Problem Chosen	2025	Team Control Number
C	MCM/ICM	2527081
	Summary Sheet	

Predicting Olympic Medal Counts

With a Random Forest Regressor Model

January 25-27/2025

Summary

Every four years, the Summer Olympics are one of the most exciting events around the world. Fans gather from around the world to watch the best athletes compete to represent their countries and earn prestigious medals. Over time, spectators have created the tradition of predicting which athletes and nations will have the most success at the Olympics and how many medals each nation will win.

We have developed a Random Forest Regression model that can accurately predict the gold medal, and total medal counts for competing teams in the 2028 Olympic Games in Los Angeles. To do this, we analyzed the historical medal data from all the past Summer Olympics, specifically accounting for each nation's past ranking in previous games. Additionally, we noted the overall impact a 'Great Coach' can have on a team.

Table Of Contents

Introduction

Our goal is to develop a model that considers a given country's previous performances at the Olympics to predict the amount of gold medals, and the total amount of medals they will achieve in the 2028 Olympic Games. Our model is a Random Forest Regression Model, inspired by previously attempted predictions of medal counts at the Olympic Games.[1] [2]

Assumptions

It is assumed that the 2028 Olympic Games will feature a program containing events like the last two Olympics, considering that the programs have rarely changed over the past few decades, apart from an occasional event selected specifically by the host nation. In addition, since the athlete data for 2028 is unavailable because qualification process hasn't begun, it is assumed that the same athletes will compete in the 2028 games that did in the 2024 games. It is also assumed that countries which competed in the 2020 and 2024 Olympics will be competing in the 2028 Olympics. Lastly the assumption is made that the teams will rank closely to where they are on track to place in the 2028 Olympics.

[1] Christoph Schlembach, Sascha L. Schmidt, Dominik Schreyer, Linus Wunderlich, *Forecasting the Olympic medal distribution – A socioeconomic machine learning model*, Technological Forecasting and Social Change, Volume 175, 2022, 121314, ISSN 0040-1625, https://doi.org/10.1016/j.techfore.2021.121314.

[2] Mengjie Jia, Yue Zhao, Furong Chang, Bofeng Zhang, Kenji Yoshigoe, *A Random Forest Regression Model Predicting the Winners of Summer Olympic Events*, Association for Computing Machinery, 2022, https://doi.org/10.1145/3404512.34045

Modeling Process

MySQL and Microsoft Excel were used to help manage and query the data to aid in the modeling process. Several Python libraries were utilized in the development of our model. The pandas library was used in order to manage the data, and the scikit-learn library was used to create the model and supplied metrics to assess the model's performance. Finally, the Matplotlib Library provided the tools for graphical representation of both model performance and model output.

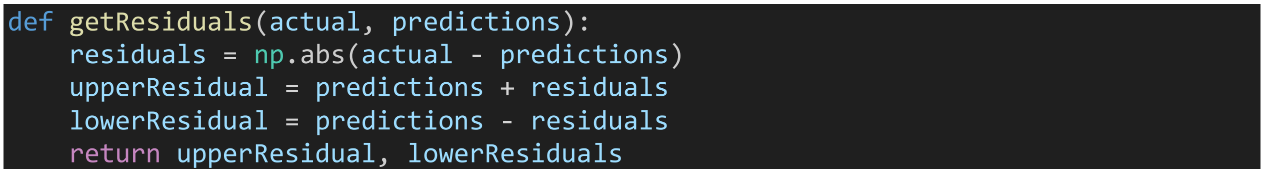
Initially, the model took in the entirety of the summerOly_medal_counts.csv Dataset, with the exception of the parameter it sought to predict. When the model was aiming to predict total medal counts, it used gold, silver, and bronze counts as context to make a prediction. This would pose a problem however, as the 2028 Olympics have yet to happen, so there would be no gold, silver, or bronze medal counts to pass through the model. Ultimately, the model would be made to focus on a given country's rank over the course of their participation in the Olympic games, and the medals they would achieve that year.

