Problem Chosen	2025	Team Control Number
C	MCM/ICM Summary Sheet	2528357

Organized once in every four years, the Olympics set a prestigious stage for countries and athletes to demonstrate their physics and athletic prowess. It brings together the people from all over the globe together to share their cultural values and to share their enthusiasm for sports, and it also acts as a great contributor to maintaining peace and harmony among countries. While most people are interested in individual sports of their interest, often, the overall total medals earned by the countries are also compared. Though the IOC doesn't release official ranks, people compare the performance of athletes, teams, and countries based on those listed medals counts. Similarly, predicting the total number of medals a country will win beforehand is also very important. It is a complex problem for the multitude of factors ranging from individual athletic-level factors to country-specific factors. In this challenge, we use the provided historical data of the medal count of countries to study and predict the total number of medals a country would win in the following countries.

After extensive research, trying, testing, and cross validating multiple models, we shifted from simple linear regression models to ARIMAX Model to finally Random Forest Regression Model. Further playing around with the model, we carved out multiple important parameters that greatly improved the accuracy of our prediction models.

Random Forest Regression Model: Using the total number of athletes, total number of medals won by the country in previous years, rolling mean of medal counts, host status, and sport effect, we developed a random forest regression model, trained it on the previous data, checked its accuracy in test set, and predicted the total number of medals a country will win in 2028. We split the first 80% of the dataset as a train set and the rest 20% of the dataset to test the mode. We calculated mean square error, mean absolute error, and R-squared error values to assess the accuracy of the predictions made by our model against the actual values in the test set of the given data and found them to be 6.87, 0.55 and 0.95 respectively for total medals count, and and 0.66, 0.16, and 0.96 respectively for the gold medal count. It reflects that our model captured the trend of the number of gold medals better than the total medal count, and the prediction for total medal counts of the countries based on provided information is also within acceptable range. We have further included a detailed interpretation of the predictions in the report.

Further Case Study: We further analyzed the provided data, browsed through the internet for further information, and analyzed the data from our models to infer further information about performance of countries in the Olympics. We analyzed two cases of host effect given in the problem, and searched for similar pattern in the historical data. Finally, we suggest that the three countries and sport types identified by this approach are the female swimming team of China, the female football team of Germany, and the men's football team of Morocco.

Ultimately after thorough research and careful consideration in the give time frame, we implemented multiple prediction models and ended up with the Random Forest Regression model that predictions better than all the models we tested. In addition to the predictions, we have analyzed the data to curate conclusions from the data.

Contents

1 Introduction

The Olympics is the most prestigious multi-sport event that features thousands of athletes from hundreds of countries. It is a platform to gather together the people from all over the world to demonstrate their physical prowess, share their cultural values, and promote solidarity. It builds an athletic and national spirit that brings people with shared enthusiasm for sports together to shape peace, build friendship, and draw lines of respect across borders.[1] Yet, it is a competition and the competitors are compared based on their performance. While the International Olympic Committee (IOC) does not officially recognize a ranking, they do publish medal tallies for informational purposes.[2] So, a most basic and straightforward way to compare the performance of a country is their medal count published in those tallies. Countries, teams, and even athletes are celebrated on the basis of how many medals a country wins, if they win gold, silver or bronze medal, and if they win a medal in the olympic for the very first time.

Beyond pride, passion, and patriotism comes economy. Predicting which country, which team, or even which athlete wins the medals in which games of olympics is a significant driver of the economy. It helps sports brands, media and entertainment industry, hotels, online marketplaces, and many others to make investment decisions.[3]

Thus, a robust prediction model that can predict how many medals a country will win in the following olympic games is very important for multiple factors.

1.1 Problem Summary

While the talent and ability of an athlete is significant in itself, there are other multitude of factors that affect the number of medals a country will win in an olympic game. Those factors can be taken into account to predict the medal count of a country in any olympic game. Most researchers point out the socio-economic factors like GDP and population of the country to predict the performance of a country in olympics.[4] Since this MCM challenge required us to make the predictions based on only the provided data, we couldn't use GDP and population into account, rather extracted multiple features from the provided historical medal count data to make the predictions of a country's performance in the following olympic matches 2028.

From the provided information, our goal was to find out multiple insights about the performance of the participating countries in the following olympics. Our goals were to analyze the provided data and make predictions for 2028 to find out the number of Gold medals and total medals a country would win, to find out if any country would win a medal for the first time, to investigate the "great coach effect" in the data and suggest if any country should invest in great coach for the upcoming olympics, and to explore the effect of the number of specific sports on the medal count of a country.

1.2 Model

This was a historical data analysis and forecasting challenge that required us to predict the performance of a country in the following olympic games. Our statistical and learning models study the historical data, find out the significant parameters, and make predictions for future data. We used the AutoRegressive Integrated Moving Average with eXogenous variables (ARIMAX) model and Random Forest Regression Model to predict the number of medals a country will win in the following olympics. We address the limitation posed by the limited data availability and small dataset in making the predictions. We conducted a case study on the medal count patterns to provide other insights such as the great coach effect.

