Theory and Practice of Science and Technology
2025, VOL. 6, NO. 3, 182-187
DOI: 10.47297/taposatWSP2633-456934.20250603

Medal prediction and uncertainty analysis of the 2028 Los Angeles Olympics based on the Lasso regression model

Ruoyu Zhang[1], Xiaoding Li[2], Zhiyuan Du[3], Yuxiang Xu[1], Yuan Hao[4]

[1]Guilin University of Technology, Guilin, China, 541000

[2]Qingdao University, Qingdao, China, 266071

[3]The Northeast Forestry University, Harbin, China, 150000

[4]Changchun University of Architecture and Civil Engineering, Changchun, China, 130607

ABSTRACT
As one of the largest sporting events in the world, the medal performance of each country in the Olympic Games not only reflects the competitive level, but also reflects the comprehensive strength of the country's sports development. With the increasingly fierce competition in the Olympic Games, accurate medal prediction provides an important reference for the strategic planning, resource allocation and preparation of each country. This paper proposes a medal prediction method based on the Lasso regression model. By analyzing the US Olympic medal data from 1986 to 2024, a regression model for predicting the medal table of the 2028 Los Angeles Summer Olympics was established. In order to improve the accuracy of the model, feature standardization, cross-validation and uncertainty estimation were performed. The results show that the United States and China are still in the leading position in the number of gold medals and total medals, with the predicted number of gold medals being 42 and 40 respectively, and the estimated number of total medals being 129 and 88 respectively. This paper also evaluates the uncertainty of the prediction results through confidence interval analysis, providing strong support for policymakers on the trend of future Olympic competitions. Finally, this paper discusses future research directions, including how to deal with emerging variables and further improve the accuracy and stability of the model.

KEYWORDS
Olympic Medal Prediction; Lasso Regression Model; Uncertainty Estimation; Feature Standardization; Regression Analysis

1 Introduction

As one of the largest sporting events in the world, the Olympic Games not only showcases the competitive level of athletes from various countries, but also reflects the overall trend of sports development in various countries. As the competition in Olympic events becomes increasingly fierce, accurate medal prediction has become a key basis for governments, sports institutions and related organizations to formulate strategic plans, allocate resources and prepare for the games. Medal prediction through data-driven methods can not only provide strong decision-making support for sports managers, but also reveal the potential and challenges of various countries in future Olympic events. Therefore, the accuracy and reliability of medal prediction directly affect policy making and the rational allocation of resources.

With the rapid development of big data technology and machine learning methods, more and more studies have begun to try to use these advanced technologies for medal prediction. Traditional prediction methods usually rely on experience and statistical analysis, but these methods often fail to fully explore the deep relationships in the data. In this context, this study introduces the Lasso regression model, with its excellent feature selection ability and performance in high-dimensional data, to try to provide a new solution for Olympic medal prediction. Through the analysis of historical data, we can not only identify the important factors affecting the number of medals, but also construct a regression model with high prediction ability.

The main purpose of this study is to use the Lasso regression model to predict the medal table of the 2028 Los Angeles Summer Olympics and provide confidence intervals for the prediction results through uncertainty estimation. This process includes steps such as feature standardization, training of the Lasso regression model, and selecting the optimal regularization parameter λ through cross-validation. Through in-depth analysis of historical data, we not only predict the medal performance of Olympic powers such as the United States, but also provide relatively accurate medal expectations for other countries. These prediction results provide policymakers and sports organizations with effective information about the competitiveness of various countries in the Olympics.

2 Related Work

The research on Olympic medal prediction began with traditional statistical methods, such as regression analysis and time series models. Although these methods provide basic prediction capabilities, their ability to handle complex data structures and multidimensional features is limited, which affects the accuracy and generalization of predictions. With the development of machine learning technology, more and more research has begun to turn to more complex models such as support vector machines, decision trees, and random forests to improve the accuracy of predictions.

Lasso regression, by introducing L1 regularization, shows obvious advantages in high-dimensional data. Related studies have shown that Lasso regression can effectively select features, avoid overfitting, and maintain good prediction capabilities, especially in Olympic medal prediction. Through historical data analysis, Lasso regression helps identify important factors affecting the number of medals, and then build an efficient prediction model.

