A Random Forest Regression Model Predicting the Winners of Summer Olympic Events

Mengjie Jia
School of Computer Engineering and Science, Shanghai University
Shanghai, China
+8618277102479
mengjiejia99@gmail.com

Yue Zhao *
School of Computer Engineering and Science, Shanghai University
Shanghai, China
+8618616612154
yxzhao@shu.edu.cn

Furong Chang
School of Computer Engineering and Science, Shanghai University
Shanghai, China
+8618817621257
cfrkashger@shu.edu.cn

Bofeng Zhang
School of Computer Engineering and Science, Shanghai University
Shanghai, China
+8613918229959
bfzhang@shu.edu.cn

Kenji Yoshigoe
Faculty of Information Networking for Innovation and Design (INIAD),
Toyo University, Tokyo, Japan
+81359242650
yoshigoe@iniad.org

ABSTRACT
From the past Olympic medal lists, we can find that the number of medals of China has been increasing steadily in recent years while we also observe that some countries always occupy the top positions of the Olympic medal list, such as the United States, Britain and Germany. In this work we take the data of the medal lists from the 18th to 31st Summer Olympic Games as a sample and selects GDP, the population, the size of national team and the home advantage as the characteristic parameters to build a random forest regression model to predict the number of medals. The FP-growth algorithm is used to analyze the association rules of the data. And the winners of some events in the 2020 Tokyo Olympic Games are predicted.

CCS Concepts
• Theory of computation→ Theory and algorithms for application domains→ Machine learning theory→ Models of learning • Information systems→ Information systems applications→ Data mining → Association rules

Keywords
Random Forest Regression model; Association rules; FP-growth algorithm

1. INTRODUCTION
The Olympic Games is an international sports event held every four years [1]. It contains various sports and is the focus of attention of all countries in the world, because it is regarded as a symbol of national identity. However, not all countries have the same ability to participate in the Olympic Games. What's more, even if they participate in the Olympic Games, they will not have the same ability to obtain medals [1]. Some media have reported that the Olympic medal list is equal to the world GDP ranking. Does the GDP of a country have a significant impact on the Olympic performance? In addition to GDP, what factors will affect the final results of the Olympic Games? Since the strength and status of athletes are unpredictable and the emergence of talented athletes or the suspension of athletes is an important reason for affecting a country's Olympic performance, it is hard to obtain a very high accuracy by optimizing the model continuously and the analysis and fitting is based on the macro factors mentioned above.

The main contributions of this work are to build a regression model of the number of Summer Olympic medals on the four influencing factors: GDP, the population, the national team size and home advantage. The training will be based on the data of the 18th to 29th Summer Olympics medals. The Random forest model outperforms and is tested using the medals of the 30th to 31st Summer Olympic Games. Besides, we will use the algorithm of association rules to mine out the superiority events of some countries and then provide the predictions for the winners of some events of the 2020 Summer Olympics in Tokyo.

Th paper is organized as follows: Section 2 discusses the related work. Section 3 introduces three kinds of regression models. FP-growth algorithm is explained in Section 4. Section 5 describes the datasets, data preprocessing and presents the data analysis as well as provides the experimental results. Section 6 concludes the paper followed by Section 7 that lists the references.

2. RELATED WORKS
A number of research have focused on the factors affecting the Olympic medal list. A survey by Daniel K. N. Johnson and Ayfer Ali investigated what factors could encourage countries to send Olympic athletes economically and politically to find out the crucial elements of their success [1]. And another paper by Andrew B. Bernard and Meghan R. Busse took into consideration population and economic resources as the determinative roles of the total number of medals from 1960 to 1996 [2]. NJ BALMER et al. studied the impact of home advantage in team games, and their purpose was to evaluate the importance of home advantage for the five event groups selected from the 1896 to 1996 Summer Olympics [3]. Tian Lei, Liu Weimin, and Liu Dan conducted

*Corresponding author

BDE 2020, May 29–31, 2020, Shanghai, China

ACM ISBN 978-1-4503-7722-5/20/05…$15.00
DOI: https://doi.org/10.1145/3404512.3404513

research on the performance of Olympic host countries through literature review, mathematical and logical reasoning [4]. The results showed that the host country won 3.6 times its own away advantage medal, which was 1.85 times the previous and next Olympics [4]. Zhang Yuhua used the results of the Chinese Olympic Games from 1984 to 2012 for linear regression to predict and analyze the results of the 31st Olympic Games [5]. Through regression analysis, Eike Emrich et al. analyzed the impact of population size and GDP per capita on sports success and found that the strongest predictor associated with the summer Olympics was population size [6]. Similarly, David Forrest, Ismael Sanz, and J.D.Tena started with an established statistical model based on a regression analysis of the total number of medals in the early games, where past performance and GDP were the main covariates [7]. However, they ended up making their own predictions based on a model that included additional regressions, including measures of public spending [7]. WANG Guofan et al. put forward a modified model combining the economic theory and the difference of competitive sports strength by introducing dummy variables, and also predicted and analyzed the number of medals in 2008 Beijing Olympic Games [8]. Their results showed that the model could reflect the general trend of the changes of Olympic medals in different countries and overcome the influence of the difference of competitive sports strength on prediction accuracy [8]. As to some special events like swimming, there are also many research on them. Timothy Heazlewood used mathematical models to predict elite performance in Olympic swimming and athletics. He found that the non-linear model showed data closer to actual data and were used to predict performance changes over the next 10, 100, and 1000 years [9]. Arkadiusz Stanula et al. analyzed the past performance of freestyle swimmers, and their predictions revealed some interesting trends that existed in the results of individual tests for men and women [10]. Engin Esme and Mustafa Servet Kiran proposed a method of football match result prediction based on classification function of learning algorithm and they derived 17 feature vectors from 8 types of data to train the model which finally show the success of the prediction results [11]. Atsushi Matsumoto et al. predicted the radiation protection function and toxicity of the radioprotectants for p53 using random forest and SVM and found that random forest is better than SVM in predicting the toxicity and they also indicate that the accuracy of result is more stable using collective-learning algorithm such as random forest than SVM [12].