The model also considers a given country's' previous rank as context to make a prediction in the given country's medal counts. Once again, as the 2028 Olympics have yet to happen, the teams are going unranked. We resolved this with the use of Microsoft Excel's '=TREND' function, in order to get an estimate of where the team will rank, based off how they had ranked in the previous Olympic games.

Model Performance

This model was developed with the use of various tools within the scikit-learn libraries. The model was developed with the 'RandomForestRegressor' class and was evaluated with the and 'r2_score' function. The model's performance for predicting an Olympic team's earned gold medals had a R^2 score of 0.9511. The model's performance for predicting an Olympic team's overall medal earnings had an R^2 score of 0.9210. The highest R^2 score achievable is a score of one. Because the model's performance is close to one, the model makes decently accurate predictions.

The model's performance for predicting an Olympic team's earned gold medals is represented graphically in **Figure 1**, and the model's performance for predicting an Olympic team's earned gold medals is represented graphically in **Figure 2**. The blue points show the predicted values plotted against the actual values. The green and red points show the upper and lower residuals respectively, plotted against the actual values. The model's residual values were determined with the following function.[3]



```
def getResiduals(actual, predictions):
    residuals = np.abs(actual - predictions)
    upperResidual = predictions + residuals
    lowerResidual = predictions - residuals
    return upperResidual, lowerResiduals
```

[3] Zach Bobbit, *What Are Residuals in Statistics?* Stratology, 2020, https://www.statology.org/residuals/

