Case 2: UCSAS 2024 USOPC Data Challenge

Sean Li, Christopher Tsai, Benjamin Thorpe, Jerry Xin

2023-11-06

Introduction

The sport of artistic gymnastics has long been a source of national pride and international recognition for the United States. The breathtaking performances, grace, and unparalleled athleticism displayed by American gymnasts have not only brought numerous Olympic medals home but also served as an inspiration for generations of aspiring athletes. As we approach the next Olympic Games in Paris 2024, the need for innovation and precision in athlete selection has never been more crucial. Our goal is to predict the best five-member lineups for both the USA Men's and Women's artistic gymnastics teams, optimizing for total medal count.

Historically, this selection process has relied heavily on the expertise and intuition of coaches and selection committees. However, in the era of data-driven decision-making, the role of predictive analytics and machine learning in the world of sports has gained unprecedented significance. The study will be of upmost significance, giving insight into lineup strategies in gymnastics and helping the USA maximize their medal count potential for the 2024 Paris Olympics.

While some prior studies have explored the use of data analytics in sports, they often fall short in the context of artistic gymnastics. Many of these studies are limited in scope and do not consider the nuanced aspects of gymnastics performance. Also, we plan to address the fact that some apparatuses (like pommel horse) are more difficult to judge by taking these higher variances in score into account when optimizing our team (Guston, 2023) . We will explore how to mathematically account for this in our analysis.

This study aims to overcome these limitations by analyzing athlete-specific data and incorporating domain expertise from the gymnastics community. It will offer a comprehensive and customized approach that addresses the unique competition format presented by artistic gymnastics and provides the USA Olympics Committee with a robust tool to make informed and data-driven decisions.

Methodology

Treating this as an optimization problem where we are maximizing for total medal count, we have developed nuanced heuristics based off analyzing past Olympic gymnastic event results to pick our top five-person lineups. Historically, we've seen that the USA, along with most other participating countries, send a combination of all-round gymnasts and event specialists in a five person lineup. Specifically, the USA Olympic Team Selection Procedure outlines that the top 2 overall all-round gymnasts will automatically qualify for the squad and then the committee will select the other 3 to complement them. A team can qualify a maximum of two individual gymnasts for each apparatus final as well as the all-around final. By this logic, it makes the most sense to send 2 all-around participants, while then having 3 specialists that would give that country the best shot at medaling in each individual apparatus. We wanted to mimic this approach, leading us to implement a two-tiered approach. First, we choose the top two overall all-round gymnasts, and then we select three more gymnasts to round out the team in an attempt to maximize medal count.

All Round Selection Method

In selecting the two all-round gymnasts, we will choose the two candidates that have the simulated highest total score average across all the events they have participated in leading up to the 2024 Olympics. Specifically, we weighted each score by artificially determining its importance using competition date and "stage" and averaged the weighted scores across all the events a US gymnast has participated in. We then sum the average weighted scores in each apparatus to find a composite all-around score.

We chose to put weights on competition date and competition stage, specifically, because we believe a heavier weight should be placed on athletes with a history of performing better under pressure in the final rounds, and more emphasis on athletes which have great recent performances. This weighting is our attempt to measure the "clutch" factor and peak/prime performance of an individual athlete. The way our weights will work is we will duplicate rows of individual scores in the dataset given the assigned weight of that score. Competition date will be split into three categories by how recent the competition was: very recent (< 6 months), somewhat recent (6 - 18 months), not recent (18+ months). A multiplier of 1 will be given to not recent competitions, a multiplier of 2 to somewhat recent competitions and a multiplier of 3 for very recent competitions. As for the competition stage, there are two stages of competition: qualifiers and finals. Qualifying round scores will receive a multiplier of 1 and final round scores will receive a multiplier of 2. For example, a row containing a score from a not recent competition and qualifying round will remain as one singular row. However, a row containing a qualifying round score, but from very recent, will be multiplied three times to get three rows containing that score. If that score was instead from the finals, it would be multiplied 2 x 3 = 6 times.

