Abstract

Utilizing data spanning from 2013 to 2023 from a diverse range of sources, this

research employs data scraping, cleaning, exploratory data analysis (EDA), and

modeling to identify the optimal Men's and Women's Teams and Individual USA

Olympic Artistic Gymnasts for the 2024 Olympic Games. The term "best" is defined as

maximizing the total medal count in artistic gymnastics for the United States. The study

integrates machine learning models, combinatorial optimization algorithms, and

advanced data analysis techniques. Initial phases focus on predicting individual

apparatus outcomes using binary classification, with evaluation based on key metrics

such as accuracy, precision, recall, F1-Score, and F2-Score. During my investigation

into team selection optimization, mixed-integer linear programming was a recurring

theme. However, my research diverges from the norm as I embrace a novel approach

that leverages Metaheuristics Algorithms, specifically Randomized Heuristic, Tabu

Search, and Variable Neighborhood Search (VNS). Grounded in data analysis and

algorithmic decision-making, the research strategically positions Team USA to

maximize their medal count in the Paris 2024 Olympics.

1. Introduction

This research focuses on selecting the optimal Men's and Women's USA

Olympic Artistic Gymnastics teams for the Paris 2024 Olympics, aiming to

position Team USA for maximum medal count. The challenge involves

identifying exceptional athletes for each team, considering both individual and

team event success. Leveraging advanced analytics models, this study

forecasts and compares medal counts, providing a data-driven foundation for

strategic decision-making in the 2024 Olympic Games.

To determine the "best" gymnasts, baseline models for individual apparatus

outcomes are established, with the ZeroR algorithm as a foundational

benchmark. Various machine learning models, including Random Forest

Classifier, AdaBoost Classifier, Support Vector Classifier (SVC), K Neighbors

Classifier, Decision Tree Classifier, Gaussian Naïve Bayes, and Neural

Networks, comprehensively assess individual gymnasts' performance and

influencing factors.

Optimal team selection involves navigating complexities with optimization

algorithms such as Randomized Heuristics, Stochastic Control, Tabu Search,

Variable Neighborhood Search, and Brute Force. These algorithms play a

pivotal role in optimizing team selection to maximize point totals.

Based on a diverse dataset spanning 2015 to 2023 from reputable platforms

like Thegymter.net, Wikipedia.org, and the UConn Sports Analytics Symposium

2024, the study integrates machine learning models, optimization algorithms,

and comprehensive data analysis. The goal is to provide Team USA with a

strategic advantage to maximize their medal count in the Paris 2024 Olympics.

2. Data Collection and Preprocessing

1

2.1. Data Sources

This research derives its data from reputable platforms such as

Thegymter.net, UCSAS 2024 USOPC Data Challenge, and

Wikipedia.org. The dataset spans from 2013 to 2023, focusing primarily

on international competitions like world cups, world championships, and

the Olympics. Key data includes individual scores, overall

performances, event rankings, athlete details, and various scores

(execution, difficulty, penalties). Due to time constraints, the primary

emphasis in data collection lies on the years 2023 and 2022, resulting

in a substantial dataset, albeit one that may not be as exhaustive as

ideal.

2.2. Data Preprocessing Steps

The data preprocessing stage is crucial for ensuring dataset integrity

and usability. In the 'Scraping' directory, web scraping code generates

CSV files of raw data. The 'Cleaning' directory addresses tasks like

filling/removing missing values and standardizing column names. The

'Combine Data' directory facilitates the integration of information from

diverse competitions for each event, offering a comprehensive

perspective. Acknowledging dataset imbalances, especially in the

minority class, poses challenges. Therefore, Robust models for

imbalanced datasets will be employed to ensure fair and accurate

2

representation across classes. These preprocessing steps transform

raw data into a clean, structured, and uniform dataset for accurate and

insightful analysis.

3. Methodology

3.1. Individual

3.1.1. Models

Various machine learning models were utilized to predict

individual apparatus outcomes. Notable models include the

Random Forest Classifier, AdaBoost Classifier, Support Vector

Classifier (SVC), K-Neighbors Classifier, Decision Tree

Classifier, Gaussian Naïve Bayes, and Neural Networks.

3.1.2. Hyperparameter Tuning

To optimize model performance, a Grid Search approach was

employed for hyperparameter tuning. This systematic method

explored predefined hyperparameter spaces to enhance model

effectiveness.

3

3.1.3. Cross-validation Strategies

Holdout validation and k-folds cross-validation were employed.

