Statistical Case Study

Garrett Wen

October 8, 2023

Contents

1 Aug 30, 2023 at 16:30:19

Figure out the plan and goal. ghggg

- Factors (from data that are influential)
- Measures about success. We can give a high dimensional vector, every dimension is an indicator about some kind of success of the athletes.
- Selection process (use the high dimensional)

2 Aug 31, 2023 at 19:35:49

Gang Wen's idea:

2.1 Basic Info of Datasets

2.2 data_2017_2021.csv and data_2022_2023.csv

data_2017_2021.csv and data_2022_2023.csv files contain gymnastics performance metrics ranging from various years. The columns in this dataset are described as follows:

- LastName: The last name of the gymnast.
- FirstName: The first name of the gymnast.
- Gender: The gender of the gymnast.
- Country: The country that the gymnast represents.
- Date: The date on which the performance took place.
- Competition: The name of the competition.
- Round: The round of the competition.
- Location: The location where the competition took place.
- Apparatus: The specific gymnastics apparatus.
- Rank: The ranking achieved in that particular performance.
- D_Score: The difficulty score assigned for the routine.
- E_Score: The execution score assigned for the routine.
- Penalty: Any penalties incurred during the performance.
- Score: The total score of the performance.

2.3 Components and Their Calculation Methods

To thoroughly assess the gymnasts, we will generate a vector with high dimensions, where each component will correspond to one of the indices listed below.:

- **Overall Score**: This is computed as the average 'Score' across all performances for each gymnast.
- **Average D-Score**: This is calculated as the average 'D_Score' for each gymnast across all performances.
- **Average E-Score**: This is calculated as the average 'E_Score' for each gymnast across all performances.
- **Average Penalty**: This is the average 'Penalty' incurred by each gymnast across all performances.
- **Specialty Score**: This is the average 'Score' on the apparatus where each gymnast has the highest average performance.
- **Consistency Score**: This is calculated as the standard deviation of the 'Score' for each gymnast across all performances.

Note that when calculating those 'averages', we may weight them according to time, i.e. more recent time score will get more weights.

2.4 Find a Metric to Define 'Success'

To synthesize these metrics into a singular measure of each gymnast's potential for future performance, we need a function to map this vector to a final metric. For example, a linear weighted model will be used. The Composite Score for each gymnast will be calculated using the following formula:

$$\textit{Composite Score} = w_1 \times \textit{Overall Score} + w_2 \times \textit{Average D-Score} + w_3 \times \textit{Average E-Score}$$

$$-w_4 \times \textit{Average Penalty} + w_5 \times \textit{Specialty Score} - w_6 \times \textit{Consistency Score}$$

Here, w_1 through w_6 are the weights assigned to each metric according to possible preferences. The gymnast with the highest Composite Score is deemed to be the most promising for future competitions. We can also use more complex models for the final metric, possibly a non-linear method by machine learning.

2.5 Final Remarks

Finally, it should be noted that we should certainly give more weights to the data for recent times (that means, we should try to weight those scores according to time). Additionally, we should also strive to find more useful data and conduct similar analyses.

3 Oct 8, 2023 at 20:27:01

Python codes written by Ruixiao Wang and Siyu Chen. Documents, R codes and plans written by Gang Wen. We use ChatGPT for the draft of the translation of python codes to R codes, and polish the English of this documents.

3.1 Introduction

The provided code is designed to analyze and predict the performance of gymnasts in different events. It simulates rankings for gymnasts across various apparatuses, both for individual and team events, and summarizes medal counts by country.

3.2 Prediction Method

3.2.1 Load and Data Preparation

Initially, the raw gymnastic data is loaded and subjected to preprocessing. This includes removing any missing values, outliers, and normalization (if required).

3.2.2 Team Selection

For each country, a team is formed by selecting the top-performing gymnasts based on their scores in previous competitions. The team size and the number of substitutes can vary based on the event's requirements.

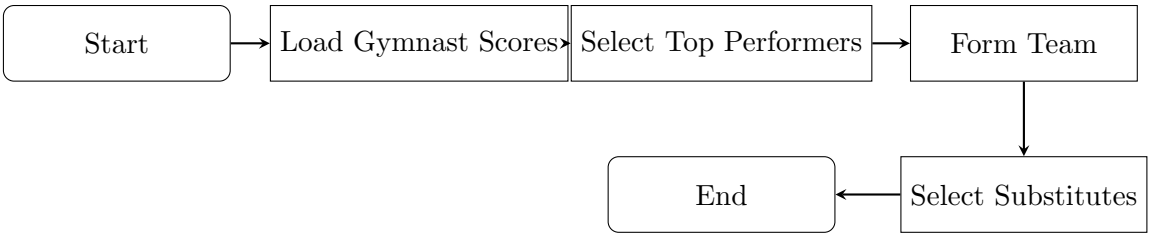

Figure 1: Flowchart for the team selection process

3.2.3 Prediction using Noise

To simulate the variability and unpredictability in real-world performances, random noise is added to the gymnasts' previous scores. This noise is generated using a Gaussian distribution with a mean of zero and a standard deviation that represents the typical variability in gymnast scores. The exact standard deviation value can be adjusted based on the event and the historical performance data.

3.2.4 Simulation and Ranking

Once the noise is added, the gymnasts' scores are used to simulate their performance in the event. Gymnasts are then ranked based on these simulated scores.

3.3 Interpreting the Prediction Results

The results are summarized in a CSV file, which has the following columns:

Column	**Description**
Country Names	Names of the participating countries.
simulationX	Total number of medals won by the country in the X^{th} simulation.

Table 1: Description of columns in the result CSV file

To interpret the results:

- Look up the 'Country Names' column to identify a specific country.
- Check the 'simulationX' columns to see the number of medals the country won in each simulation.
- The average across all simulation columns provides an estimate of the expected medal count for that country.

3.4 Data Analysis Flowchart

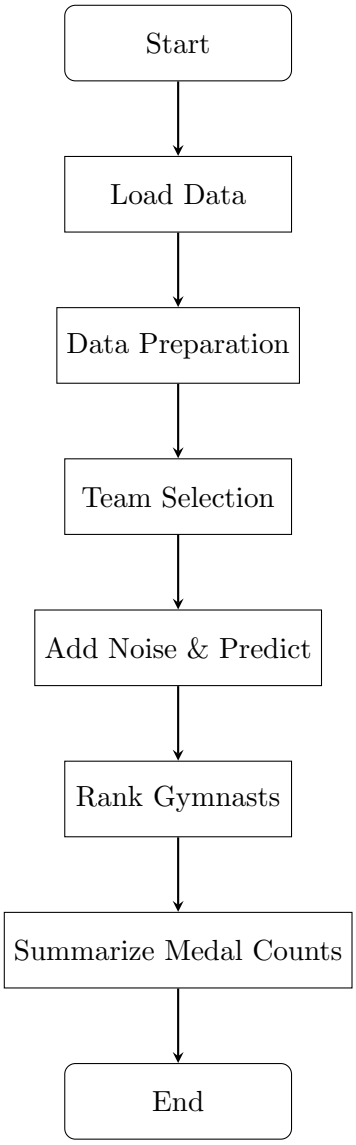

Figure 2: Flowchart for the detailed data analysis process

3.5 Conclusion

The code provides a comprehensive method for analyzing gymnastic data, predicting performance, and summarizing results. It's important to note that the predictions are based on past performance data and include random noise, so they should be interpreted as estimates rather than definitive outcomes.