this we decided to square the number of athletes (creating a new variable **Athletes^2**) and we got a better R^2 of 0.709.

Model

The model we decided fit best with our dataset was an AR-X model. This model contained an Auto-Regressive variable which was the percentage of medals won in the previous olympics (denoted medals_{L1} where L1 stands for lag 1), as well as an eXogeneous variable which was athletes^2. We used a linear regression approach to estimate the coefficients and evaluate their significance.

Assumptions

In order to forecast the number of medals won by each country in the Paris 2024 Olympic Games we had to outline certain assumptions in order to calculate some estimates.

1. **Number of Athletes per country:** Since we do not have access yet to this information (selections for the Olympic games are still ongoing so no data is available on this), we had to estimate this figure. We used a linear regression analysis to record the number of athletes representing each country in the previous Olympic Games which we had access to from our dataset. We then used this trend to more accurately estimate the number of athletes to represent each country at the upcoming 2024 Olympic Games.
2. **Number of Total Medals Issued:** In order to get an accurate estimate of the number of total medals that will be issued at the next Olympic Games we completed the following process: we were able to see how the number of medals issued have increased year on year from our dataset as more events had been added. Similarly, after some research we discovered that 4 new events were going to be introduced for the first time at the 2024 Paris Olympics (break dancing, climbing, surfing, skateboarding). We then took this into consideration and added this to the total number of medals issued factoring in the year on year increase observed from analysing our dataset.

Outcome

With these assumptions in place we were able to input these independent variables into our AR-X model equation :

$$Medals_t = 0.6 \times Athletes_t^2 + 0.3 \times Medals_{t-1} + error_t$$

We obtained an intercept of 0.00052 which is basically 0. This makes sense as a country that sends 0 athletes should receive 0 medals.

Model Evaluation

To analyse our models performance, we trained it on data from the 1964 to the 2012 Olympics (80% of our dataset) and then tested it on the 2016 and 2020 Olympics (20% of our dataset).

The first metric we decided to evaluate was the Adjusted R^2 of our model which was 83.6%. An Adjusted R^2 of 83.6% highlighted that our independent variables ($Athletes^2$ and $Medals_{L1}$) were able to account for 83.6% of the variability of the dependent variable (Tot_Medals).

The second evaluation metric we observed was a Mean Absolute Error (MAE) of 2.69. This denoted that our model on average was 2.69 medals away from the true number of medals. From figure 2 we can see that our AR-X model succeeded in predicting both countries that won a low number of medals (Ireland and India) but also the countries that consistently won a high amount of medals (United States and Great Britain).

Some countries consistently overperformed/underperformed their expectations and our model likewise under/over predicted the amount of medals that these countries would win. China for example consistently overperformed and thus our model underpredicted the medals they would win. Germany and France are examples of countries that displayed the opposite trend.

Error Modelling

In order to try and improve our model to better deal with the anomalies detailed above we tried to analyse the errors of these countries. We initially adopted a time-series model approach in order to try and get a better understanding of these errors, however, after plotting the Auto-Correlation Function (ACF) and the Partial-Auto-Correlation Function (PACF) (Figure 3) we quickly discovered that due to their being no significant lags we would have to adopt a different approach.

Using what we had learned in class we decided our next approach would be to model the errors as a Random-Walk where the error at time t is defined as $error_t = error_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(0, \sigma^2)$. A larger σ^2 would result in a wider range of error values and hence a larger range of possible medal forecasts. We only applied this way of modelling errors to countries that were consistently under/overperforming. We noticed that with our adjusted medal forecasts, our MAE decreased from 3.12 to 2.8.

Other predictors

Before finalising our AR-X model with the chosen independent variables ($Athletes^2$ and $Medals_{L1}$) we analysed a few other variables to see if they improved our model (Figure 5).

Host Country

An especially topical variable would be the influence of the host country on the number of medals won (especially with the 2024 Olympics happening in Paris!). This variable was already present in our original dataset (1 denoting if a specific country was the host in that Olympic Year and 0 otherwise). As would be expected, this is an extremely imbalanced variable with only one country being host to 149 other non-hosts. We were disappointed to see that the host variable only had a 0.0031 impact on the total number of medals won.