2 Assumptions

The given dataset includes data from various countries spanning 1896 to 2024. Over this period, significant geopolitical and economic changes, especially during and after the World Wars, have greatly influenced Olympic performances. Some countries no longer exist, while new ones have emerged. Based on our observations, we make the following assumptions:

- Several geopolitical entities no longer exist or compete in the Olympics, and we have excluded them from the dataset. These include the Russian Empire, Soviet Union, Czechoslovakia, Yugoslavia, British West Indies, Netherlands Antilles, and Australasia. Similarly, teams like the Unified Team (1992 Olympics), Independent Olympic Participants, and the Refugee Olympic Team, which are not tied to specific nations, have also been excluded. Additionally, due to ongoing sanctions, Russia and the Russian Olympic Committee (ROC) are excluded from the 2028 Olympics and our dataset.
- A total of 25 medals were awarded to mixed teams in the first three Summer Olympic Games (1896–1904) and one medal in the 1924 Winter Olympics. Mixed teams represented athletes from different countries competing together, which is no longer the norm. Hence, we will exclude these data points.
- We standardized naming conventions in the dataset for consistency. Bohemia is treated as part of the Czech Republic, Ceylon as Sri Lanka, and Formosa as Chinese Taipei. References to East and West Germany were initially unified, but only East Germany's data is retained under "Germany" due to its stronger performance. Yugoslavia is reflected as Serbia and Montenegro, while Macedonia is referred to as North Macedonia. United Kingdom replaces all mentions of Great Britain, and Malaya is standardized to Malaysia.
- We excluded data for the canceled Olympics in 1916, 1940, and 1944, as well as the 1906 Intercalated Games, which are not officially recognized by the IOC.
- We assumed that the type of event does not impact the number of medals achieved. Instead, we prioritized the total number of sports events over individual event types.
- We assumed that the city in which the game is organized has no impact on the medals. Therefore, we excluded the city from consideration in our model.

3 Data

The dataset comprises five files, with one file serving as a reference that provides information about the others. The athletes dataset contains details about the medals won by athletes in specific sports and events. Upon analysis, we noticed that while the events vary, the same athletes often compete in multiple events within the same sport. For instance, athletes participating in the 100m race may also compete in the 200m and 400m races.

Additionally, we made the following assumptions:

- To simplify the data, we combined all events within the same sport for each country in a given year and calculated the total number of medals won in that sport.
- Athletes who have participated in four or more Olympics are unlikely to return for subsequent games.

- Athletes who have consistently won medals in previous Olympics are more likely to secure a medal in the next Olympics.

3.1 Host Effect

In 1904, 1932, and 1984, the USA hosted the Olympics, which is reflected in noticeable spikes in the number of medals won by the country during those years. Similarly, we observe spikes in medal counts for other host countries, such as France in 2024, Greece in 2004, and Australia in 1956 and 2000. This aligns with the "host advantage," where host nations tend to perform exceptionally.

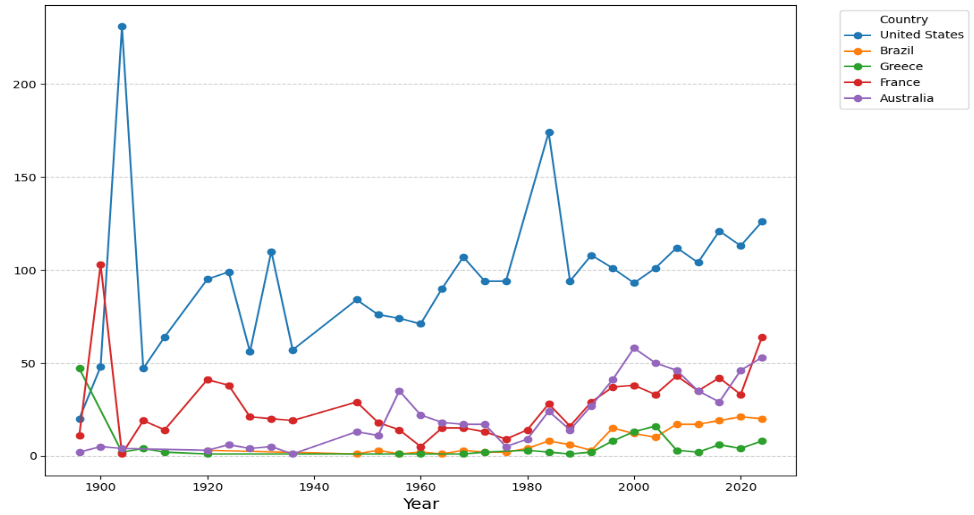


Figure 1: Total Medals for each Year

To account for this, we assigned larger weight to the host country feature in our model. Additionally, we extracted a new feature indicating whether a country is set to host the next Olympics. This allows us to analyze whether countries perform better in the Olympics preceding their hosting year, possibly as a result of increased preparation and investment in their athletes, knowing they will be the next host.

3.2 Number of Athletes

The graph below shows the relationship between the number of athletes and the number of medals won, specifically for the United States in Wrestling. From the graph, we observe a clear trend: as the number of participating athletes increases, there is generally a corresponding rise in the number of medals won.