3. REGRESSION MODELS

Based on the related works above and the analysis of the correlation between different variables and medal counts, we find that they are not simply linear related. Three regression models adaptable to different regressions are used in the study to fit the medal list, which are described briefly next.

3.1 GBR

GBR is a machine learning technique used to solve regression problems. It generates a prediction model by assembling weak prediction models, usually decision trees and gradually generalizes the model to optimize their functions by allowing optimization of arbitrary differential losses. Perhaps the accuracy of each learning algorithm is not high, but they can be integrated together to achieve a good goal.

3.2 Polynomial Regression

Polynomial regression is to study the regression analysis method of polynomials between a dependent variable and one or more independent variables. It is used to describe nonlinear phenomena such as the growth rate of tissues [13], the distribution of carbon isotopes in lake sediments [14], and the course of disease epidemics [15]. The maximum index is generally set to 2 or 3 in the model for if the index is too large, it will lead to over-fitting.

3.3 Random Forest

Random forest is an integrated learning method which constructs a large number of decision trees during training and outputs the model of class (classification) or average prediction (regression) of individual trees [16] [17]. The random decision forest corrects the inadequacies of the decision tree over adaptation to its training set [18].

4. FP-GROWTH ALGORITHM

FP-growth (Frequent pattern growth) algorithm is an improvement to the Apriori method to mine association rules which maps the transactions in the data set to a FP-Tree, and then find the frequent item-sets based on this tree.

The two important filter criteria are support and confidence. Support is defined as support = P(AB), which refers to the probability that event A and event B occur simultaneously. Confidence is defined as confidence = P(B|A) = P(AB)/P(A), which refers to the probability that event B will occur on the premise of event A.

5. EXPERIMENTS

This part is to do some data preprocessing. Dataset ***U*** is statistically analyzed and is visualized and the features of data in it are selected and verified. Then we evaluated the regression models trained by the dataset and used FP-Growth algorithm to find associated rules.

5.1 Preprocessing

The original data and the two imported external datasets are described at first. The data preprocessing is mainly divided into the missing value processing, the introduction of external data sets, logical deduplication and data integration.

5.1.1 Data Sources

The original dataset named ‘athlete_events’ is the information and awards of all athletes participating in the Olympic Games from 1896 to 2016. The external datasets are GDP of different countries in the world from 1964 to 2016 and the total population of the world from 1964 to 2016. The GDP dataset is named ‘world_gdp’ and the population dataset is called ‘world_pop’. The data source of both of them is World Development Indicators. The above three datasets are all collected and downloaded from Kaggle and saved as CSV files.

5.1.2 Original Dataset

The original dataset contains information on all participating athletes from the 1896 to 2016 Winter and Summer Olympic Games. There are 271116 pieces of data and each contains 11 pieces of information, which are the athlete’s name, sex, age, height, weight, team, nationality, the Olympic game he participated in, the city the game was held, the event he competed in and the medal he won.

5.1.3 External Datasets

Since the original dataset has few meaningful features for predicting the number of a country’s medals, it is decided to introduce the two external datasets of the world's total population and the world's GDP based on the existing academic research results.

5.1.4 Missing Value Processing

The research purpose of this project is to analyze the number of Olympic medals in various countries over the years. Therefore, the athletes' personal information, such as height, weight and so on, are irrelevant information which are removed from the dataset. Since not all contestants will win, the missing value of the medal is justified, which can be filled with "DNW" (Do Not Win).

5.1.5 Introduce External Datasets

Whether a country's comprehensive strength is strong or not determines how much it can invest in sports. If the probability distribution of various types of talents in different countries is the same, then under the same conditions, countries with large populations will have an absolute number of outstanding athletes, which will greatly increase the possibility of winning the Olympic Games [2]. Thus, a huge population and high per capita GDP are required to win a large number of medals [2].

Since the two datasets of national GDP and total population both started from 1964 while the original dataset is started from 1896, the original dataset is screened and only the summer Olympics data from 1964 to 2016 is selected as the new dataset, which is denoted as ***E***. And then the national GDP and total population datasets are introduced into ***E***.