Holdout validation divided the dataset into training and validation

sets, while k-folds cross-validation involved dividing samples

into equal-sized folds for training and testing.

3.1.4. Base Model

The ZeroR classifier, a simple method predicting the majority

class, served as a baseline for comparing other classification

methods.

3.1.5. Confusion Matrix

Utilized for performance evaluation, the confusion matrix

provided insights into accuracy, precision, F1-Score, and

F2-Score, crucial for assessing model proficiency.

These comprehensive evaluation metrics provide a robust analysis of

the individual machine learning models' performance in predicting

gymnastics outcomes.

3.2. Team

In the process of team selection, two distinct datasets were employed.

The initial dataset prioritized comprehensive all-around scores, while

the second dataset focused on individual apparatus scores for USA

gymnasts.

3.2.1. Qualification Round

The qualification round uses 4 up 3 counts, where four athletes

will compete on each apparatus. The cumulative total of the

three highest scores on each apparatus determines the

advancing teams. Several algorithms were used with the same

goal in mind; to get the three highest scores from four gymnasts.

3.2.1.1. Men's team selection

Tabu Search, Randomized Heuristic, and Variable

Neighborhood Search algorithms were used,

incorporating data from 2022 to 2023. In case of multiple

entries, the maximum score was chosen, capturing each

gymnast's optimal potential.

3.2.1.2. Women's team selection

Tabu Search, Randomized Heuristic, Variable

Neighborhood Search, and Brute Force algorithms were

used with datasets from 2022 to 2023. For multiple

entries, the mean score was used for Team USA,

ensuring reliability in assessing each gymnast's

performance. In case of multiple entries for other

countries, the maximum score was chosen, capturing

each gymnast's optimal potential. While it may appear

unfair, the forthcoming observation will demonstrate that

Team USA simply outperforms all other countries by a

4

significant margin. (The maximum scores for Team USA

were initially employed, establishing a significant lead of

over 4.5 points ahead of the second-place contender.)

3.2.2. Finals

A 3 up 3 count structure in the finals involves three gymnasts

per apparatus, with cumulative scores contributing to the final

team score. Selection criteria were based on Individual

All-Around (IAA) scores, concurrently considering the inclusion

of potential individual medalists in the final roster.

4. Evaluation and Results

4.1. Individual Apparatus Evaluation and Results

This analysis draws upon historical data encompassing Men's and

Women's international gymnastics competitions. The models for

Women were trained on data spanning from 2013 to 2021, while the

models for Men were trained on a more limited timeframe from 2018 to

2020, due to time constraints. The evaluation of the models involves the

utilization of unseen data from the years 2022 and 2023 for inference.

4.2. Model Performance Metrics

For each apparatus, the data was run through the various models

(3.1.1.), and their accuracy, precision, F1-scores, and F2-scores were

compared. Once the best model was selected, judged by its accuracy

and F2-score, it was optimized using grid search for parameter tuning.

From there the data was scrutinized with feature importance. Although

the option of feature scaling was contemplated, the utilization of

AdaBoost and Random Forest classifiers for binary classification

demonstrated resilience to fluctuations in feature scales. As a result,

explicit scaling was deemed unnecessary in the analytical process. The

models underwent additional iterations to evaluate the potential for

improvements.

In addition to accuracy, I also evaluated the classifiers based on the

F2-score metric. The F2-score considers both precision and recall, with

a higher weight given to recall. It measures the overall effectiveness of

the classifier in selecting top gymnasts while minimizing false negatives.

The findings indicate that, across various gymnastics events, both the

Random Forest and AdaBoost classifiers consistently demonstrated

high accuracy. Moreover, both classifiers consistently attained elevated

F2-scores, emphasizing their proficiency in accurately identifying top

performers.

In summary, the Random Forest and AdaBoost classifiers consistently

demonstrated high accuracy and F2-scores in the selection process for

Team USA gymnasts. These findings highlight the effectiveness of

these classifiers in accurately identifying the top gymnasts while

minimizing false negatives. The results provide valuable insights for

improving the selection process and ensuring the inclusion of the most

deserving gymnasts in Team USA.

5

4.3. Individuals likely to Medal

The gymnasts identified in the preceding models are deemed probable

contenders for securing medals on behalf of Team USA. It is essential

to note that the models were designed with a primary focus on

minimizing false negatives rather than false positives. Consequently,

there exists a likelihood that certain gymnasts identified in the following

list may not necessarily be destined to secure a medal.