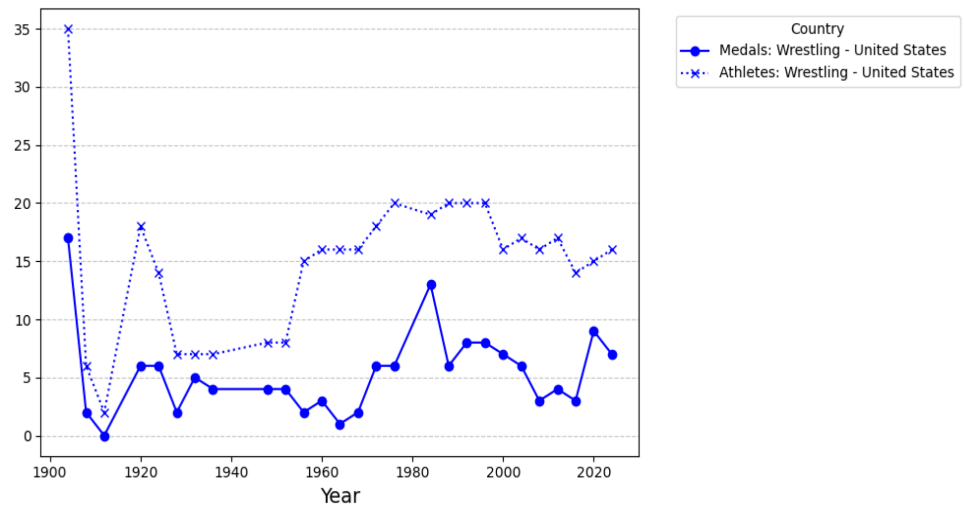


Figure 2: Number of Medals Won and Number of Athletes in Wrestling from US

This trend highlights the impact of larger athlete delegations on a country's performance in a specific sport. It suggests that a greater number of athletes competing in various events within the same sport improves the overall chances of winning medals. This insight reinforces the importance of factoring in the number of participating athletes as a significant feature in our analysis.

3.3 Number of Sports

We also explored whether there is a relationship between the number of events held within a sports category in a given year and the number of medals won in that category. In the graph below, we illustrate this relationship for the United States in Wrestling.

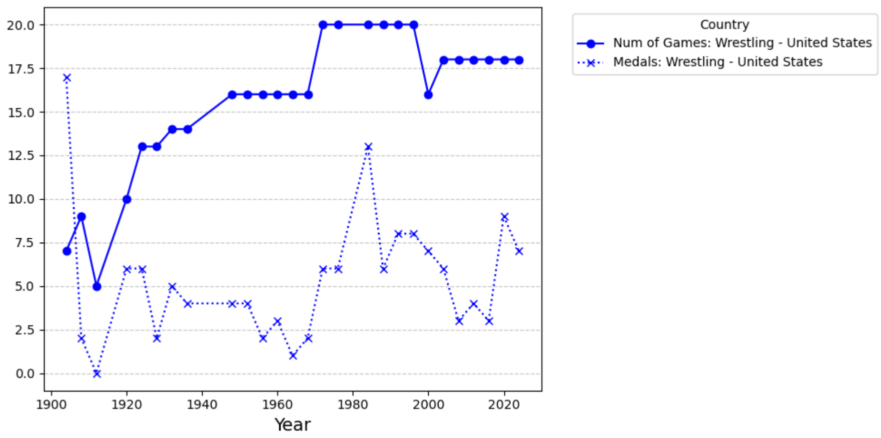


Figure 3: Number of Events Held and Number of Medals in Wrestling from US

From the graph, we observe a pattern: as the number of events within the category decreases, there is a sharp drop in the number of medals won, and vice versa. This trend suggests that the total number of events significantly impacts a country's medal count in a specific sport.

4 Prediction model selection

Due to the very small size of our dataset, there is a high risk of overfitting as the data has a high bias. Due to these limitations, we initially developed a linear regression model as a baseline model for our prediction problem. Since this baseline resulted in overfitting of the data, we chose an ensemble machine learning approach and also tested on ARIMAX which is a statistical time series model that incorporates exogenous variables. While ARIMAX captured some of the time series dynamics, the ensemble method provided a better balance between bias and overfitting, resulting in improved predictions.

4.1 Data Pipelining

From our data, we selected the most important features based on intuition, contemporary research, and manual feature engineering based on the performance in the Linear Regression model.[5] Name of Country (NOC) and Year were selected as identifiers. Total athletes of the countries, Host status, Lag Medal Count, and Winning Ratio were selected as the most important features for creating a baseline with the Linear Regression model. Total athletes of NOC was created by grouping each unique athlete with respect to NOC and Year from the athletes dataset. Lag medal count was created from the total medal count by shifting the medal count to a set number of Olympics cycles before the current medal count. A rolling mean feature of the lag features can be included for capturing the smooth trend of the lag features. Host status was handled by mapping the year and NOC value and was binary encoded as 1 for the country which is the host in the corresponding olympic year and 0 if not. Similarly the Future host effect feature was represented by taking the lag value of this host effect feature. This future host effect was created in consideration to the fact that the host country gets informed about their host status 7 years prior to their hosting olympic year so we decided to incorporate this feature to see its impact on prediction for our model.

The best shift value for the Lag medal count was calculated by testing between the different lag medal counts for individual countries on the ARIMAX model. Winning Ratio was created by dividing the Total Medal count by Total Athlete. This feature was selected because it effectively represented the quality of athletes by year and NOC. The prediction feature or the target variable for our approach was either the total number of medals or the total number of gold medals. Lag medal and Winning Ratio are both features dependent on the target variable and were adjusted based on the variable we are predicting. Both these target variables were tested using the Random Forest Regression, ARIMAX, and Linear Regression Models.

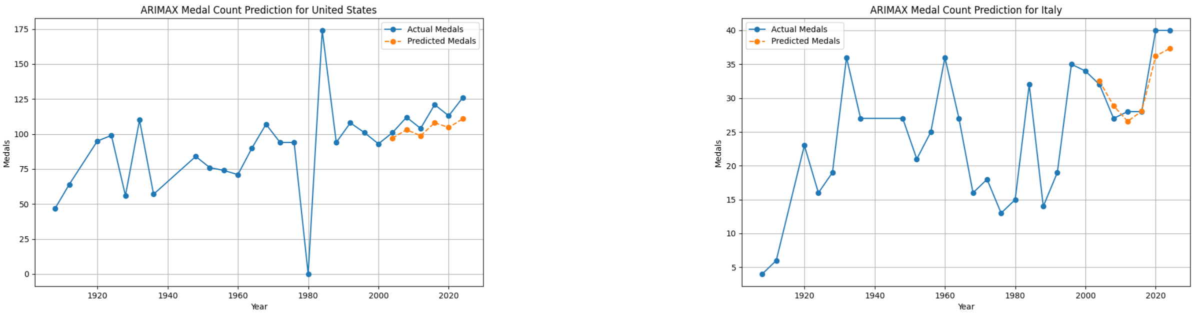
Lag medal count, which is the previous Olympic medal count of the Name of Country (NOC), was later considered an important feature. Lag medal count and Winning Ratio were also considered important features based on manual feature engineering and their positive influence on Mean Absolute Error (MAE), Mean Squared Error(MSE) and R-squared values on the ARIMAX model. The features like sex and city were ommited because they depended on specific other socio-economic factors which the dataset didnt contain.[**?**]