5.1.6 Logical Deduplication

The new dataset ***E*** still collects data on an individual basis, so group events will be awarded multiple medals resulting inaccuracy in each country's medal statistics. Thus, when counting the number of national medals, it is necessary to find the group item and make the item's medal count to 1.

Firstly, a column named "Medal_Won" is added into ***E***, and the value is 1 for the award, otherwise 0. Secondly, the total number of athletes sent by each country is counted as a new characteristic parameter named "Total_Athletes" and added as a new feature into ***E***. Thirdly, two new feature parameters, "Country_Host" and "Home_adv" are generated. The first one gets its value from which country held the game and the second one is used as the host country flag. If a country is the host country of the corresponding year, its value is marked as 1, otherwise 0.

The threshold is defined as 1 to obtain a list of group events because if the number of gold medals awarded by the same event is greater than 1, the event is mostly like to be a group event. However, not all events whose gold medal counts greater than 1 are group events because there will be situations where results are tied. Through screening, four items are found to be individual athletic events in the total group event list and we finally get 101 group events.

After the medals of each Olympic Games are revised according to the country and the events in the dataset ***E***, two new datasets respectively denoted as ***S*** and ***U*** are generated and both of them contains 7740 pieces of data. They differ in the kinds of features of their data. ***S*** only has two new features named "Medal_Won" and "Medal_Won_Corrected" respectively as well as deleting some features while ***U*** gets other two new features, "Total_Athletes", "Country_Host" and "Home_adv" in addition. It is noted that the feature "Medal_Won" is replaced by "Medal_Won_Corrected" and its data is corrected. The association rule, Country-Event-Medal, will be mined based on ***S***.

5.2 Data Analysis

The dataset ***U*** is statistically analyzed and is visualized and the features of data in it are selected and verified. In order to get a good result of experiment, the data is changed and standardized.

5.2.1 Correlation Coefficient

Figure 1 shows the correlation coefficient matrix between variables. It can be seen that the correlation coefficient between "Medal_Won_Corrected" and "Total_Athletes" is up to 0.86, which is a highly linear positive correlation. The correlation coefficient between "Medal_Won_Corrected" and "GDP" is 0.65, which is moderately related, but "Medal_Won_Corrected" is weakly related to "Home_adv" and "population".

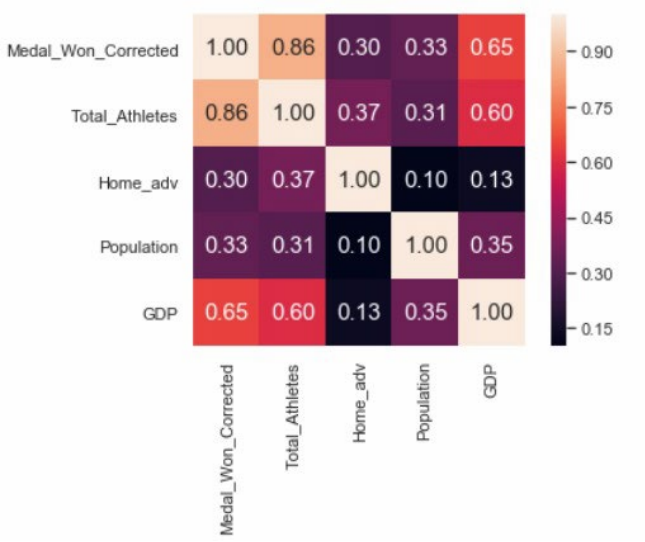

Figure 1. Correlation coefficient matrix between variables

5.2.2 Feature Verification

Since the correlation coefficient reflects the degree of linear correlation between variables, a small correlation coefficient only means a poor linear correlation between the two variables and does not indicate that the two are irrelevant. Through the calculation of correlation coefficient, it is found that the linear relationship between "Medal_Won_Corrected" and "GDP" and "population" is not obvious.

High productivity (measured by GDP) shows the ability to pay for athletes to participate in the Olympics, and may also be linked to better training and equipment [1]. If a country has a good economic foundation, it can provide athletes with advanced training conditions, generous rewards, and high quality of life, so that athletes can participate in competitions and training without any worries as well as have more motivation to win more honors, which will lay a solid foundation for achieving good results.

As can be seen from Figure 2, where columns represent the counts of medals and the line represents GDP, in the "Top 30 countries" of the 2016 Brazil Olympic medal list, the GDP can be roughly divided into three gradients, which are distinguished by the red vertical line. It shows that the number of national medals and GDP are simultaneously falling indicating a country's Olympic performance is positively related to GDP.