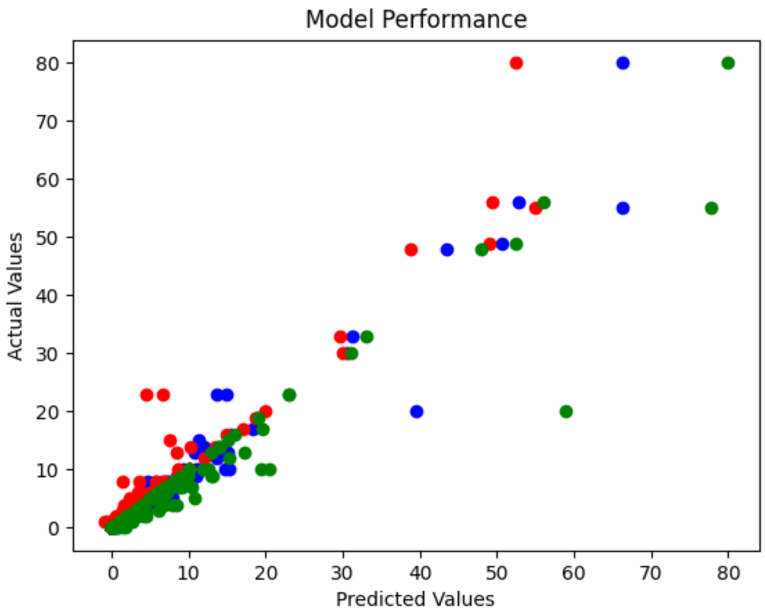


Figure 1: Gold Medal Model Performance

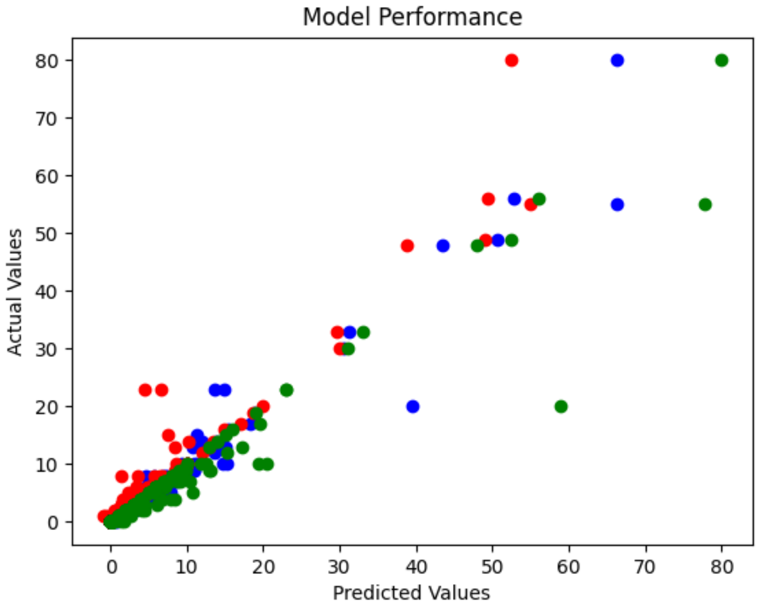


Figure 2: Total Medal Model Performance

Results

Our model's predictions for the number of gold medals an Olympic team is likely to earn in the 2028 shown in **Figure 3**. The Total number of medals our model predicts an Olympic team is likely to earn in 2028 is shown in **Figure 4**. These predictions are merely a single estimate, so a prediction interval was developed to get a better idea of the number of medals a given country is likely to earn. The prediction interval was developed with the use of the root mean squared of the model.[4] The root mean squared was found with the 'root_mean_squarred_error' function from the scikit-learn library. The upper bound is defined by the model's prediction plus the root means squared, and the lower bound is defined by the prediction minus the root mean squared; unless this would result in a negative quantity, in which case the lower bound is simply reported as zero. These calculations were made using Microsoft Excel. Predictions including prediction intervals for gold medals and total medals are found in **Figure 5** and **Figure 6**, respectively. For sake of space, the list has been shortened. The upper bound is denoted as 'UB' and the lower bound is denoted as 'LB.'

[4] Zach Bobbit, *What Are Residuals in Statistics?* Stratology, 2021, https://www.statology.org/how-to-interpret-rmse/

