UCSAS Methodologies Used Report

Yihao Li

Akkasit Khongtong

Albert Lu

Instructor: Luke Cherveny

1 Introduction

The selection of athletes for the Olympic Games is a critical decision that combines art and science, subjective judgment, and objective data analysis. In preparation for the 2024 Olympics, we embarked on a data-driven project to optimize the selection of Team USA's gymnastics squad. Our goal was to maximize the potential medal count using a combination of statistical analysis, machine learning modeling, and mathematical optimization.

2 Overview

The project was structured into several key phases:

- Data Collection and Preparation: We compiled a comprehensive dataset of Team USA gymnasts, including scores, rankings, and medal histories from recent competitions.
- Exploratory Data Analysis (EDA): We conducted an initial analysis to understand the distribution of scores and ranks and identify top performers.
- Machine Learning Model Development: We developed a predictive model to estimate the potential medal points for each gymnast.
- Optimization of Team Selection: We used linear programming to select a team that maximizes the predicted medal count, considering constraints like team composition.
- Gymnast Prediction Results: We provided the prediction results from our model as the list of gymnasts that would perform best among other American gymnasts in each competition category, which consists of Team All-Around, Individual All-Around, and Individual Event.

- Scenario Analysis and Interpretation: We interpreted the results, considering various scenarios and constraints, to provide recommendations for team selection.

The following sections delve into each phase, outlining the methodologies, analyses conducted, and the insights gained.

3 Data Collection and Preparation

We began by gathering detailed performance data of Team USA gymnasts over recent years. This dataset included individual scores in various apparatuses, overall rankings in competitions, and the medals won. Key to our approach was ensuring data accuracy and comprehensiveness, as these would form the foundation of our subsequent analyses.

3.1 Data Cleaning and Structuring

The dataset required meticulous cleaning and structuring. We handled missing values, normalized score ranges, and aggregated data at the gymnast level, ensuring a clean and analyzable dataset.

4 Exploratory Data Analysis (EDA)

4.1 Score and Rank Distribution (female gymnasts)

We first visualized the distribution of scores and ranks among Team USA gymnasts. Histograms and boxplots revealed patterns in performance across different apparatuses, highlighting areas of strength and potential improvement.

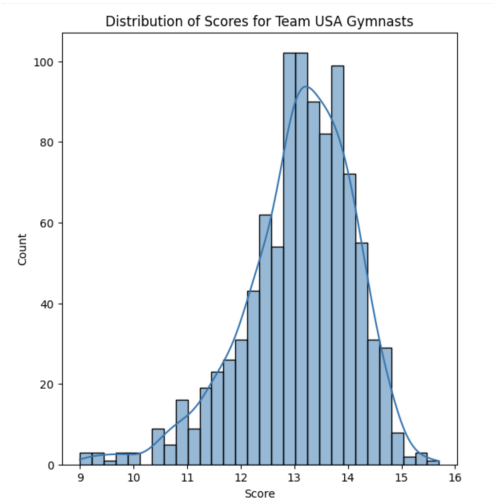

Figure 1: Female Distribution of Scores for Team USA Gymnasts

This histogram overlaid with a kernel density estimate shows the distribution of scores for Team USA female gymnasts. The distribution appears to be roughly normal, centering around a score of 13, which suggests that most gymnasts score near this value during their performances.