4.2 ARIMAX Model

The prediction of Olympic medal counts is a complex problem that is affected by a multitude of factors. The olympic medal counts have temporal dependency, i.e. we can predict how a country will perform in future based on how it performed in the past. This creates an autoregressive component in the data. In addition, there can be short-term fluctuations in the medal counts due to factors like host effect, great coach effect, socio-economic effects, and other athlete-level effects. This creates a moving average effect on the data. Additionally, countries show long-term trends in their medal counts that requires integration to make the data stationery. Additionally, there are other factors like winning ratio, number of events, and sport-specific effects on the total number of medals a country would win in the following years. All of these features are best captured by this AutoRegressive Integrated Moving Average with eXogenous variables (ARIMAX) model.[6] The ARIAMX model has mainly three parts—autoregressive (AR) part, moving average (MA) part, and differencing part. The AR part models the relation between current medals and past medals. The MA part models the relation between the current medals and past unexpected variations to capture exceptional performance of certain countries. Lastly, the differencing part accounts for the recent general improvement of a country. Mathematically, ARIMAX model can be represented as:

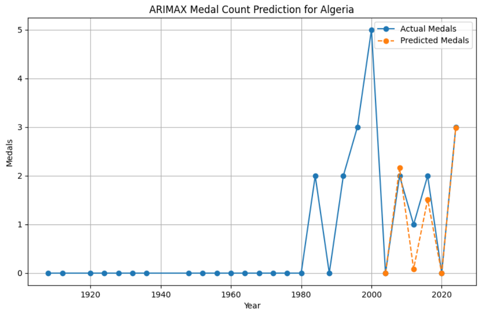
$$Y_i = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_i \epsilon_{t-j} + \sum_{k=1}^{d} \beta_k X_{t,k} + \epsilon_t \tag{1}$$

where, $X_{t,i}$ represents the set of exogenous variables that we have considered for prediction in the model, which included total number of athletes, total medals won by the country previously, host status, and winning ratio of the country. The Y_t is the dependent parameter in our model, which is the predicted total number of medals of the country. The ϕ_i is the autoregressive part, the θ_i is the moving average part, and the β_i is the integration difference part. We split the country-wise data into train and test sets, and ran Bayesian optimization for hyperparameter tuining to get a better and precise prediction.

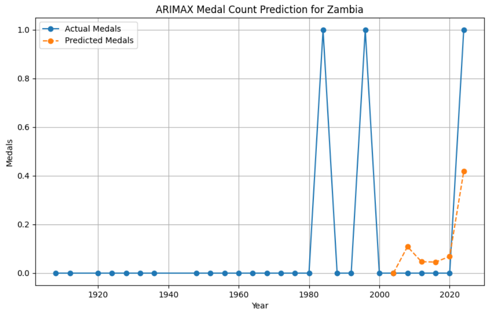


(a) Total Prediction of United States

(b) Total Prediction of Italy



(a) Total Medal Prediction of Algeria



(b) Total Medal Prediction of Zambia

4.3 Random Forest Regressor

Random Forest Regressor (RFR) is a decision tree based ensemble learning algorithm that can capture the non-linear pattern of the data while being able to handle both categorical and numerical data. RFR also effectively handles the problem of overfitting due to its use of Bagging (Bootstrap Aggregating). Bagging ensures that individual decision trees are trained on different subsets of the data, which effectively handles the problem of overfitting.[7] The large number of decision trees and the random sample of data on which the trees are trained essentially make RFR an approach with minimal overfitting and bias.

The feature selected for the input parameters were rolling mean, total athletes count, lag medal count, Host status. Extensive feature selection was done by validation with the ARIMAX model and also by manually testing the set of potentially important features. One interesting fact our model showed was that the future host effect which many contemporary research pointed out to be important didnt affect the metrics like Mean Squared Error(MSE) or R-squared which measures Bias and overfitting. Since Random Forest is a Descision tree learning model, we did a feature importance test where feature importance is a normalized score for measuring how much each feature contributes to reducing Impurities across all trees in the ensemble.