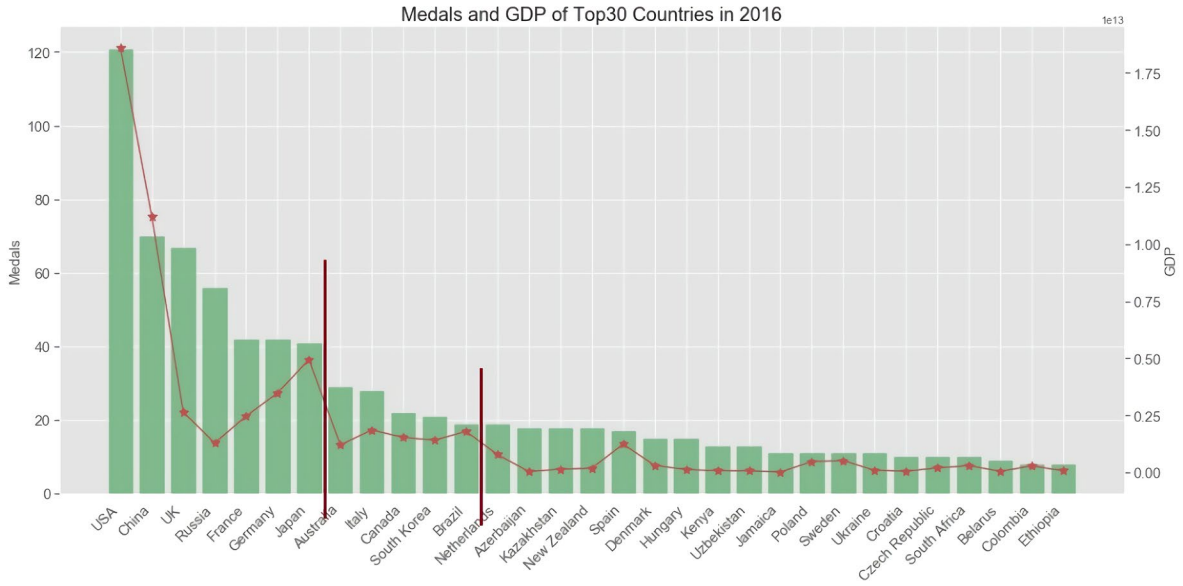

Figure 2. Medals and GDP of Top30 Countries in 2016

Countries with larger populations win more medals (and more gold medals) at the Summer Olympics, averaging one medal per 10 million inhabitants (one gold medal per 30 million inhabitants) [1]. A large population means that the fixed costs of training, such as infrastructure and facilities, are more efficiently shared and there are more potential athletes to choose from [1]. However, developed countries tend to have strong economic strength and a small population, making the impact of economic strength on the Olympic performance more significant and weakening the impact of the population on it.

Therefore, according to the principle of control variables, in the verification of the correlation between the population and the number of medals, the comparison is made with 25 countries that participated in the Olympic Games in 2016 whose GDP is of the order of magnitude of e8. The 25 countries can be divided into two levels according to the number of medals won, which can be seen in Figure 3, where columns represent the counts of medals and the line represents population. It can be found that the population of the first-tier countries is generally larger than that of the second-tier countries, proving that the population of a country enhances its Olympic performance under the control of the GDP factor.

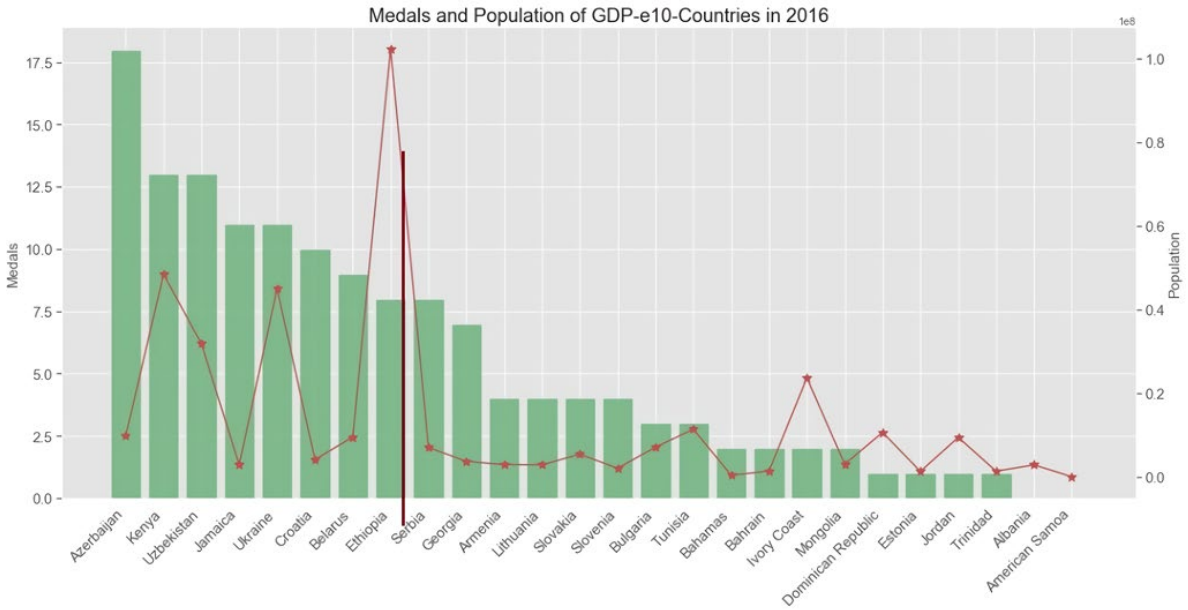

Figure 3. Medal and Population of GDP e10 countries in 2016

Clearly, the host country and its neighbors also have a competitive advantage in terms of lower transport costs, climate and training [1]. The results of the host countries of each Olympic Games from 1964 to 2016 are output in Figure 4. In order to form a comparison, the three columns on the far-right of the table are the medals won by the host country in the previous, hosting and next Summer Olympic Games. It can be seen from Figure 4 that except for the countries that did not participate in the 1980 or 1984 Olympic Games, the total medals in host countries in their own Olympic Games is almost the highest among the three Olympic Games, which indicates that the hosting country do have some advantages promoting their Olympic performance.