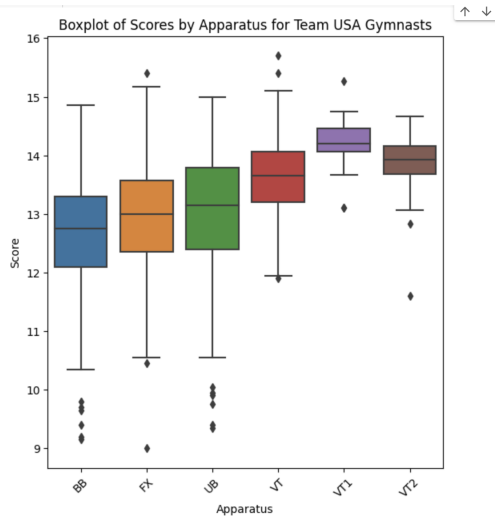

Figure 2: Female Boxplot of Ranks by Apparatus for Team USA Gymnast

The boxplot provides a comparison of scores across different apparatuses. Each box represents the interquartile range (IQR) of scores, with the median score represented by a line in the middle of each box. The "whiskers" extend to the furthest points that are not considered outliers. Points beyond the whiskers are plotted individually as outliers. This visualization shows variability in scores across apparatuses, with some (like VT1 and VT2) showing higher median scores and less variability.

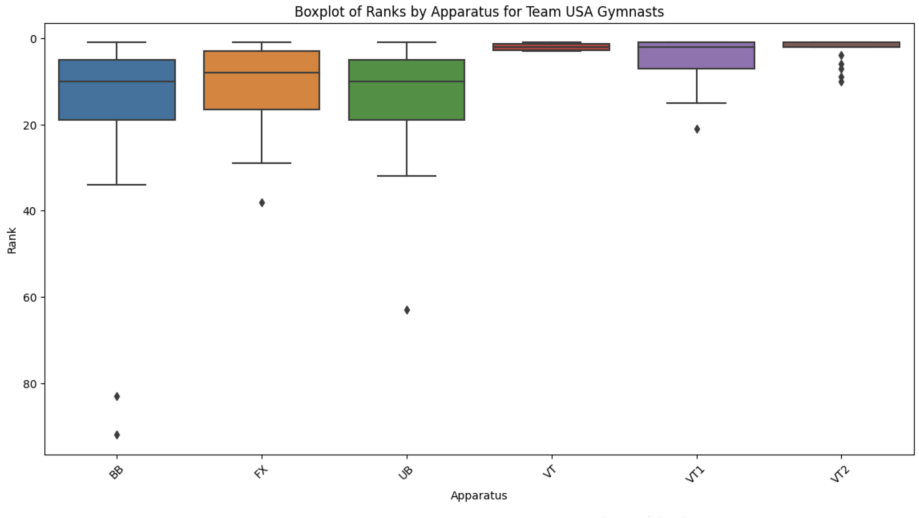

Figure 3: Female Boxplot of Ranks by Apparatus for Team USA Gymnast

Similar to the scores boxplot, this visualization shows the ranks of gymnasts across different apparatuses. Lower ranks are better, indicating a higher placement. This chart shows that for certain apparatuses, there's

a wide range of ranks among gymnasts, while others have a tighter grouping, suggesting more consistent performances across athletes.

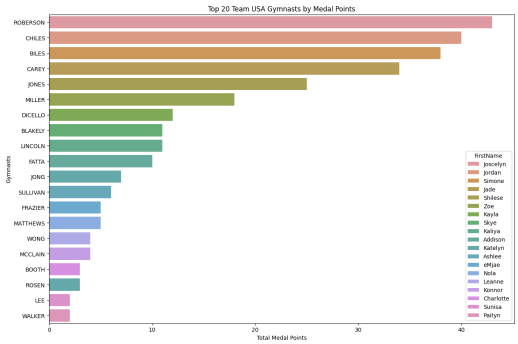

Figure 4: Female Boxplot of Ranks by Apparatus for Team USA Gymnast

This bar chart ranks the top 20 USA female gymnasts by predicted medal points, likely derived from the RandomForest model's predictions. Joscelyn Roberson, Jordan Chiles, and Simone Biles have higher predicted medal points, indicating they are expected to perform well in competitions.

4.2 Score and Rank Distribution (male gymnasts)

For the male gymnast data exploratory, we reiterated the steps similar to the female; however, more apparatuses were added in the male competition, which were High Bar(HB), Parallel Bars (PB), Pommel Horse (PH), and Still Rings (SR).