4.3.1. Women

Apparatus Gymnast

Balance Beam Simone Biles

Balance Beam Joscelyn Roberson

Vault Simone Biles

Uneven Bars Shilese Jones

Uneven Bars Zoe Miller

IAA Jordan Chiles

Floor Simone Biles

Floor Ashlee Sullivan

4.3.2. Men

Apparatus Gymnast

Parallel Bars Yul Moldauer

4.4. Team Evaluation and Results

Initially, the algorithms were implemented with data sets containing all

American women or men from the data scraped from international

events. This resulted in optimized teams for the men and women. The

implementation of individual event modeling led to the inclusion of

gymnasts who were not part of the initially selected teams. This

necessitated the adoption of different approaches for forming the Men's

and Women's teams.

4.4.1. Men’s Team Evaluation and Results

The top-performing algorithms for men’s team selection were

Tabu Search and Variable Neighborhood Search (VNS),

consistently favoring Individual All-around (IAA) data sets. Tabu

Search, particularly with IAA Scores, emerged as the

highest-scoring algorithm.

6

(Note that the y-axis does not initiate at zero for improved visibility of nuances in high-scoring

algorithmic predictions.)

The initial phase involves evaluating the likelihood of Team USA

securing a medal in the team competition and identifying the

gymnasts pivotal to the team's success. Subsequent stages of

analysis focus on pinpointing individual gymnasts with the

highest probability of medaling in specific individual events. The

investigation identified one gymnast expected to excel in

individual apparatus events, and this gymnast was already

included in our team selection.

The algorithms for team selection were extended to other

nations showcasing convincing performance for team USA in

the qualification round for the 2024 Olympics. This extension

provides a valuable framework for predicting scores among

top-performing nations, aiding in the selection of gymnasts for

Team USA. The main goal was to optimize the individual medal

potential of gymnasts while ensuring the overall success of the

team.

The men's team selection process began by the algorithms

choosing four athletes— Khoi Young, Asher Hong, Fred Richard,

and Yul Moldauer—for the qualification round, yielding a

combined score of approximately 260.579, which should pass

the qualifying rounds to the finals. The only gymnast with the

potential to medal in individual events is Moldauer, who is

already selected for the team. From here we have a space for

Colt Walker, who boasted the second highest IAA score among

Americans at 85.00. The team of Walker, Moldauer, and

Richards could potentially score 254.992 in the final round.

These scores were then compared with other top

7

Olympic-qualified countries, including Japan, China, the United

Kingdom, Switzerland, and Germany.

(Germany and Switzerland scores were below 245, and not in the following chart)

(Note that the y-axis does not begin at zero to emphasize variations in the upper range

of scores.)

Above we can see that Team USA should potentially medal in

the 2024 Olympics.

4.4.2. Women’s Team Evaluation and Results

The comparative evaluation of diverse algorithms across the IAA

and Apparatus data sets provided nuanced insights into their

respective performances in finding the highest scoring USA

team. Tabu Search and Variable Neighborhood Search (VNS)

algorithms demonstrated the highest scores, with the advantage

leaning towards data sets using IAA scores. Specifically, the

VNS algorithm stood out as the top-performer.

8

(Note that the y-axis does not initiate at zero to provide a more detailed view of the performance

distinctions between the algorithms.)

The initial step involves determining the likelihood of Team USA

medaling in the team competition and identifying the gymnasts

for the team. Subsequent stages include identifying individual

gymnasts most likely to medal on specific apparatus. Analysis

revealed several gymnasts likely to medal on individual

apparatus, some of whom may not align with the predicted Team

USA lineup.

The team selection algorithms were applied to other countries

that performed well in team qualification for the 2024 Olympics,

providing a guideline for estimating the scores of top-performing

nations. This information assisted in the selection of gymnasts

for Team USA, enhancing the potential for individual medals

while ensuring the overall success of the team.\

The initially optimized women's team was examined with the

goal of achieving a medal. Further data exploration and

modeling were conducted due to the consideration of both the

team's likelihood of winning a medal and the potential for six

different women to possibly earn individual medals. To maximize

medals, the women's team should include Simone Biles,

predicted to win three individual medals, as well as four of the

five remaining women predicted to potentially medal.