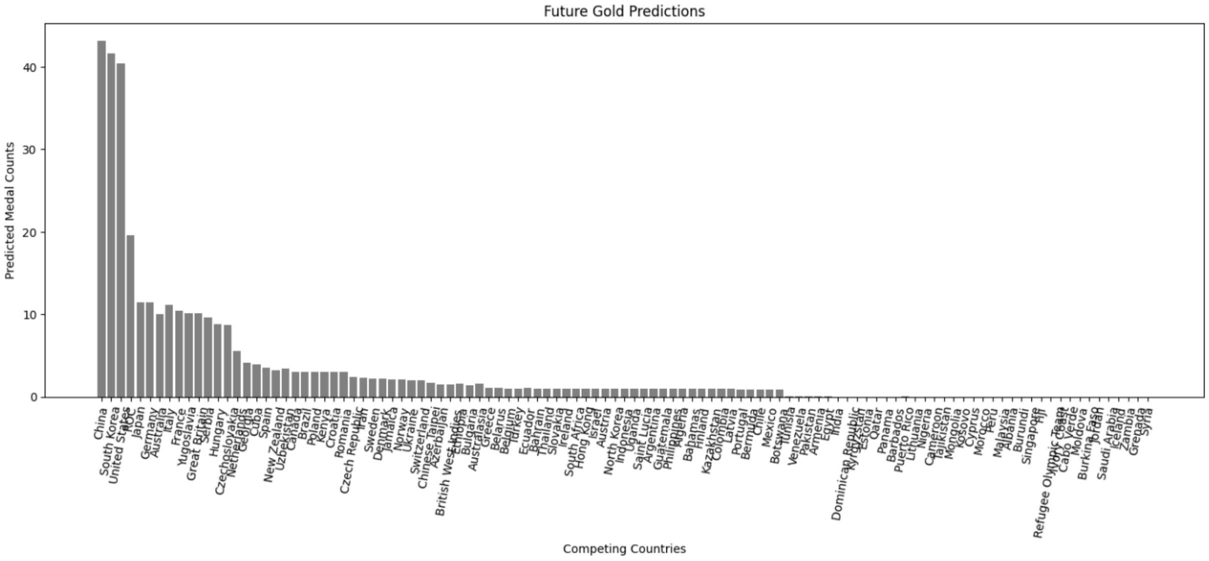


Figure 3: Predictions of Future Gold Medals

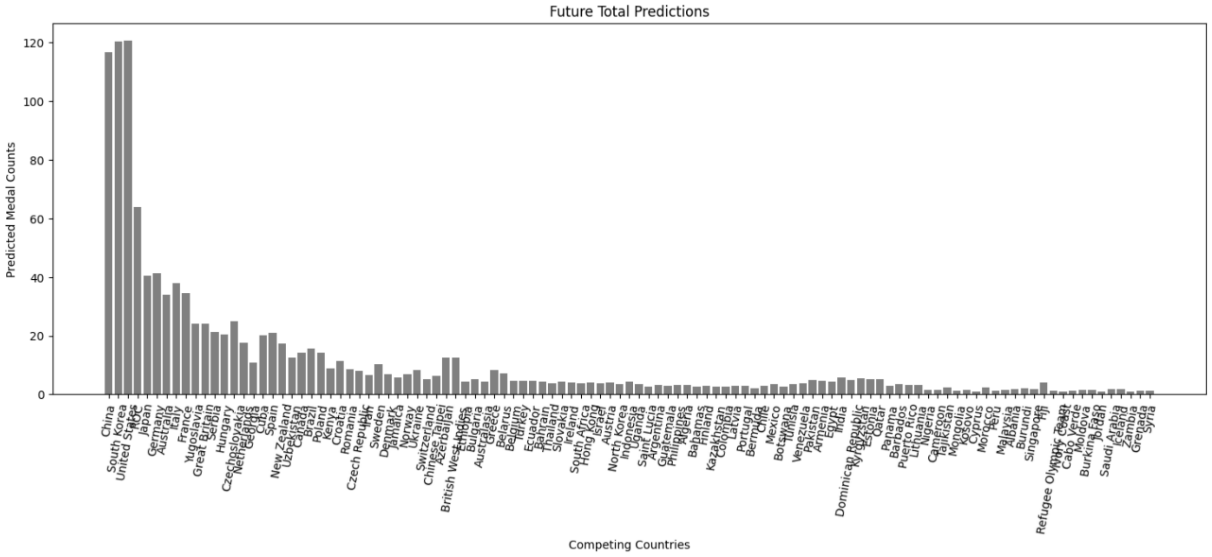


Figure 4: Predictions of future Total Medals

Figure 5: Gold Medals

Rank	NOC	Gold	UB	LB
1	China	43	45	41
1	South Korea	42	44	40
1	United States	40	42	38
5	ROC	20	22	18
7	Germany	11	13	9
7	Japan	11	13	9
8	Australia	10	12	8
9	Italy	11	13	9
10	France	10	12	8
11	Great Britain	10	12	8
11	Yugoslavia	10	12	8
12	Czechoslovakia	10	12	8
12	Hungary	9	11	7
12	Serbia	9	11	7
16	Netherlands	6	8	4
17	Cuba	4	6	2
17	Georgia	4	6	2
19	Spain	4	6	2
20	New Zealand	3	5	1
21	Uzbekistan	3	5	1
22	Brazil	3	5	1
22	Canada	3	5	1
22	Croatia	3	5	1
22	Kenya	3	5	1
22	Poland	3	5	1
24	Romania	3	5	1
29	Czech Republic	2	4	0
30	Iran	2	4	0
30	Sweden	2	4	0
31	Denmark	2	4	0
32	Jamaica	2	4	0
32	Norway	2	4	0
35	Ukraine	2	4	0