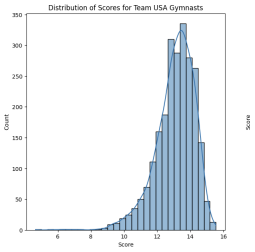

Figure 5: Male Distribution of Scores for Team USA Gymnasts

This histogram overlaid with a kernel density estimate shows the distribution of scores for Team USA male gymnasts. The histogram presents the left-skewed distribution, centering around a score of 13, which we can conclude that most male gymnasts perform near this score value.

This boxplot provides a comparison of scores across different apparatuses for male gymnasts. Similar to female performance, male gymnasts perform on VT1 and VT2 with high median scores and less variance; most scores are between 14 and 15. On the other hand, for Pommel Horse, has the lowest mean score (around 13.0) with more score variability.

This visualization shows the ranks of gymnasts across different apparatuses. Lower ranks are better,

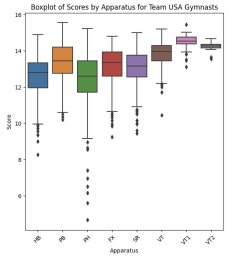

Figure 6: Male Boxplot of Ranks by Apparatus for Team USA Gymnast

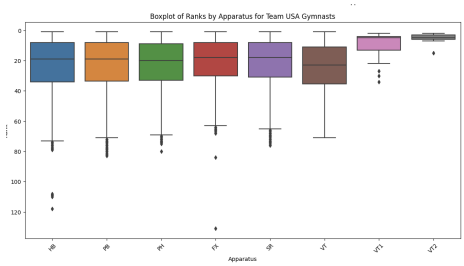

Figure 7: Male Boxplot of Ranks by Apparatus for Team USA Gymnast

indicating a higher placement. This chart shows that American male gymnasts are more likely to perform well on VT1 and VT2 and place on the higher rank. If we send the candidate that is doing well on VT1 VT2, the US team will have more chance to win the medal.

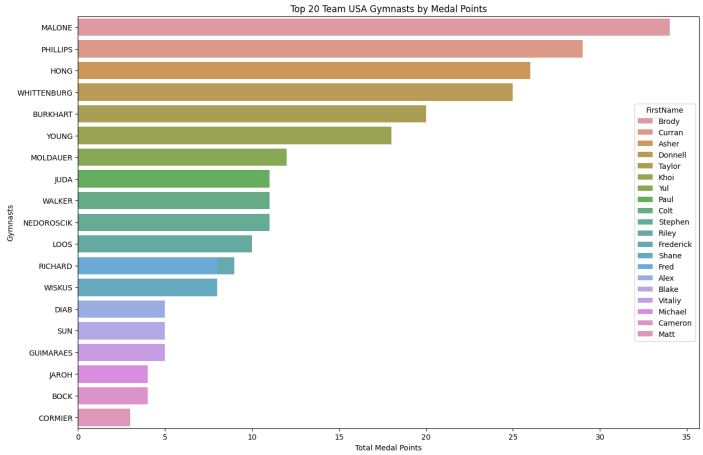

Figure 8: Bar Charts on Male Top 20 Team USA Gymnasts by Medal Points

This bar chart ranks the top 20 USA male gymnasts by predicted medal points from the RandomForest model's predictions. Gymnasts like Brody Malone, Curran Phillips, and Asher Hong have higher predicted medal points, indicating they are expected to perform well in competitions.

4.3 Medal Points Analysis

A crucial part of our EDA was analyzing the distribution of medal points. We assigned weighted values to different medal types (gold, silver, bronze) and identified gymnasts who had accumulated the highest medal points, indicating consistent high performance.

5 Machine Learning Model Development

To predict the potential medal points for each gymnast, we employed machine learning techniques. This predictive modeling was aimed at quantifying the likely contribution of each gymnast to the team's overall medal count.

5.1 Feature Engineering and Model Selection

We identified key features from the data, including average scores, consistency (standard deviation of scores), and historical medal counts. These features were used to train a RandomForestRegressor, a robust model capable of capturing complex, non-linear relationships in the data.