The algorithms were employed to select four gymnasts from the

six potential medalists for qualifying. Team scores were

compared with those of other top countries: Great Britain, China,

Brazil, France, and Canada. In multiple scenarios, Simone Biles

and Shilese Jones were consistently chosen, with Jordan Chiles

being selected seven times. The graph below illustrates Team

9

USA is likely to pass the qualifying round easily, regardless of

which two of the remaining three gymnasts are chosen.

Consequently, Team USA should consist of Simone Biles,

Shilese Jones, and Jordan Chiles, who should also compete in

the final round, along with any two of Ashlee Sullivan, Joscelyn

Roberson, or Zoe Miller. In the chart below the error bar for

Team USA illustrates the range of scores associated with

different team compositions.

(Note: x-axis does not commence at zero for a more detailed representation of differences in

performance among nations.)

The Women’s Team is predicted to have a phenomenal 2024

Olympics with a potential gold medal in the team competition.

Below is a graph showing their predicted score in the finals

compared to other top nations. The significance is heightened

when considering that the scores for the listed countries

represent the top scores of each gymnast from the past two

years, while Team USA's scores are derived from the mean

values of their performances.

10

(Note that the y-axis does not begin at zero for a closer examination of score differences among

top countries.)

5. Conclusion

5.1. Findings

The exhaustive exploration of data scraping, cleaning, exploratory data

analysis (EDA), and modeling has unveiled significant insights into the

performance dynamics of both Men's and Women's USA Olympic

Artistic Gymnastics teams for the upcoming 2024 Paris Olympics. My

operational definition of "best," emphasizing the maximization of the

total medal count, has guided a meticulous investigation across

individual apparatus outcomes, team compositions, and strategic

positioning for Team USA.

In terms of individual events, a diverse collection of machine learning

models, encompassing Random Forest Classifier, AdaBoost Classifier,

Support Vector Classifier, among others, showcased nuanced

performances across various gymnastic disciplines. Results

underscored the effectiveness of the Random Forest and AdaBoost

classifiers, consistently displaying high accuracy and F2-scores,

affirming their robustness in the intricate process of selecting

top-performing gymnasts. These models demonstrated both reliability in

identifying winners and adaptability to handle inherent imbalances

within gymnastics datasets.

On the men's side, the optimization algorithms, including Tabu Search,

Randomized Heuristic, and Variable Neighborhood Search, contributed

valuable insights into the complexities of team composition. Results

indicated that Tabu Search and Variable Neighborhood Search (VNS)

algorithms, particularly when utilizing Individual All-Around (IAA) scores,

emerged as the highest-scoring algorithms. These findings provide a

comprehensive understanding of the intricate decision-making

11

processes involved in the selection of gymnasts for the Men's Team

USA.

Meanwhile, the evaluation of the Women's Team revealed a potential

gymnastics powerhouse for the 2024 Olympics, projecting an

impressive total of eight medals. This includes both individual accolades

and team achievements. The combination of individual gymnast

predictions and team dynamics highlighted the multidimensional nature

of gymnastics performance, showcasing the prowess of athletes like

Simone Biles, Shilese Jones, and Jordan Chiles, among others.

5.2. Limitations

It is imperative to acknowledge the inherent constraints and limitations

that accompany the study. This section aims to transparently specify the

boundaries and potential sources of bias within the analysis, providing a

nuanced perspective on the scope and applicability of the findings.

This study does not consider the impact of injuries, or the physical toll

associated with participating in numerous events within a condensed

time frame. Omitting these factors may limit the comprehensive

understanding of athletes' capabilities, as injuries and physical fatigue

can significantly influence performance outcomes.

The complexity of team sports, such as gymnastics, extends beyond

individual performances. This analysis does not account for the

dynamics of team cohesion, which plays a crucial role in team-based

events.

The study employed historical data for Men's and Women's gymnastics,

with Women's models trained on a dataset spanning from 2013 to 2021

and Men's models trained on a dataset from 2018 to 2020. Including

data from the 2016 and 2020 Olympics provided a comprehensive

testing ground. However, the discrepancy, resulting from time

constraints, in the time span of data collection for Men's and Women's

models should not introduce bias, but is noteworthy to mention.

5.3. Future Research

Conducting a more fine-grained analysis of individual features that

contribute to gymnastic performance could refine the models.

Investigating specific skills, techniques, or routine components and their

impact on outcomes could lead to more targeted and actionable insights

for athletes and coaches. Having a data set not only with the athlete’s

name, scores, penalties, and dates, but also which skills were

performed and at what point in their performance, as well as noting

combinations of skills and in which order.

12