Although we didn't have the time to complete this for the current project, we had considered applying resampling techniques to our data in order to balance out the two groups (host/non-host more evenly).

GDP / GDP per capita

Another interesting variable was to see if a country's Gross Domestic Product (GDP) could influence the number of total medals won in an Olympic Year. We also computed this figure per capita so that we could compare countries like the United States to that of Zimbabwe. Similarly to the host variable, it was insignificant in the model.

The Impact of COVID-19

Something which we had not considered but could be interesting to analyse was how countries performed post the COVID-19 pandemic. We had observed that the variance of medals won was noticeably larger in 2020 than in previous years, yet we couldn't confidently conclude that this was due to COVID. Germany spiked our interest after it significantly underperformed compared to its expectations in 2020, but we cannot conclude that this is because of COVID.

2024 Forecasts & Conclusion

Our model enabled us to forecast the following results:

- 1st place - United States (102 Medals)
- 2nd place - China (78 Medals)
- 3rd place - Russia (59 Medals)

We also computed the forecast for other countries of interest (Figure 4).

We really enjoyed completing this project as a part of the Forecasting and Predictive Analytics course. It was extremely interesting to utilise the skills and knowledge we gained from this course and apply it to a real-world example. We, as a team, are all very excited to see how close our model was to predicting the true medal counts of the Paris 2024 Olympic Games!