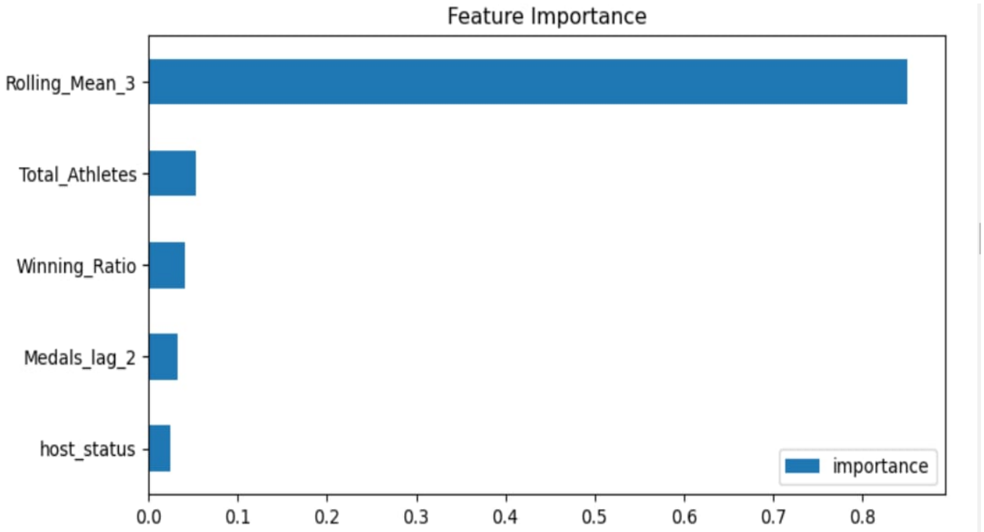


Figure 6: Feature Importance

The Hyper-parameters selected for the model were the n estimator(a) which is the number of descision trees, max depth(b) which specifies the longest path from root node to leaf node, min samples split(c) which represents the number of samples required to split an internal node in a descision tree, and min samples leaf(d) which represents the number of samples required. Tuning was done in these hyper-parameters by applying extensive GridSearchCV with a cross validation fold of 5. The best hyper parameters were found to be [a,b,c,d] = [200,10,1,2]

A Random Forest Regressor is made up of B decision trees. For a new input x, Each tree gives its own prediction. The final prediction is the average of all these B predicitons. This show in the prediction equation (2) below.

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{2}$$

where, B is the total number of trees, and $T_b(x)$ is the prediction of the b^{th} tree.

The Mean Square Error is an estimate of how close predictions are to actual values.

$$MSE(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

where, y is the actual medal count, $\hat{y}$ is the medal count predicted by our model.

5 Results and Discussions

In this section we discuss about the results of prediction models in 4

5.1 Random Forest Regression Results

The prediction of total olympics medal and total gold medals won by each country are shown below. The top 5 countries were selected to properly showcase the fitting and the projections made by RFR in fig 5. and fig 7. meanwhile the projection of the all the countries for every year is shown in fig 6. and fig 8. these plots shows that the RFR can properly capture the trend of the medal counts of countries based on the selected input features. The MSE, MAE and R-squared values for the total medal predictions were 6.87, 0.55 and 0.95 and for gold medal predictions were 0.66, 0.16, 0.96 showing that the model captured the trend of number of gold medals better than that of total medal count. This might be the case because countries who dont have a history of winning gold will have a harder time getting gold medals due to factors like gdp and population size making the trend of gold medal winners easier to capture.

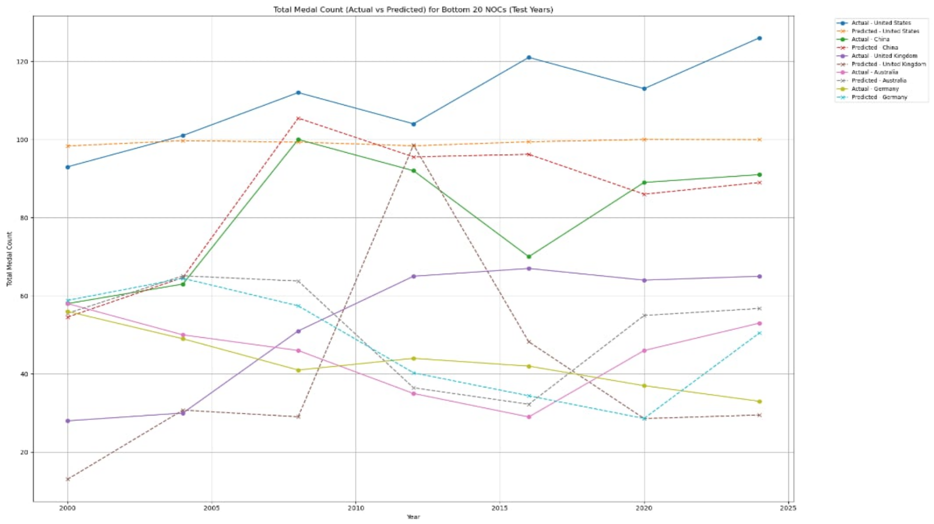


Figure 7: Total Medal Count (actual v/s predicted) of top 5 countries

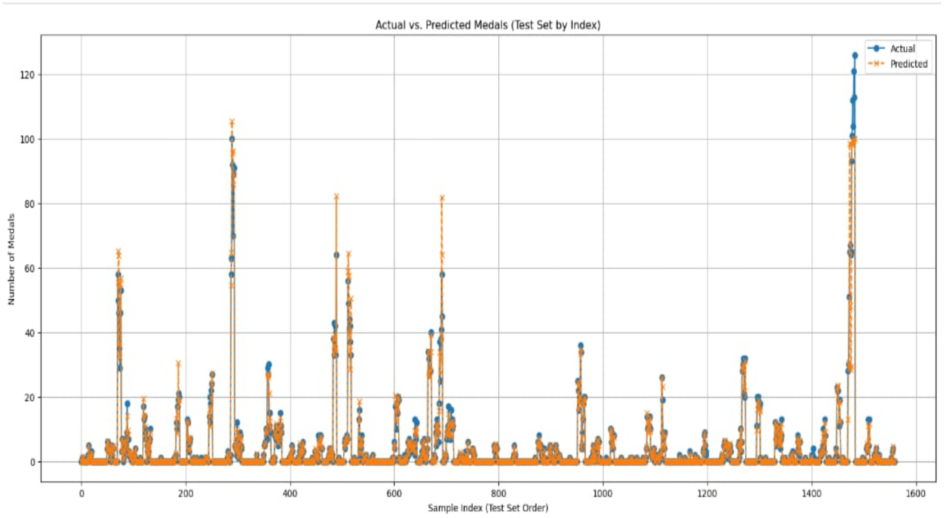


Figure 8: Actual v/s Predicted total medals (Test Set by Index)