In recent years, deep learning methods have gradually been applied to medal prediction due to their strong ability to handle nonlinear and high-dimensional data. However, although these methods can provide higher accuracy, they lack good interpretability due to their "black box" characteristics and still need further optimization. In this context, some studies have combined Bayesian reasoning with traditional regression models, using prior knowledge and probability models to quantify the uncertainty of predictions, thereby improving the confidence and reliability of prediction results.

Although some progress has been made in existing research, Olympic medal predictions still face many challenges, such as how to deal with external factors such as individual differences among athletes, fluctuations in competition conditions, and how to incorporate emerging sports or changes in competition rules into prediction models. Therefore, how to continuously improve the accuracy and robustness of models in dynamic and complex systems remains a key direction for future research.

3 Medal table projections for the U.S. Summer Olympics in Los Angeles

3.1 Feature standardization

In Lasso regression, it is essential to perform feature standardization, and in this problem, standardization for features such as the number of all-time medals, the number of sport types, and the number of athletes can help the Lasso regression model to more accurately select features that contribute significantly to the prediction results. The goal of standardization is to adjust the mean of each feature to 0 and the standard deviation to 1 so that all features are compared under the same measure. The standardization formula is as follows:

$$X_{scaled} = \frac{X - \mu}{\sigma} \quad (1)$$

Where X_{scaled} is the standardized feature value and X is the original feature. μ is the mean of the feature. σ is the standard deviation of the feature. After standardization, all features will have the same scale, making them have an equal impact on the regularization term during training. Using the number of Olympic sports in the United States from 1986 to 2024.

3.2 Defining the Lasso regression model

Lasso (Least Absolute Shrinkage and Selection Operator) regression is a linear regression model with L1 regularization. It enables feature selection and prediction by adding a penalty term to the model that makes certain regression coefficients converge to zero. The goal of linear regression is to minimize the mean square error (MSE).

$$C_{OLS}(\beta) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2 \quad (2)$$

Where Y_i is the actual value of the target variable, X_i is the input feature, and β is the regression coefficient. Its objective function obtained by adding the L1 regularization term (i.e., the sum of the absolute values of the regression coefficients) to the loss function is as follows:

$$L(\beta) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3)$$

λ is the regularization parameter, which controls the strength of the regularization term. β_j is the regression coefficient. p is the dimension of the feature. The effect of the L1 regularization term is to drive some of the regression coefficients β_j towards zero, so that in the optimization process the Lasso regression selects the most important features and sets the regression coefficients of the unimportant features to zero.

The regularization term $\sum_{j=1}^{p} |\beta_j|$ in Lasso regression limits the size of the regression coefficients and prevents model overfitting. When $\lambda = 0$, Lasso regression degenerates into ordinary linear regression. And when λ tends to infinity, all regression coefficients tend to zero and the model becomes extremely simple. Through cross-validation, we can choose the most appropriate λ, which balances the complexity and fitting ability of the model. The objective function is to minimize the error while adding a regularization term to get the final regression coefficients:

$$\hat{\beta} = arcsin_{\beta} \left(\sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad (4)$$

3.3 Training the Lasso regression model

In this paper, we use Gradient Descent as an optimization method to find the optimal solution by calculating the gradient of the loss function and updating the coefficients along the gradient direction.

$$\beta_j^{(k+1)} = \beta_j^{(k)} - \eta\left[-2\sum_{i=1}^{n} X_{ij}\left(Y_i - X_i\beta\right) + sgn\left(\beta_j\right)\right] \tag{5}$$

The gradient of the loss function consists of two parts: the residual sum of squares and the L1 regularization term. The L1 regularization is not trivial at $\beta_j = 0$ and is therefore handled by the subgradient method. The learning rate η controls the step size of each parameter update. In this way, the gradient descent method is able to optimize the regression coefficients in the Lasso regression, and although the regularization term is not trivial at some points, the subgradient method is effective in solving this problem and makes the optimization process go smoothly.