Figures

Figure 1 - Athletes vs Athletes^2

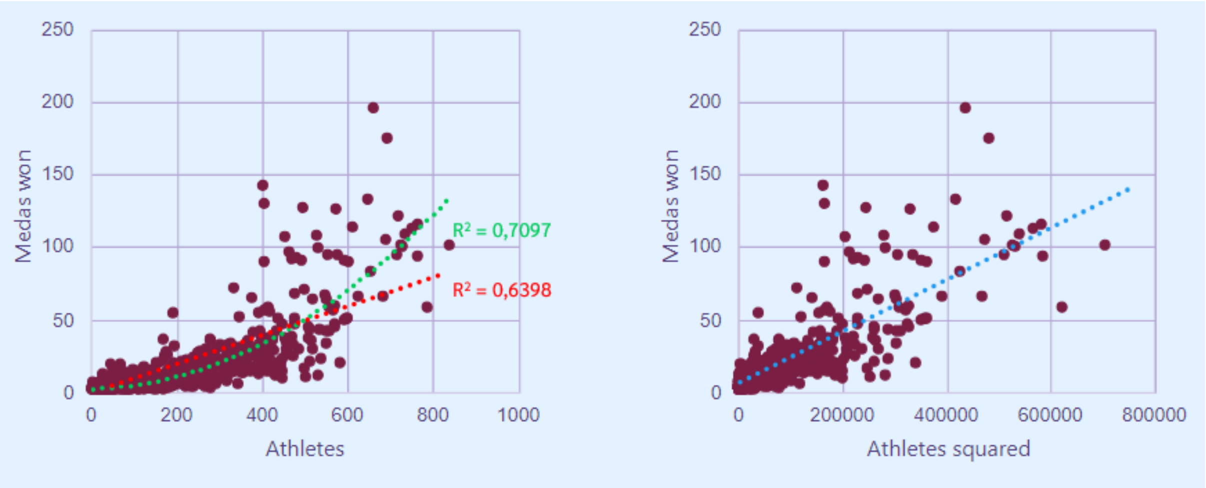
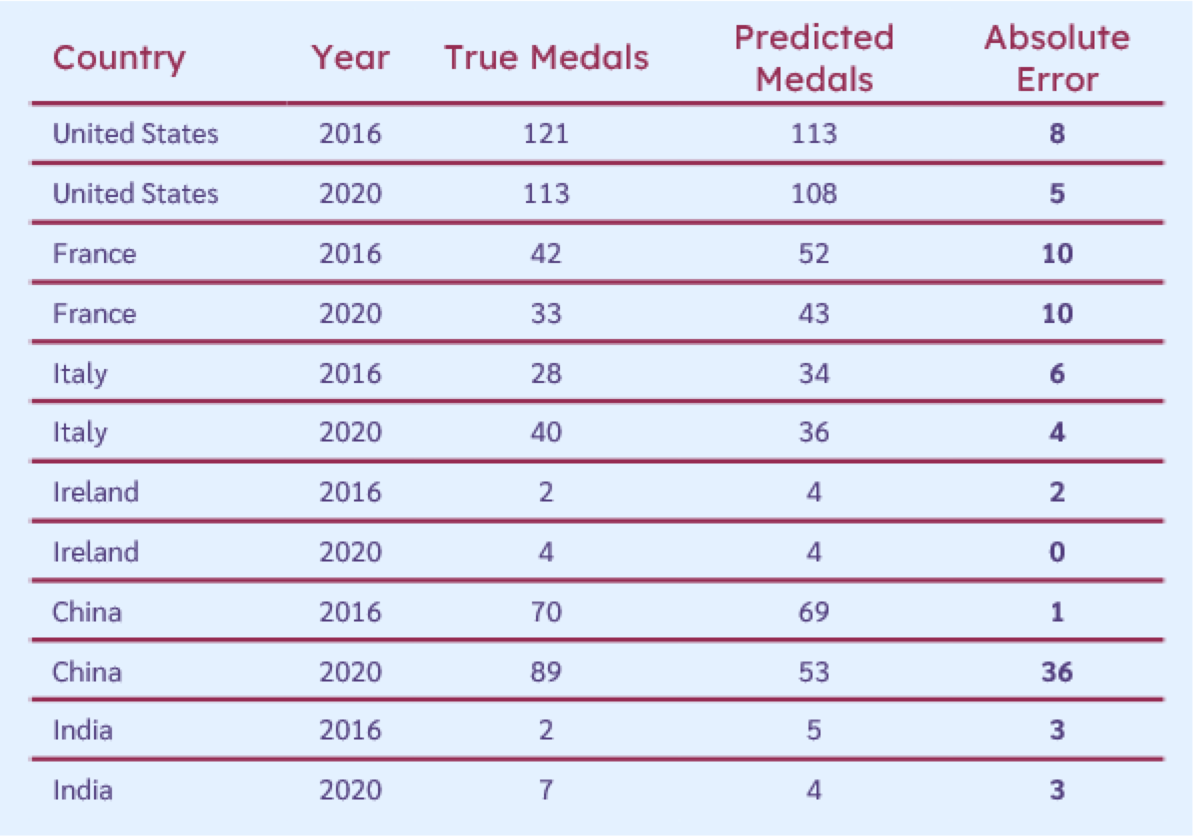


Figure 2 - Model Evaluation



Country	Year	True Medals	Predicted Medals	Absolute Error
United States	2016	121	113	**8**
United States	2020	113	108	**5**
France	2016	42	52	**10**
France	2020	33	43	**10**
Italy	2016	28	34	**6**
Italy	2020	40	36	**4**
Ireland	2016	2	4	**2**
Ireland	2020	4	4	**0**
China	2016	70	69	**1**
China	2020	89	53	**36**
India	2016	2	5	**3**
India	2020	7	4	**3**

Figure 3 - ACF and PACF (China)