	Year	Country_Host	Team	Medal_Won_Prev_Year	Medal_Won_Host_Year	Medal_Won_Next_Year
0	1964	Japan	Japan	NaN	29.0	25.0
1	1968	Mexico	Mexico	1.0	9.0	1.0
2	1972	Germany	Germany	51.0	106.0	129.0
3	1976	Canada	Canada	5.0	11.0	NaN
4	1980	Russia	Russia	125.0	195.0	NaN
5	1984	USA	USA	NaN	173.0	94.0
6	1988	South Korea	South Korea	19.0	33.0	28.0
7	1992	Spain	Spain	4.0	22.0	17.0
8	1996	USA	USA	108.0	101.0	91.0
9	2000	Australia	Australia	41.0	58.0	50.0
10	2004	Greece	Greece	13.0	16.0	4.0
11	2008	China	China	64.0	100.0	89.0
12	2012	UK	UK	48.0	65.0	67.0
13	2016	Brazil	Brazil	17.0	19.0	NaN

Figure 4. Performance of the hosting countries

5.2.3 Data Transformation

The data of GDP and population both have a positive skewness and the distribution of GDP is taken as an example. Figure 5 depicts the distribution of national GDP where the vertical and horizontal axis stand for country counts and GDP respectively. After taking logarithm of the raw data of the country's GDP and population, the distribution of data is uniform for subsequent model building, which can be seen in Figure 6 (still taking GDP as an example). The logarithmic operation does not change the nature and correlation of the data, but compresses the scale of the variables, making the data more stable, and also weakens the collinearity and heteroscedasticity of the model the data will be used in.

The original value of the indicators needs to be standardized to ensure the reliability of the results since the levels of the indicators differ greatly. The method used in this paper is z-score standardization which standardizes the data based on the mean and standard deviation of the raw data. After z-score normalization of the four features' value in ***U***, the new dataset denoted as ***V***.Experimental Results

After fitting the three regression models on the train data and evaluating their performance with the criteria which is described next, we select Random forest as the model with the best regression fitting and use it to predict each country's medals with the test data.

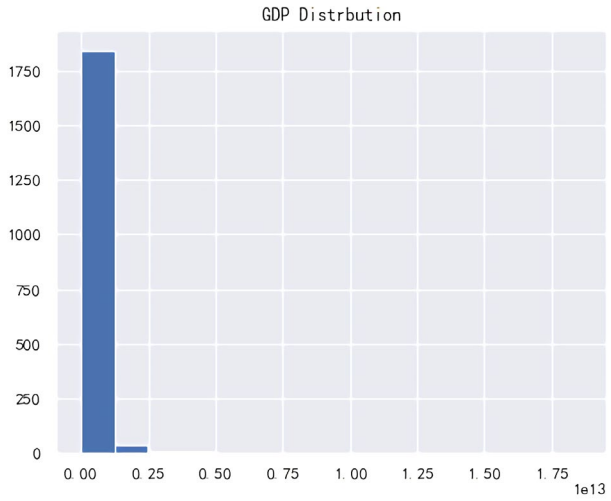

Figure 5. Distribution of GDP

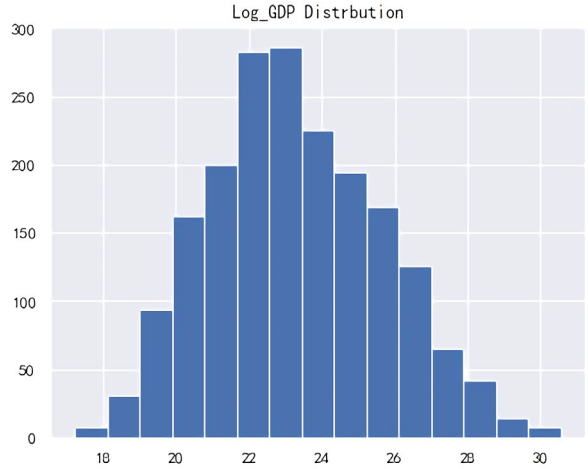

Figure 6. Distribution of Log_GDP

5.2.4 Training Set and Test Set

Considering the Olympic dataset ***V*** is combined in each session, the data of the 18th to 29th Olympic Games is used as the training set, and the data of the 30th to 31st Olympic Games is the test set. After division, the training set has 1756 pieces of data and the test set has 376 pieces of data.