Also Matlab based cross validation is used to select the best λ. The regularization parameter λ controls the complexity of the Lasso regression model. K-fold cross-validation is a technique that can effectively assess the generalization ability of a model by dividing the dataset into multiple subsets and training and validating the model multiple times. The results are shown below:

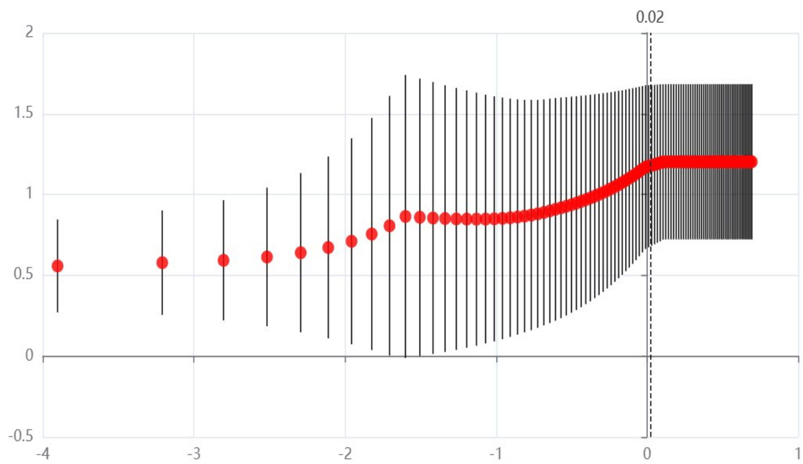

Figure 1 Lasso regression cross-validation plot

Figure 1 shows in visual form the selection of λ values using cross-validation. To minimize the mean square error determine λ = 0.02, log(λ) = −3.902.

4 Forecasting and uncertainty estimation

Having established the foundational models and analyzed the historical data, we now turn our attention to the critical task of forecasting future Olympic medal counts and estimating the associated uncertainties. Accurate forecasting is essential for strategic planning and resource allocation, while uncertainty estimation provides a measure of confidence in our predictions. In this section, we will detail the methods used to generate forecasts and quantify the uncertainties, ensuring that our predictions are both reliable and actionable.

Once the regression model is trained, we can use the trained regression coefficients to make predictions. For the new input data X_{new}, we can compute the predicted values $\hat{Y}_{new}$:

$$\hat{Y}_{new} = X_{new}\hat{\beta} \tag{6}$$

Where X_{new} is the new input data and $\hat{\beta}$ is the regression coefficients obtained from training.

To further assess the uncertainty, this paper calculates the standard error (SE) of the regression coefficients to derive the confidence intervals for the predictions. The formula for the confidence interval of the prediction is:

$$\hat{Y}_{new} \pm 1.96 \times SE(\hat{Y}_{new}) \tag{7}$$

Where, 1.96 is a constant of standard normal distribution indicating 95% confidence interval. The distribution of projected gold medal forecasts for the 2028 Los Angeles Olympics shows that the United States and China continue to lead the way in terms of gold medals and total medals, with a projected 42 and 40 gold medals and 129 and 88 total medals, respectively. Traditional sporting powerhouses such as Great Britain, France and Russia also perform well in the overall medal count, with projections of 63, 63 and 59 total medals respectively. Australia and Japan are performing well in specific events, while Italy, Germany and Canada are competitive in some events. The Netherlands, Korea and Brazil are competitive in individual events, while New Zealand, Spain and Ukraine are competitive in terms of total medals but have relatively few gold medals.

Figure 2 below shows the confidence intervals for the top 10 countries in terms of gold medals and total medals in the 2028 medal table:

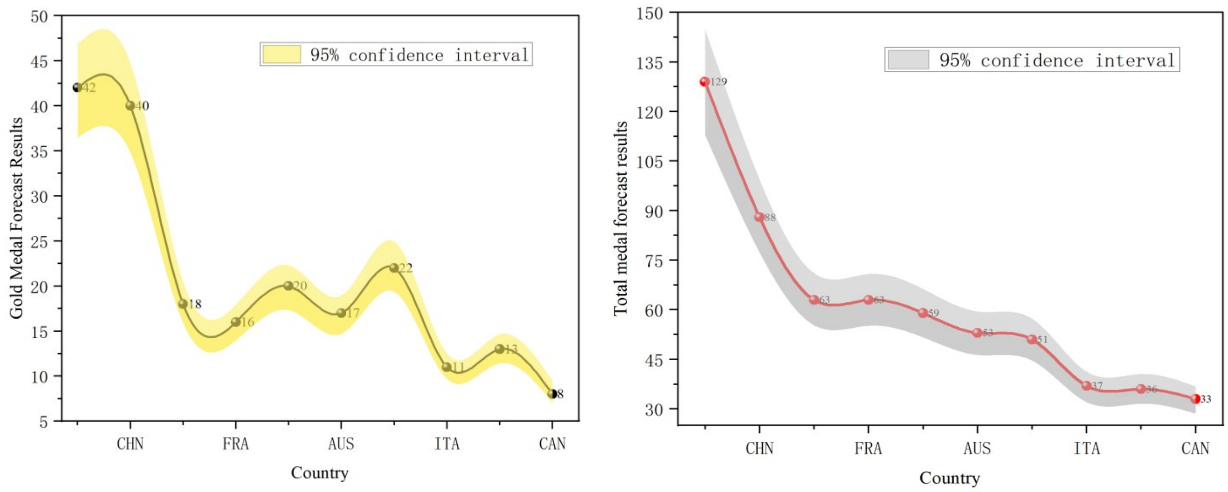

Figure 2 Example of Gold and Total Medal Forecast Intervals for a Country

Compared to the forecast range, we can also see that the gold medal and total medal counts of the United States are fluctuating between 42 and 129, while the gold medal and total medal counts of China are fluctuating between 40 and 89.

Using these analyses, we predict which countries are likely to achieve improved medal performances at the 2028 Olympics, identifying some countries that may be at risk of declining. The table 1 below shows the results of our predictions based on these factors.

Table 1 Improved and declining country projections

NOC							
Progress				Decline			
USA	EST	NGA	CZE	FRA	EGY	CHN	GEO
GRC	BRA	MNG	BLR	FRA	IRL	MAR	MOZ
DEU	JPN	UGA	SVK	AUT	CHL	SYR	SAU
GBR	POL	CMR	ARM	AUS	PHL	DZA	KWT
HUN	TUR	PRK	BDI	ITA	PER	DOM	KGZ
HUN	LVA	COL	ECU	CAN	KOR	ZMB	SRB
DNK	MEX	THA	VNM	NOR	LBN	IDN	AFG
CHE	JAM	ZWE	ARE	NLD	BGR	NAM	BHR
BEL	TTO	SUR	TJK	SWE	ISL	SVN	BWA
CUB	PAN	CRI	SGP	NZL	PAK	MYS	CYP
ESP	IRN	LTU	GRD	ARG	TUN	UKR	GTM
LUX	PRI	HRV	FJI	URY	GHA	KAZ	JOR
IND	VEN	ISR	XKS	HTI	KEN	MDA	
FIN	BHS	QAT		PRT	NER	UZB	
ZAF	ETH	RUS		ROU	BMU	AZE	

5 Model evaluation

The purpose of the model assessment was to evaluate the performance of the Lasso regression model to ensure its ability to generalize over unseen data. The evaluation metrics used included mean square error MSE and coefficient of determination R^2, Mean Square Error (MSE) is an important indicator to assess the performance of a regression model, which indicates the difference between the predicted and actual values of the model. The smaller the MSE is, the better the predictive ability of the model. The formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{8}$$

The coefficient of determination (R^2) is a measure of the goodness of fit of a model and indicates the proportion of the variance of the target variable that is explained by the model.The closer the R^2 is to 1, the better the model fit. The formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2} \tag{9}$$

The parameters of the final prediction results assessment are shown in the table 2 below:

Table 2 Model Evaluation

variable name	Standardized coefficient	Non-standardized coefficient	R^2	MSE
intercept (the point at which a line crosses the x- or y-axis)	719.128	719.128	0.943	15.0209
Athlete Count	0.079	0.079		
Total events	0.173	0.173		
Year	-0.385	-0.385		

6 Discussion

In this study, we used the Lasso regression model to predict the number of medals for the 2028 Los Angeles Summer Olympics in the United States, and further estimated the uncertainty of the prediction results. By standardizing the relevant features, building the Lasso regression model, and applying cross-validation, we were able to obtain relatively accurate predictions and quantify the results through uncertainty analysis. Although our model has been able to effectively capture the trends in the data and provide specific medal prediction intervals, there are still some challenges and potential room for improvement in practical applications.

First, although the Lasso regression can effectively select important features and avoid overfitting through its L1 regularization mechanism, the prediction ability of the model is still limited by some external factors. For example, uncertain factors in the Olympic Games, such as fluctuations in athlete performance, changes in stadium conditions, and even political factors, can affect the final number of medals. Therefore, although our model provides a reasonable framework for prediction, these external factors may affect the accuracy of the model to a certain extent. In future studies, it is possible to consider introducing more variables or even using more complex models (such as random forests or neural networks) to further improve the accuracy of predictions.