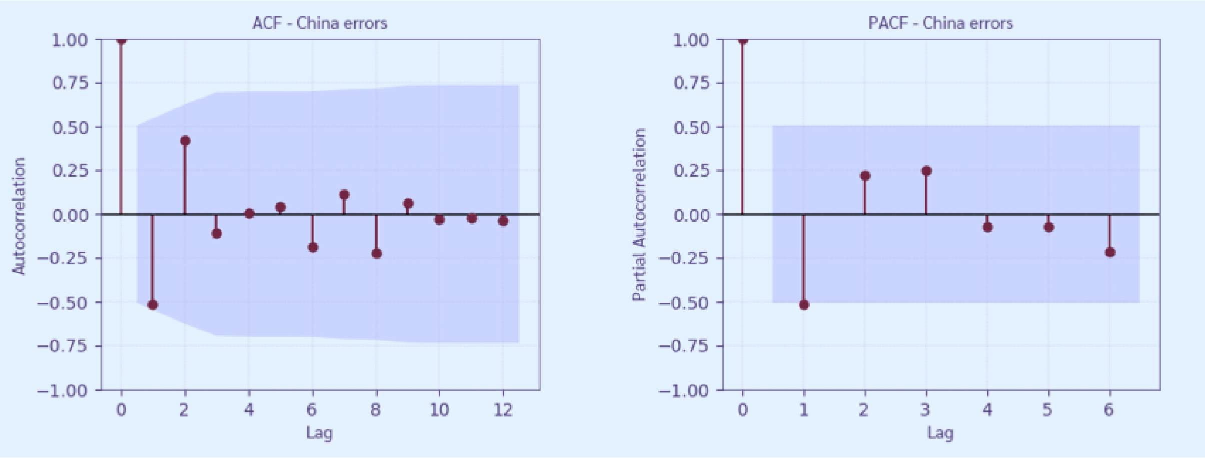
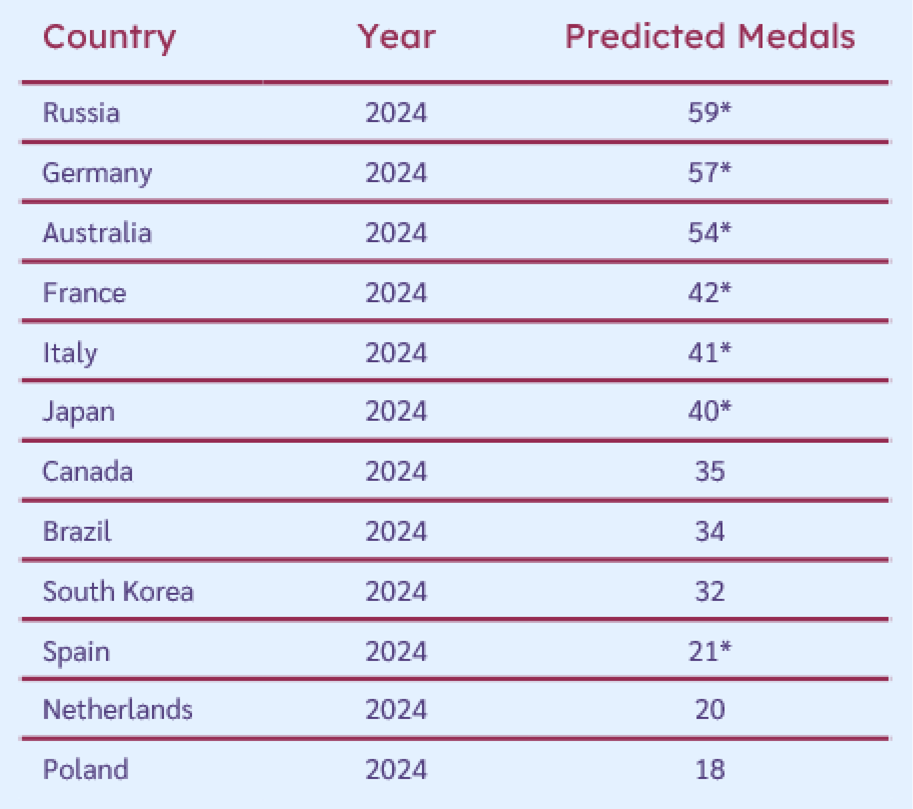
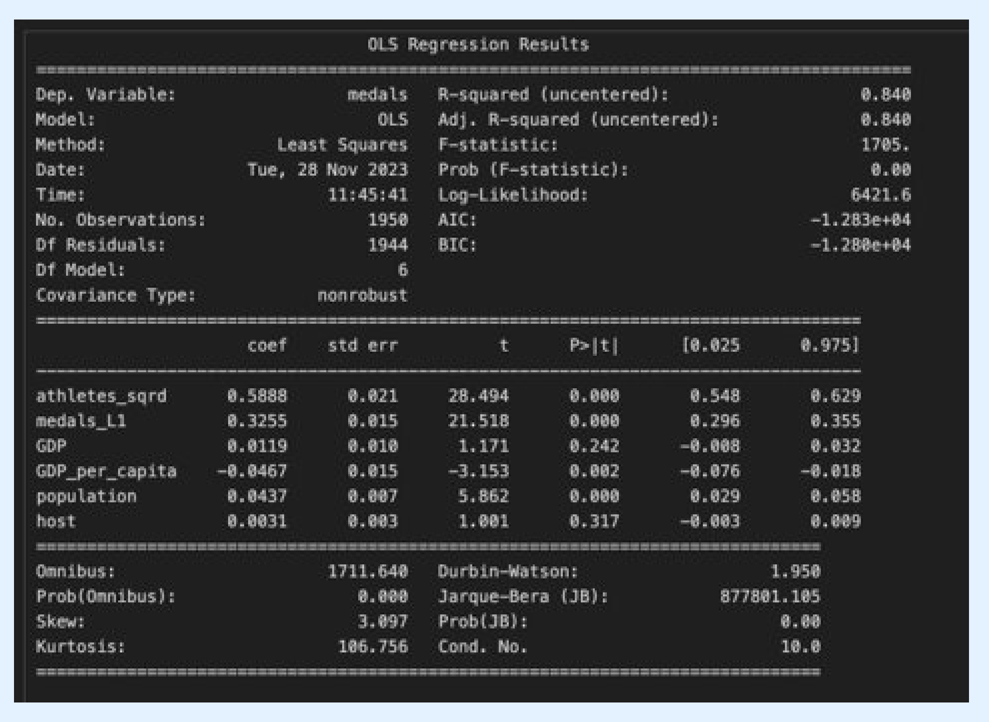