5.2 Model Training and Evaluation

The model was trained on a subset of the data, with the remaining portion used for validation. We used Mean Squared Error (MSE) as our primary metric to evaluate the model's accuracy. A lower MSE indicated a model with better predictive capability.

5.3 Optimization of Team Selection

With the model predictions in hand, our next step was to select the optimal team composition using mathematical optimization.

5.3.1 Setting Up the Optimization Problem

We used the PuLP library for linear programming to solve our optimization problem. The objective was to maximize the sum of predicted medal points while selecting a team of five gymnasts.

5.3.2 Constraints

The primary constraint was the team size - exactly five gymnasts had to be selected. We also explored additional constraints, such as including at least one all-around gymnast or a specialist in a particular apparatus.

5.4 Solving the Problem

The optimization problem was solved, providing a set of gymnasts that, according to the model's predictions, would maximize the total medal count.

6 Scenario Analysis and Interpretation

The results from the optimization provided valuable insights but also raised important considerations.

6.1 Interpretation of Results

The selected team comprised gymnasts who were predicted to contribute maximally to the medal tally. However, it was essential to interpret these results within the broader context of athletic performance, which encompasses unpredictable elements like form, fitness, and psychological readiness.

6.2 Consideration of Different Scenarios

We examined various scenarios, adjusting constraints and re-running the optimization. For example, we considered the impact of including a specialist gymnast and evaluated if their inclusion would increase the predicted medal count.

6.3 Incorporating Gender into the Model

Recognizing the distinct differences in men's and women's gymnastics, we extended our analysis to separately optimize the male and female teams. This approach allowed us to tailor strategies to the unique strengths and competition formats of each gender's events.

6.4 Separate Models for Men and Women

We developed separate machine-learning models for male and female gymnasts. This differentiation acknowledged the variations in scoring patterns, event types, and competition dynamics between men's and women's gymnastics.

6.4.1 Gender-Specific Optimization

We conducted separate optimizations for the men's and women's teams using the same linear programming technique. Each optimization aimed to maximize the respective team's total predicted medal count, with constraints and selections tailored to the specific needs and strengths of each gender's team.

7 Candidate Prediction Results

7.1 Team all-around

For team all-around competition, the team's score will depend on the sum of scores from all gymnasts who participate across all apparatuses. Here is the list of the top gymnasts with the highest total medal points from our prediction for male and female gymnast team.

7.1.1 Female Gymnast all-around Prediction, one medal

Number	FirstName	LastName	Predicted Medal Points	Age
1	Simone	Biles	33.01	26
2	Jordan	Chiles	32.29	23
3	Jade	Carey	21.86	23
4	Shilese	Jones	21.40	21
5	Joscelyn	Roberson	15.87	17

7.1.2 Male Gymnast Team all- around Prediction, one medal

Number	FirstName	LastName	Predicted Medal Points	Age
1	Brody	Malone	28.41	23
2	Asher	Hong	23.85	19
3	Donnell	Whittenburg	21.86	29
4	Curran	Phillips	21.40	23
5	Taylor	Burkhart	15.87	21

The prediction results of Team All-around, for female gymnasts Simone, Jordan, and Jade have the predicted medal points of 33.01, 32.29, and 21.86, respectively. For the male team, Brody, Asher, and Dornnell hs 28.41, 23.85, and 21.86, respectively. We are curious if the teams have a suitable variety of age, we found that both male and female teams have a good combination of age. For instance, Simone, who has the highest predicted score, has the highest age of 26, while Joyscelyn has 17 years old. Which is the good combination that can enhance the team strength.

7.2 Individual all-around

As the individual all-around will count on the sum of scores a gymnast performs in all apparatus, here is the prediction result for the male and female gymnasts with the highest predicted medal point for all instruments.