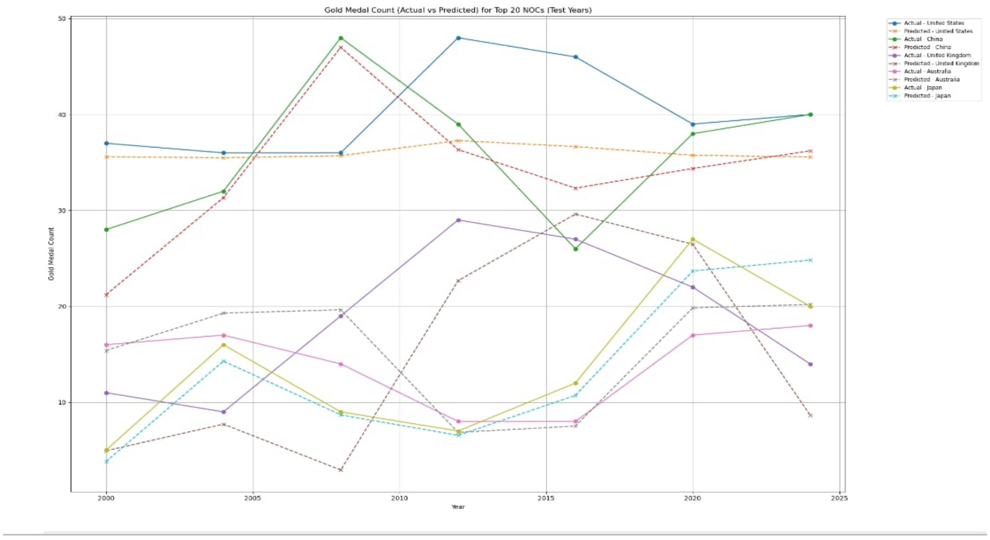


Figure 9: Gold Medal Count (actual v/s predicted) of top 5 countries

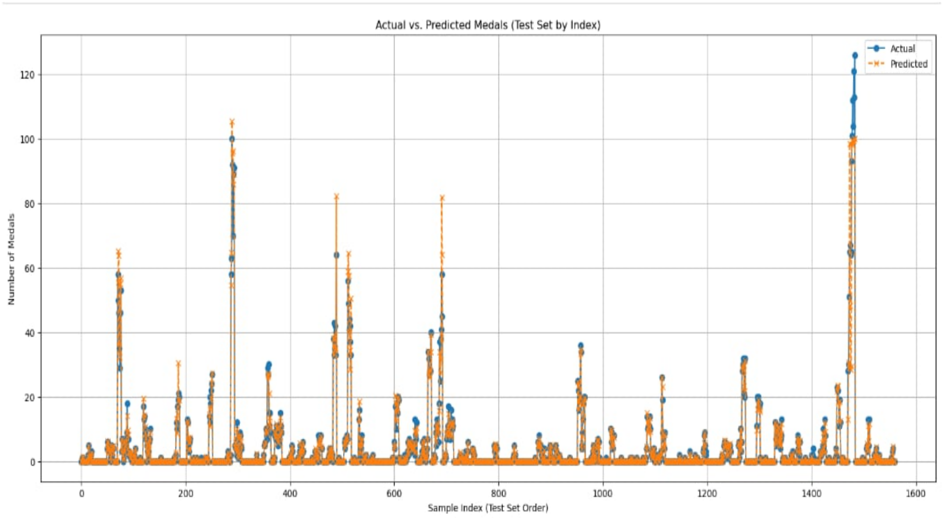


Figure 10: Actual v/s Predicted gold medals (Test Set by Index)

The projections for the 2028 olympics held in USA is shown in the table below it includes projections from the RFR model.

The projections for the countries that have yet to earn a medal is included in the table above and through manual inspection we have projected that only 8 countries had a total medal prediction of above 0 and lag medal value of 0 what that meant was that only these 8 countries had never won an olympic medals before 2024 in that list. And of those 8 countries all of them had a won an olympic medal atleast once. So what we concluded was that our model predicted that none of the countries who have never won before will win in 2028.

Table 1: Total Medal Prediction for top 20 countries for 2028

NOC	Total Athletes	Host Status	Rolling Mean (3)	Winning Ratio	Medals$_{l}ag_{1}$	Predicted Medals
United States	619	0	120.0	0.2	126	90.73
China	396	0	83.33	0.23	91	86.55
France	601	0	46.33	0.11	64	68.37
Japan	430	0	48.0	0.1	45	66.16
Australia	474	0	42.67	0.11	53	64.96
United Kingdom	0	0	65.33	0.0	65	62.89
Germany	457	0	37.33	0.07	33	61.71
Italy	397	0	36.0	0.1	40	41.43
Netherlands	288	0	29.67	0.12	34	33.31
South Korea	147	0	24.33	0.22	32	31.98
Canada	332	0	24.33	0.08	27	28.28
Brazil	289	0	20.0	0.07	20	19.74
New Zealand	208	0	19.33	0.1	20	19.26
Hungary	177	0	18.0	0.11	19	19.15
Spain	400	0	17.33	0.04	18	15.72
Ukraine	141	0	14.0	0.09	12	12.73
Uzbekistan	88	0	10.33	0.15	13	12.24
Kenya	74	0	11.33	0.15	11	11.73
Cuba	61	0	11.67	0.15	9	10.86
Sweden	125	0	10.33	0.09	11	10.5