Figure 4 - 2024 Medal Forecasts (*Adjusted with Random Walk Approach)



Country	Year	Predicted Medals
Russia	2024	59*
Germany	2024	57*
Australia	2024	54*
France	2024	42*
Italy	2024	41*
Japan	2024	40*
Canada	2024	35
Brazil	2024	34
South Korea	2024	32
Spain	2024	21*
Netherlands	2024	20
Poland	2024	18

Figure 5 - Analysis of Other Predictors



```
                                 OLS Regression Results
========================================================================================
Dep. Variable:                 medals   R-squared (uncentered):                   0.840
Model:                            OLS   Adj. R-squared (uncentered):              0.840
Method:                 Least Squares   F-statistic:                              1705.
Date:                Tue, 28 Nov 2023   Prob (F-statistic):                        0.00
Time:                        11:45:41   Log-Likelihood:                          6421.6
No. Observations:                1950   AIC:                                 -1.283e+04
Df Residuals:                    1944   BIC:                                 -1.280e+04
Df Model:                           6
Covariance Type:            nonrobust
===================================================================================
                     coef    std err          t      P>|t|      [0.025       0.975]
-----------------------------------------------------------------------------------
athletes_sqrd      0.5888      0.021     28.494      0.000       0.548        0.629
medals_L1          0.3255      0.015     21.518      0.000       0.296        0.355
GDP                0.0119      0.010      1.171      0.242      -0.008        0.032
GDP_per_capita    -0.0467      0.015     -3.153      0.002      -0.076       -0.018
population         0.0437      0.007      5.862      0.000       0.029        0.058
host               0.0031      0.003      1.001      0.317      -0.003        0.009
===============================================================================
Omnibus:                     1711.640   Durbin-Watson:                    1.950
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            877801.105
Skew:                           3.097   Prob(JB):                          0.00
Kurtosis:                     106.756   Cond. No.                          10.0
===============================================================================
```

References

- 120 years of Olympic history: athletes and results. https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data
- Medium. Olympic Medal Numbers Predictions with Time Series, Part 3: Time Series Forecasting. https://medium.com/databulls/olympic-medal-numbers-predictions-with-time-series-part-3-time-series-forecasting-255e732616d5
- Forecasting national team medal totals at the Summer Olympic Games. https://www.sciencedirect.com/science/article/abs/pii/S0169207009002088?fr=RR-2&ref=pdf_download&rr=82fd7fecea2c027f