5.2.5 Evaluation Criteria

In this project, the 10-fold cross-validation method is used as one of the evaluation criteria. Each test yields the correct rate and the average of the correctness of the results of 10 times is used as an estimate of the accuracy of the algorithm. We also take RMSE to evaluate the error and R^2 to assess the model quality. The smaller the value of RMSE, the better the regression prediction of the model and the closer the value of R^2 is to 1, the better the fit.

5.2.6 Model Training

We use the GBR model, the polynomial regression model (coefficient set to 2) and the Random forest model in the Scikit-Learn library to train the training set and then evaluate the three models with the three metrics. The results are shown in Table 1 below which indicates that the Random forest is the best among the three models. Thus, the Random forest model is selected as the best to predict the test set.

Table 1. Comparison of evaluation metrics of the three models

Model	RMSE	Average score of 10-Fold-Cross-Validation	R^2
GBR	5.9311	0.8941	0.8479
Polynomial regression	6.7952	0.8442	0.8061
Random Forest	5.6047	0.8976	0.8538

5.2.7 Prediction of the Test Set

The medal lists in 2012 and 2016 Summer Olympic Games in the test set are predicted. We present the results of 2016 Summer Olympic Games in Figure 7-12. The abscissa is country codes of the participating countries which is arranged in ascending alphabetical order and the ordinate is the number of medals. The red polyline indicates the correct value while the green one indicates the predicted value.

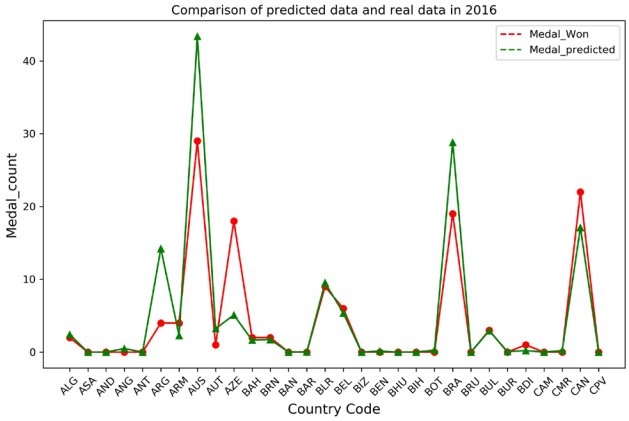

Figure 7. Results of medal counts for the 2016 Olympic Games (1)

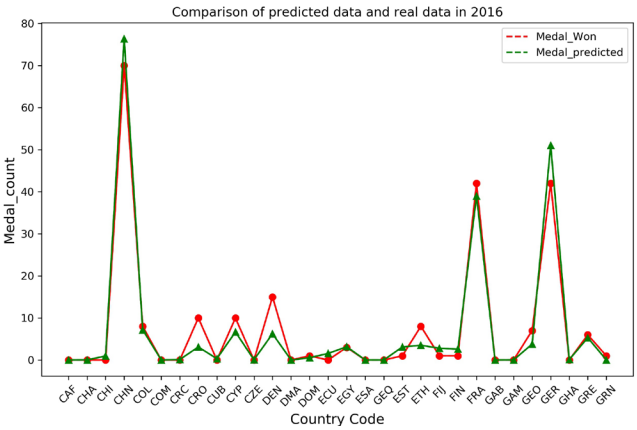

Figure 8. Results of medal counts for the 2016 Olympic Games (2)

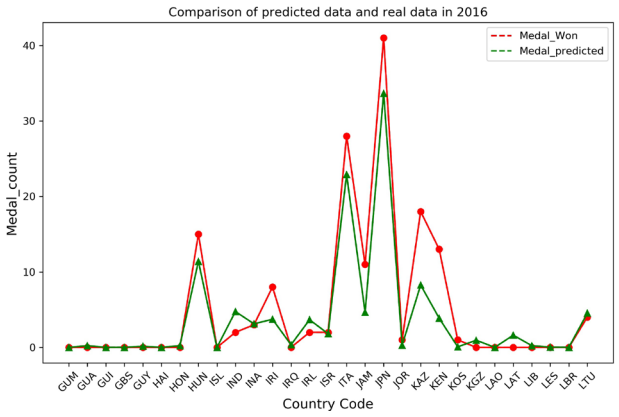

Figure 9. Results of medal counts for the 2016 Olympic Games (3)

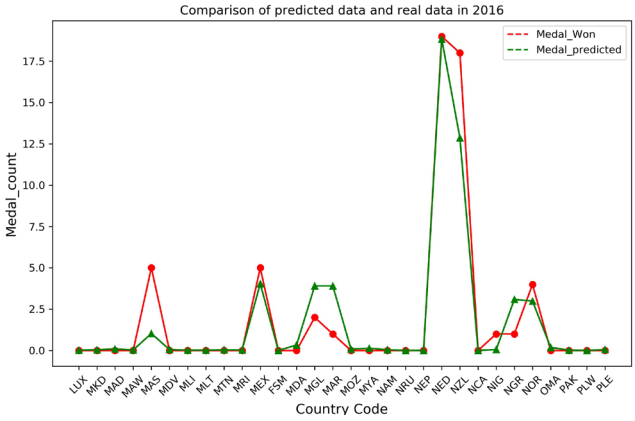

Figure 10. Results of medal counts for the 2016 Olympic Games (4)