Table 2: Gold Medal Prediction for top 20 countries for 2028

NOC	Total Athletes	Host Status	Rolling Mean (3)	Winning Ratio	Medals Lag(1)	Predicted Medals
United States	619	0	41.67	0.06	40	38.81
China	396	0	34.67	0.1	40	36.45
Japan	430	0	19.67	0.05	20	24.02
Australia	474	0	14.33	0.04	18	21.15
France	601	0	12.0	0.03	16	15.79
Germany	457	0	13.0	0.03	12	15.74
Netherlands	288	0	11.0	0.05	15	14.48
United Kingdom	0	0	21.0	0.0	14	12.51
Italy	397	0	10.0	0.03	12	11.02
New Zealand	208	0	7.0	0.05	10	10.29
South Korea	147	0	9.33	0.09	13	10.03
Canada	332	0	6.67	0.03	9	8.18
Uzbekistan	88	0	5.0	0.09	8	7.05
Hungary	177	0	6.67	0.03	6	6.38
Kenya	74	0	4.67	0.05	4	5.15
Spain	400	0	5.0	0.01	5	5.14
Sweden	125	0	3.0	0.03	4	4.22
Ireland	143	0	2.0	0.03	4	3.81
Norway	109	0	2.67	0.04	4	3.7
Cuba	61	0	4.67	0.03	2	3.56

5.2 Strengths

Our model performs very well on metrics like MSE,MAE and R-squared. This showcases that our model can capture non linear and complex patterns in the data. The validation MSE also steadily decreases as more training samples are added which means that through better feature selection and engineering and better parameter tuning overfitting can be reduced.

5.3 Challenges

There were several challenges that we faced particularly due to the small dataset and the limited time that we had for developing the model. For the last part of the prediction problem where the goal was to explore the relationship between sports and medal count. Due to the time constraint and our feature testing finding that features related to sports couldnt be properly captured by our model we unfortunately couldnt asses this metric.

5.4 Weakness

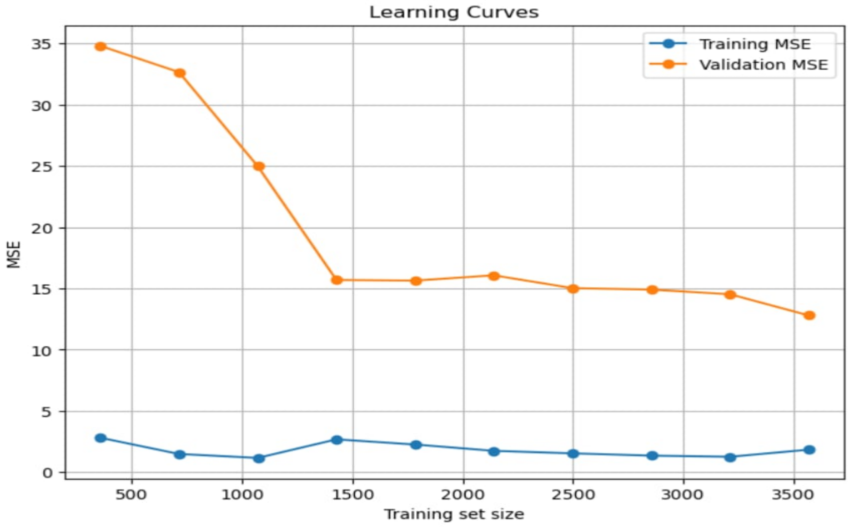


Figure 11: Actual v/s Predicted gold medals (Test Set by Index)

The above is the cross-validation of our RFR model and it shows that there is still some over fitting as the curves dont converge this is due to the extremely small dataset and lack of relevant features for capturing the target variable of predicting medals.

5.5 Sensitivity

The SHAP summary plot below shows how each feature affects the RFR model's medal prediction. Interpreting this SHAP summary plot is simple. The further the shap values are from the 0 means that the more feature contributes to medal prediction. Rolling mean has the largest spread of SHAP values so it contributes the most to medal prediction also pointed by the feature importance bar graph. Winning ratio, medal lag and total athletes all have moderate impact on medal prediction. Host effect have a very small range showing that it impacts the medal prediction the least among the chosen feature. This is also shown in the feature importance graph and also through the projection of United states where they are the next host for 2028 but it still projects that the United States will have fewer number of total medals in 2028 when they are the host.

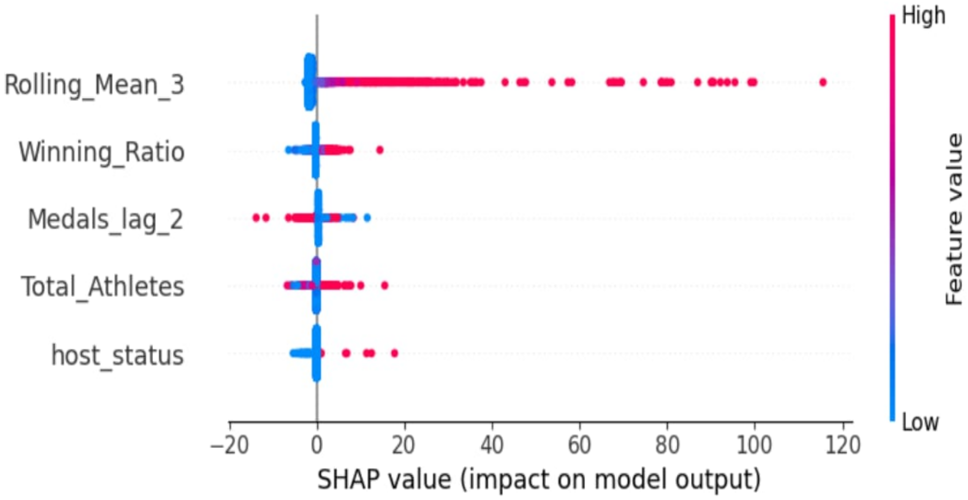


Figure 12: SHAP Summary Plot)

5.6 Discussion

There are several improvements that can be made on our current model. The features selected can be improved to include one-hot encoding of several categorical data to better capture the linear patterns of our data. Also proper tuning of the parameters and possibly an unsupervised approach can be explored for feature selection and engineering to handle better overfitting of data.