7.2.1 Female Individual all-around Prediction, one medal

Number	FirstName	LastName	Predicted Medal Points	Age
1	Simone	Biles	33.01	26
2	Jordan	Chiles	32.29	23

7.2.2 Male Individual all-around Prediction, one medal

Number	FirstName	LastName	Predicted Medal Points	Age
1	Brody	Malone	28.41	23
2	Asher	Hong	23.85	19

The candidates of the US team for the individual all-around event are Simone Biles and Jordance Chiles for the female competition, Broder and Asher for male competition; all of them are a good choice since winning the individual all-around gymnastics competition generally requires a combination of factors, including skills, techniques, consistency, and experience, which are reflected on their high predicted score.

7.3 Individual Event

For a single apparatus competition that emphasizes a specific apparatus, the total score, which is the summation of the difficulty score and the execution score, will determine a gymnast's final score and rank on that particular apparatus. Here is the list of the first and second-best American gymnasts with the highest medal point for each apparatuses.

7.3.1 Female Individual Event, four medals

Apparatus	1st Gymnast	Age	1st Score	2nd Gymnast	Age	2nd Score
floor exercise	Biles	26	14.87	Roberson	17	14.08
vault	Biles	26	15.17	Carey	23	14.58
balance beam	Biles	26	14.60	Mcclain	18	14.50
uneven bar	Jones	21	14.51	Biles	26	14.26

7.3.2 Male Individual Event, six medals

Apparatus	1st Gymnast	Age	1st Score	2nd Gymnast	Age	2nd Score
floor exercise	Juda	22	14.34	Moldauer	27	14.30
pommel horse	Young	20	14.39	Guimaraes	23	14.35
still rings	Whittenburg	29	14.54	Diab	26	14.20
vault	Young	20	14.77	Karnes	19	14.73
parallel bars	Phillips	23	14.89	Walker	22	14.67
high bar	Malone	23	14.37	Phillips	23	14.12

For female team, Simone Biles has the top scores in floor exercise, vault, and balance beam, and a second-best score in the uneven bars, would be the leading candidate for the competition. Shilese Jones is predicted to be the best on the uneven bars. Joscelyn Roberson, Jade Carey, and Konnor McClain are anticipated to have the second-best scores in their respective events, making them strong contenders as well. Biles and Jones would likely be the top selections based on these predictions. For male team, Paul Juda is predicted best in floor exercise, Khio Young for pommel horse, Donnell Whittenburg for still rings, Khio Young for vault, Curran Phillips for parallel bars, and Brodyn Malone for high bars, Yul Moldauer, Vitaliy Guimaraes, Alex Diab, Josh Karnes, Colt Walker are anticipated to have the second-best scores in their respective events. All of them are good choices. Male team is more competitive to choose the candidate since their scores are close. Female team might need to consider choose second choice player other than Simone Biles.

8 Final Recommendations and Summary

8.1 Recommendations

- Balanced Team Composition: For both the men's and women's teams, a mix of all-around gymnasts and event specialists is recommended to cover all events effectively and maximize medal potential.
- Utilization of Specialists: Inclusion of specialists, particularly in events where Team USA historically lacks medal contenders, could be a strategic advantage.
- Continuous Data Monitoring: As new performance data becomes available, it's crucial to update the models and re-optimize the team selection, considering current form and fitness levels.
- Consideration of Qualitative Factors: Beyond data-driven insights, factors such as team dynamics, mental toughness, and experience should be integral to the final team selection.
- Age and other health data for each gymnast should be included since it can reflect the experience and overall readiness of a gymnast for competition.

8.2 Summary

Our comprehensive data-driven approach combined exploratory data analysis, machine learning predictions, and mathematical optimization to recommend an optimal team composition for Team USA in the upcoming Olympics. While our models provided quantitative insights into potential medal-winning gymnasts, the final team selection should blend these insights with expert judgment and qualitative evaluations. This balanced approach aims to ensure that Team USA fields the strongest possible gymnastics team, maximizing the chances of success at the Olympics.

9 Citation

https://statds.org/events/ucsas2024/challenge.html