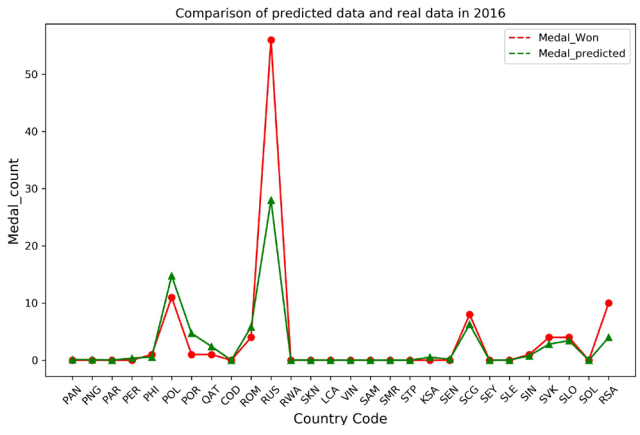

Figure 11. Results of medal counts for the 2016 Olympic Games (5)

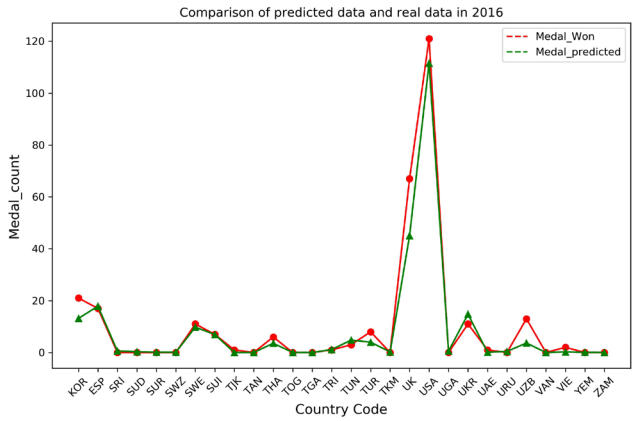

Figure 12. Results of medal counts for the 2016 Olympic Games (6)

As can be seen from Figure 7 to Figure12, the regression model is more accurate in predicting the number of medals below 10, and there is a deviation of 5-25 for those above 10. Except for developing countries such as China and Brazil, the countries whose predicted value differs greatly from the actual value are almost developed countries, which proves once again that the influence of a country's economic strength on a country's Olympic performance is quite significant. Under the influence of economic globalization, the flow of sports elites in the world today is becoming more and more frequent. Although developed countries may not have a large population, they can attract sports talents with their rich salary and superior training conditions. Consequently, the impact of population factor on the Olympic performance of this part of the country has decreased.

5.2.8 Association Rules

Based on the dataset ***S,*** the set generated, denoted as ***I***, contains a total of 7740 item-sets. ***I*** includes the data from the 18th to the 31st, a total of 14 Olympic Games, so if a country participated in the same event more than 10 times (including 10 times), the country can be considered to be connected with this event frequently. Therefore, the support threshold is 10/14. Assume that for each event, the probability of each country to win is the same, then if the frequency of a country winning for one event more than 0.7, we consider that the country has a big advantage in this event. So, the confidence threshold is set to be 0.7.

Statistically, the frequency of {(event, country) => (medal)} is not exactly the same as the frequency of {(event, medal) => (country)}. So we use the frequency based on the historical data to estimate the probability a country will win an event and the probability is represented by the frequency. According to the definition of conditional probability, the calculation formulas are as follows:

$$P((event, country) \Rightarrow (medal)) = \frac{P(event,country,medal)}{P(event,country)} = \frac{F(event,country,\ medal)}{F(event,country)}$$

$$P((event, medal) \Rightarrow (country)) = \frac{P(event,country,medal)}{P(event,medal)} = \frac{F(event,country,medal)}{F(event,medal)}$$

P ((event, country) => (medal)), denoted as P_1, means the frequency a country won this kind of medal in the event it participated in.

P ((event, medal) => (country)), denoted as P_2, means the frequency a country won this kind of medal in the record of the event.

Both of the kinds of frequency are important and needed to be considered when predicting the probability of an event's winner. According to the definition of *F-score*, a measure of a test's accuracy, which considers both the precision *p* and the recall *r* of the test, we define a new parameter P_3 to calculate the comprehensive probability. The formula is as follows:

$$P_3 = \frac{P_1 * P_2 * 2}{P_1 + P_2}$$

After calculating the value of P_3 for all the association rules, 62 strong association rules are obtained. The rules whose P_3 value is greater than 0.7 are presented in Table 2, which declares that South Korea and the USA respectively has a very high probability in "Archery Women's Team" and "Swimming Men's 4×100 meters Medley Relay", almost in a monopoly position. The P_3 value of all association rules is the probability prediction value of the events' winner of the 2020 Tokyo Olympic Games, which is based on the analysis and mining of historical data.