6 Great Coach Effect

Due to the lack of data about coaches in our dataset and the limited time we had for this contest, we decided to perform a case study on the two given examples of the Great Coach Effect. In both cases, we found that the total medal count had little impact on the Great Coach Effect (GCE). We inferred from the data that a Great Coach would train athletes to be good enough that even when they leave the team, their lasting effect would not significantly drop the medal count but instead drop the quality of the medal, and vice versa. From this inference, we can surmise that a sudden large change in the quality of medals would be a GCE. As such, the feature ”Total Gold Medals of NOC in an Olympic year” would best represent the Great Coach Effect. Because of the time constraint, we couldn't validate these claims by testing them in our RFR model. However, an approach to capture this effect in RFR would be to create a Lag feature for gold, silver, and bronze medals and, similar to the winning ratio (referenced above), create a gold-bronze and gold-silver ratio to capture the change in the quality of medals. Then, set weights for the gold-bronze and gold-silver ratio features. The target value, in this case, should capture the Great Coach Effect. One assumption made is that Great Coaches/GCE will not be particularly successful if there is no potential in the team.

To address the question of countries and sports where they should consider investing in a Great Coach, we begin by making the assumption that this effect works best when there is potential for the team to improve. The term ”potential for the team to improve” is captured by the feature of the number of bronze and silver medals. Thus, a country not successful in getting medals in a particular sport or with a low number of athletes in the sport will benefit little from a Great Coach.

6.1 Data Pipelining for Great Coach Effect

To identify the countries that would benefit most from great coaches, we begin by grouping the athlete dataset by NOC, Sport, and Sex as identifiers. The number of athletes in a sport by male and female category can then be calculated by summing up the unique athletes. The number of gold, silver, and bronze medals can be calculated by summing up each entry. Now we can give a coach benefit score for each sport type of a country by taking a weighted sum of the number of bronze and silver medals. Because of our assumption, this approach heavily favors successful athletes, as seen in the difference between the 4th and the 3rd ranking sport types by country based on the coach benefit score. This is the direct result of the inference from the case study of the two examples of the GCE. Finally, the three countries and sport types identified by this approach are the female swimming team of China, the female football team of Germany, and the men's football team of Morocco.

Table 3: Coach benefit score by NOC

NOC	Sport	Sex	Number of Athletes	Gold Medals	Silver Medals	Bronze Medals	Coach Benefit Score
CHN	Swimming	F	18	0	4	21	46
GER	Football	F	20	0	0	20	40
MAR	Football	M	19	0	0	19	38
GBR	Athletics	F	34	1	7	14	35
GBR	Athletics	M	34	0	2	16	34
ARG	Hockey	F	16	0	0	16	32
IND	Hockey	M	16	0	0	16	32
ESP	Handball	M	16	0	0	15	30
DEN	Handball	F	17	0	0	15	30
AUS	Swimming	M	21	1	8	10	28

7 Conclusions

At the end, we restate everything.

The Olympics is considered as the prestigious athletic event that brings people with shared athletic spirit together to showcase physical prowess and to promote solidarity. Yet, the number of medals a country wins is taken as a factor to compare the success of countries, and lots of research has been done to predict the medal count of countries even before the olympics.

We studied, tried, tested, and implemented Random Forest Regression Model to predict the medal count of the countries. We calculated mean square error, mean absolute error, and R-squared error values to assess the accuracy of the predictions made by our model against the actual values in the test set of the given data and found them to be 6.87, 0.55 and 0.95 respectively for total medals count, and and 0.66, 0.16, and 0.96 respectively for the gold medal count. Further, the three countries and sport types identified by this approach are the female swimming team of China, the female football team of Germany, and the men's football team of Morocco.

References

[1] Peace and development through sport. URL https://www.olympics.com/ioc/peace-and-development.

[2] Olympic Medal Table. URL https://www.olympics.com/en/olympic-games/paris-2024/medals.

[3] How Brands are ‘Winning’ Olympic Gold. URL https://poole.ncsu.edu/thought-leadership/article/how-brands-are-winning-olympic-gold/.

[4] Who Wins the Olympic Games: Economic Resources and Medal Totals. URL https://faculty.tuck.dartmouth.edu/images/uploads/faculty/andrew-bernard/olymp60restat_finaljournalversion.pdf.

[5] Vagenas, G., Vlachokyriakou, E. (2011). Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the “population-GDP” model revisited. Sport Management Review, 15(2), 211–217. URL https://doi.org/10.1016/j.smr.2011.07.001.

[6] Shibiao Dong. The application research of arma forecasting model in prediction of medals and ranking for 2016 olympic games. *Journal of Chemical and Pharmaceutical Research*, 6(7):1383–1389, 2014. URL http://www.jocpr.com/articles/the-application-research-of-arma-forecasting-model-in-prediction-of-medals-and-ranki
pdf.

[7] Altman, N., Krzywinski, M. Ensemble methods: bagging and random forests. Nat Methods 14, 933–934 (2017). URL https://doi.org/10.1038/nmeth.4438.