Maximizing Medal Potential: A Multi-Method Data-Driven Model for US Gymnastics Team Selection

A Project for UCSAS 2024 USOPC DATA CHALLENGE | Team: Trojans Fighting From USC | Shilong Ren

A Data-Driven Model for Team Selection in Paris Olympic Gymnastics

1.Introduction: Problem and Motivation

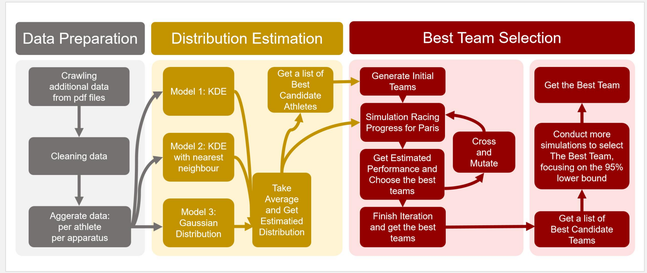
In the UCSAS 2024 USOPC Data Challenge, we developed a model to select the most promising US men's and women's gymnastics teams for the 2024 Paris Olympics, using competition data from 2017-2023 to optimize their overall medal performance.

Why Do We Combine Data-Driven Methods in Team Selection?

- Objective and Transparent Selection: Data-driven models ensure objectivity, fairness, and transparency in athlete evaluation and team selection.
- Balanced Team Creation: Data analysis identifies strengths and weaknesses across events, facilitating well-balanced team composition and maximizing medal potential.

2.1.Execution Overview

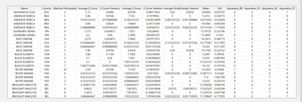
First, we constructed probability densities for individual athletes' scores on each apparatus. Next, we input these into a simulation to predict the expected medal performance for various team configurations and identify the team with optimal performance.



2.2.Data Preprocessing

Initially, we constructed a Python PDF crawler to source additional gymnastics data from the Tokyo Olympics at '*https://gymnasticsresults.com/results/2021/olympics/*'.

We cleaned and aggregated the data to obtain 'per athlete per apparatus' insights and preserved its average and variance.

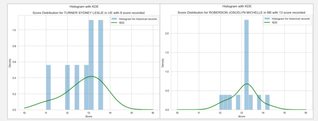


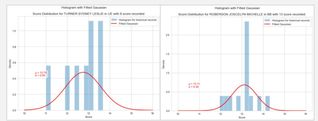
An Innovative Integration of Randmo Foest, XGboost Gaussian Distrubution, KDE, KNN and Genetic Algorithm for Simulation

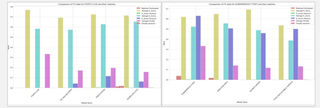
2.3.Distribution Estimation

We estimated probability distributions using scores from all previous competitions, generating and averaging three density curves per athlete/apparatus combination to balance predictions and account for varying data availability.

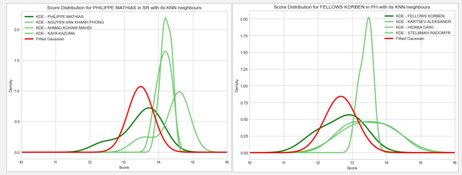
- **KDE (Kernel Density Estimation)** uses historical scores to show an athlete's score distribution for each apparatus; single scores are adjusted by adding ±0.25 for KDE fitting.
- **Gaussian Distribution** needs at least two scores per athlete/apparatus. To ensure fairness, we predicted the mean and standard deviation for all athletes. We tested Linear Regression, Random Forest, and XGBoost using Per-Athlete-Apparatus data, and selected the most accurate model.
- **KNN-inspired approach** employs variables as before to identify the three athletes most similar to each target across apparatuses, then averages their KDE curves.







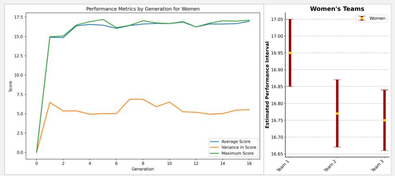
Athlete scores are generated by resampling and averaging, using as much data as possible to avoid bias and ensure fair outcomes. Hence, we complete our distribution estimation.



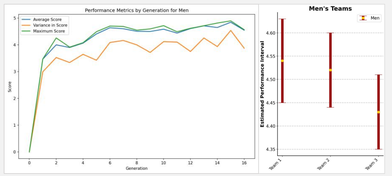
2.4.Genetic Algorithm & Simulation

Utilizing estimated probability distributions for each athlete and apparatus, we selected male and female teams to maximize the expected medal count, based on data from the Tokyo Olympics and incorporating adjustments for recent rule changes.

Simulations and a Genetic Algorithm efficiently explored team combinations, identifying the highest-scoring team, as evaluating all options was impractically extensive.



Selected Best Team (Women):
Biles Simone, Lee Sunisa, Lincoln Kaliya, McClain Konnor, Skinner Mykayla.



Selected Best Team (Men):
Hale Dallas, Mikulak Samuel, Richard Frederick Nathaniel, Wiskus Shane, Yoder Alec.

An Approach Ensuring Selection Fairness with Further Applications

3.Validity Verification

Our model was based on three fundamental assumptions:

- An athlete's performance could be partially predicted from their historical record and followed a distribution.
- Over 2-3 years, an athlete's performance tends to remain relatively stable.
- The overall competitor level in the Olympics remains stable, regardless of the countries and athletes participating.

To maintain fairness, we applied the same distribution estimation model to all athletes, emphasizing relative rankings over absolute scores for reliable performance comparisons. A large scale of data was used to predict individual athlete performance, minimizing bias from limited score records

A genetic algorithm conducted initial team selection, then identified the best team based on the lower bound of their predicted medal score's confidence interval. We then confirmed the current condition of each selected athlete and found no infeasible factors.

4.Further Contributions

Innovative Gymnastics Team Selection Techniques:
Our model, blending KDE, Gaussian models, and KNN, subtly enhances methods for gymnastics team selection and athlete evaluation, potentially contributing to the field of sports analytics.

Gymnastics Lineups Optimization:
Our application of Genetic Algorithms in gymnastics team selection offers insights for lineup optimization, potentially impacting sports strategy and planning.

Sports Analytics Educational Tool:
Our Python code, integrating efficient parallel processing, serves as a potential educational tool in sports analytics and coding practices. It highlights how parallel processing can optimize data science tasks in sports analytics, providing students with a practical implementation example.

Github & Google Drive Links:

Code & Poster

Full Report