Table 2. Rules whose P_3 value is greater than 0.7

Event	Team	Medal	P_3
Archery Women's Team	South Korea	Gold	1
Swimming Men's 4 x 100 metres Medley Relay	USA	Gold	0.9637
Weightlifting Men's Lightweight	China	Gold	0.86
Tennis Women's Doubles	USA	Gold	0.8571
Equestrianism Mixed Dressage, Team	USA	Bronze	0.8235
Equestrianism Mixed Dressage, Team	Germany	Gold	0.8189
Athletics Men's 4 x 400 metres Relay	USA	Gold	0.7976
Athletics Women's 4 x 100 metres Relay	USA	Gold	0.78
Table Tennis Women's Singles	China	Gold	0.7654
Basketball Women's Basketball	USA	Gold	0.7634
Swimming Men's 4 x 100 metres Freestyle Relay	USA	Gold	0.75
Rowing Women's Quadruple Sculls	Germany	Gold	0.75
Wrestling Men's Heavyweight, Freestyle	Russia	Gold	0.7481
Basketball Men's Basketball	USA	Gold	0.7388
Table Tennis Women's Singles	China	Silver	0.7132

6. CONCLUSION

According to the analysis of the Olympic medal list and the correlations between the factors and it, we found that GDP, population, the national team size and home advantage all have a positive impact on a country's Olympic performance. After the comparison of three regression models, Random forest regression model is selected to predict the medal lists of 30th and 31st. We define a new parameter P_3 of association rules mined by FP-growth to predict the probability of an event's winner in the 2020 Tokyo Olympic Games.

7. REFERENCES

[1] Daniel K. N. Johnson, Ayfer Ali. "A Tale of Two Seasons: Participation and Medal Counts at the Summer and Winter Olympic Games". Wellesley College Department of Economics Working Paper 2002-02.

[2] Andrew B. Bernard, Meghan R. Busse, Review of Economics and Statistics, p.413-417, 2004.

[3] NJ BALMER, AM NEVILL, AM WILLIAMS. "Modelling home advantage in the Summer Olympic Games". Journal of Sports Sciences Volume 21, 2003 - Issue 6, Pages 469-478 | Published online: 07 Feb 2011. DOI = https://doi.org/10.1080/0264041031000101890

[4] TIAN Lei, LIU Wei-min, LIU Dan. Study on Home Advantage in the Summer Olympic Games. China Sport Science and Technology 2008-01

[5] ZHANG Yuhua. Prediction of Chinese Delegation Medal Number in the Thirty-first session of Olympic Games by Linear Regression Dynamic Model. Journal of Henan Normal University(Natural Science Edition 2013-02

[6] Eike Emrich, Markus Klein, Werner Pitsch, Christian Pierdzioch, 2012. "On the determinants of sporting success – A note on the Olympic Games". Economics Bulletin, AccessEcon, vol. 32(3), pages 1890-1901.

[7] David Forrest, Ismael Sanz, J.D.Tena. "Forecasting national team medal totals at the Summer Olympic Games". International Journal of Forecasting, Volume 26, Issue 3, July–September 2010, Pages 576-588

[8] WANG Guofan, XUE Erjian, TANG Xuefeng. Prediction of the Number of Medals in International General Games: With Beijing Olympic Games as an Example. Journal of Tianjin University of Sport 2010-01

[9] Timothy Heazlewood. Prediction Versus Reality: The Use of Mathematical Models to Predict Elite Performance in Swimming and Athletics at the Olympic Games. *The 8th Australasian Conference on Mathematics and Computers in Sport, 3-5 July 2006, Queensland, Australia.* J Sports Sci Med. 2006 Dec; 5(4): 480–487.

[10] Arkadiusz Stanula1, Adam Maszczyk, Robert Roczniok, Przemysław Pietraszewski, Andrzej Ostrowski, Adam Zając, Marek Strzała. The Development and Prediction of Athletic Performance in Freestyle Swimming. Journal of Human Kinetics volume 32/2012, 97-107

[11] Engin Esme and Mustafa Servet Kiran, "Prediction of Football Match Outcomes Based On Bookmaker Odds by Using k-Nearest Neighbor Algorithm," International Journal of Machine Learning and Computing vol. 8, no. 1, pp. 26-32, 2018

[12] Atsushi Matsumoto, Shin Aoki, and Hayato Ohwada, "Comparison of Random Forest and SVM for Raw Data in Drug Discovery: Prediction of Radiation Protection and Toxicity Case Study," International Journal of Machine Learning and Computing vol.6, no. 2, pp. 145-148, 2016

[13] Shaw, P; et al. (2006). "Intellectual ability and cortical development in children and adolescents". Nature. 440(7084): 676–679

[14] Barker, PA; Street-Perrott, FA; Leng, MJ; Greenwood, PB; Swain, DL; Perrott, RA; Telford, RJ; Ficken, KJ (2001). "A 14,000-Year Oxygen Isotope Record from Diatom Silica in Two Alpine Lakes on Mt. Kenya". Science. 292 (5525):2307–2310.

[15] Greenland, Sander (1995). "Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis". Epidemiology. 6 (4): 356–365.

[16] Ho, Tin Kam (1995), Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995.

[17] Ho TK (1998), The Random Subspace Method for Constructing Decision Forests, IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.

[18] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5