Figure 6: Total Medals

Rank	NOC	Total	UB	LB
1	China	117	123	110
1	South Korea	120	126	114
1	United States	120	127	114
5	ROC	64	70	58
7	Germany	40	47	34
7	Japan	41	47	35
8	Australia	34	40	28
9	Italy	38	44	32
10	France	35	41	28
11	Great Britain	24	30	18
11	Yugoslavia	24	30	18
12	Czechoslovakia	21	28	15
12	Hungary	20	27	14
12	Serbia	25	31	19
16	Netherlands	18	24	11
17	Cuba	11	17	5
17	Georgia	20	26	14
19	Spain	21	27	15
20	New Zealand	17	24	11
21	Uzbekistan	13	19	6
22	Brazil	14	20	8
22	Canada	16	22	9
22	Croatia	14	20	8
22	Kenya	9	15	3
22	Poland	11	18	5
24	Romania	9	15	2
29	Czech Republic	8	14	2
30	Iran	7	13	0
30	Sweden	10	16	4
31	Denmark	7	13	1
32	Jamaica	6	12	0
32	Norway	7	13	1
35	Ukraine	8	15	2

Great Coach Effect

Using MySQL, the potential effects of a potential “Great Coach” were investigated. Lang Ping was the one of the notable coaches we researched. Before Lang’s first stint as coach of the Chinese National Team from 1995-1998, they had failed to win a medal at the 1992 Olympic Games in Barcelona. However, under the guidance of Lang, the team won Silver in Atlanta the following Olympics and failed to make the podium in Sydney when Lang wasn’t the coach. During her second stint from 2013-2021, she also helped the team win gold in Rio, further showing her impact as a great coach.

Lang’s greatness can also be shown with other teams. From 1996-2004, the United States Women’s Volleyball Team didn’t win a single Olympic Medal, but after Lang took over as the head coach, they won Silver at the very next games. Even after Lang’s departure, team USA has remained a powerhouse, winning medals at every Olympic Games since.

To help quantify the impact of a coach, we first looked at the US Women’s team’s historical results prior to Lang joining the team. Team USA competed eight times from 1964-2004, winning one bronze and one silver in 40 years. While Lang was the coach, the US won silver, and following her tenure, they have won gold, silver twice, and bronze once. We will assign a score of 0-3 depending on the team’s placement, which results in an average placement score of 0.375 from 1964-2004. While Lang’s coaching tenure during the Beijing games in 2008 the team’s average is 2.000, and following Lang’s coaching stint, their average score is 2.000, showing a significant improvement by a factor of 5.333. For China, the results differ significantly. Since the Chinese team was already one of the strongest teams, especially since Lang was a player on a gold medal winning team in 1984, her impact as a coach isn’t going to be as strong compared to a team that was never as historically dominant. The Chinese team had an average score of 1.125 without Lang as their coach and had a score of 1.667 with her as coach, showing an improvement of 1.482. While not as drastic as her impact with the USA, her effect on a strong team like China shows how beneficial a great coach can be.

Conclusion

We have developed a model which is able to predict gold medal and total medal outcomes for teams competing in the 2028 Olympics. Through the use of the model's root mean squared error metric, we were also able to define prediction intervals for the gold and total medal predictions.

Although our model does fairly well in predicting medal outcomes, it only takes into account the competing team, and their prior standings. Our predictive model does not take into consideration the athletes competing, the coaches, or the events that will be taking place. An area of improvement for this model would be implementing the consideration of these parameters. Due to such limitations in our model, we were not able to deploy it to investigate the effects of a 'great coach.' Improvements in the model would also potentially allow for such an investigation.

Bibliography

Christoph Schlembach, et al.

Mengjie Jia, et al.

Nicolai Meinshausen, ed. Greg Ridgeway, *Quantile Regression Forests,* Journal of Machine Learning Research 7, 2006, https://jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf

pandas Python Library and Documentation, https://pandas.pydata.org/pandas-docs/stable/index.html

scikit-learn Python Library and Documentation, https://scikit-learn.org/stable/index.html

General Statistic Materials, https://www.statology.org/