Secondly, although we used standardization and cross-validation to optimize the parameters of the Lasso regression model, the results showed that the medal prediction intervals for some countries were still relatively wide. For example, for the number of medals for the United States and China, although the predicted total number and number of gold medals were relatively clear, the final prediction intervals showed large fluctuations due to the large uncertainty in the competitive intensity and performance of these countries. This uncertainty shows that although the current prediction model provides an important reference for strategic planning, in practical applications, policymakers and resource allocators need to consider more temporary factors, such as athletes' injuries and training status. Therefore, in practical applications, the prediction results should be used as a reference rather than an absolute basis.

In addition, in terms of uncertainty estimation, we used standard errors and confidence intervals to quantify the reliability of the predictions. Although this method can provide a certain degree of confidence, there is still room for improvement. For example, we can combine Bayesian statistical methods to further refine the uncertainty analysis to better handle complex systems and dynamically changing situations. With these further improvements, the model can more accurately predict the number of medals that may appear in different situations and provide more accurate data support for the preparations of various countries.

Finally, although this study made predictions based on historical data, future predictions may be affected by emerging factors, such as the addition of new sports, changes in competition systems, and the evolution of global sports trends. Therefore, our model needs to be updated regularly to reflect new data and changing trends. This is especially important because the Olympic Games are not only a showcase for sports competitions, but also an important stage for the intersection of global politics, economy, and culture, and many variables may have an impact on the final medal distribution.

7 Conclusion

In this study, we proposed a method based on the Lasso regression model to predict the medal table of the 2028 Los Angeles Summer Olympics and conducted a detailed analysis of the uncertainty of the prediction results. By standardizing features, building a Lasso regression model, and performing cross-validation, we are able to effectively extract predictive information from historical data. Our model successfully predicts the number of medals for major

Olympic powers, including the United States and China, and provides valuable prediction intervals for other countries.

However, although the model can provide reasonable prediction results, we also recognize that the complexity of Olympic competitions is far beyond the scope that can be captured by simple data modeling. In addition to the trends in historical data, athletes' performance fluctuations, changes in competition venue conditions, and other unforeseen external factors may affect the final medal distribution. Therefore, although the prediction provides strong support for strategic decision-making, it is still necessary to combine real-time data and other dynamic factors in practical applications to make more flexible and effective resource allocation and decision-making.

In addition, our research also reveals the space for model optimization. For example, by combining more factors, adopting more complex prediction models, and even introducing advanced uncertainty assessment techniques such as Bayesian methods, the accuracy of the model and the reliability of predictions may be further improved. Nevertheless, the current Lasso regression model provides a solid foundation for understanding medal prediction and uncertainty estimation, and demonstrates the potential of data-driven methods in large-scale sports event prediction.

In summary, this study not only provides an effective framework for Olympic medal prediction, but also demonstrates how to evaluate the reliability of predictions through quantitative methods. Future research should continue to explore different modeling methods and adapt to the changing data environment to better provide decision makers and Olympic-related organizations with accurate and forward-looking prediction tools. This will help improve the efficiency of the preparation and planning of the Olympic Games, while promoting fair and reasonable distribution of sports resources worldwide.

About the Author

Ruoyu Zhang Undergraduate study, mechanical design, manufacturing and its automation.

References

[1] Tschisgale, P., Kubsch, M., Wulff, P., Petersen, S., & Neumann, K. (2025). Exploring the sequential structure of students' physics problem-solving approaches using process mining and sequence analysis. Physical Review Physics Education Research, 21(1), 010111.

[2] Yan, X., & Li, L. (2024). Censoring the intellectual public space in China: What topics are not allowed and who gets blacklisted?. Perspectives on Politics, 22(3), 753-770.

[3] Ramiller, A., Song, T., Parker, M., & Chapple, K. (2024). Residential Mobility and Big Data. Cityscape, 26(3), 227-240.

[4] Serebrennikov, D., & Skougarevskiy, D. (2024). A tale of four cities: Exploring security through environmental characteristics of CCTV equipment placement. Journal of Computational Social Science, 7(3), 2735-2766.

[5] Jiang, Y., Si, J., Wang, Y., Zhang, H., Zhou, F., Lu, X., ... & Wang, Z. (2024). The Relationship Between PM2. 5 and Eight Common Lung Diseases: A Two-Sample Mendelian Randomization Analysis. Toxics, 12(12), 851.