Taking the average of an individual athlete's scores after this duplication process will give us a mean expected score on each apparatus for that athlete. Still, we are aware that each gymnast will have variance in their scores for each apparatus and we wanted to take that into account. For each USA gymnast, we will also develop an individual total score distribution from their weighted scores to account for the variance for each individual on each apparatus. A density plot of all the scores after the duplication process will give us a distribution centered around the mean score for an athlete as well as a variance.

Because the main goal of analysis is to determine the probability that an individual athlete medals, we also had to determine a way to get medal probabilities. We will accomplish this goal by determining medal thresholds, or thresholds necessary to achieve a certain medal. This will first require us to get a distribution of scores on individual apparatuses. We do this by creating a distribution from the top 24 individuals' weighted mean score. Of the 24 people, however, no more than 4 competitors from the same country will be included in this distribution as a maximum of 4 individuals from each country can compete in qualifying for individual apparatuses. We then will sample from this distribution to get medal cutoffs.

In order to find an individual's probability of earning a medal we will use a two step simulation and sampling process. We will simulate the following process 10,000 times. We first will sample from the distribution of the top 24 individuals 24 times. The top three scores of the 24 will represent the medal cutoffs with the 3rd highest sample score corresponding to a Bronze medal, the 2nd highest sample score corresponding to a Silver medal cutoff, and the highest sample score corresponding to a Gold medal cutoff. Because we are only worried about the placing of gymnasts from the USA, we will sample once from each of the top 8 USA gymnasts in each individual apparatus given their unique distributions. If that gymnast's sampled score surpasses one of the cutoffs they will have 'earned' that medal for the given simulation. For example, if when sampling from the 24 gymnasts distribution we get medal cutoffs of 14.5, 14.8, and 14.9 and gymnast A gets a sampled score of 14.6 they will "earn" a Bronze medal. A sampled score of 14.81 would "earn" a Silver medal and a sampled score of 14.91 would "earn" a Gold medal. The only caveat is if multiple American gymnasts fall within the threshold for a certain medal. In that case the gymnasts would be ordered and the top score earns that medal and the next highest score would earn the next highest medal if it is available. We again simulate this two-tiered sampling approach 10,000 times to get each individual gymnast's probability of getting a specific medal in a specific apparatus. We will use this approach for the individual all-around, as well, and the two women with the highest medal probabilities will be automatically selected for our team.

From there we will use optimization to determine the 5-person lineup that maximizes expected medals earned including the team all-around medal. In determining the team all-around medal we will use a system

identical to the way the Olympics will work in 2024. We will use many permutations of 3 specialists along with our two guaranteed all-around gymnasts and sample from every gymnast's individual distributions to get all-around scores. For each apparatus, the top three US gymnasts on the teams distributions will be sampled from (because only the top 3 gymnasts participate in the team all-around finals). The scores will be aggregated to get a total team all-around score. The same method of medal cutoffs will be used in the team all-around medal cutoffs, but using empirical team all-around score data to create a distribution to be sampled from. The 5-person team that maximizes total medals earned will be the team we recommend the USA send to Paris for the 2024 Olympics.

#CHRIS EDITING STOPPED HERE

Specialist Selection Method

For the remaining spots, we again choose the three specialist gymnasts based on a simulation of each apparatus. We first obtain the top 24 gymnasts for a specific event per gender by aggregating a weighted score of the most recent results. Then we similarly make individual (USA athletes) distributions and grouped (all athletes) distributions per event and run the same procedure to pick the best US athletes that have the highest probability of medaling per event.

Now, we must take into account the weaknesses and strengths of the USA gymnastics team against the competition. For example, if Japan / China have specialists that are dominant in vault and balance beam, our total medals are strategically lower if we choose to place our specialists in those events as well. Instead, we should optimize for events where it is relatively easier to place. [not sure how this will statistically work, maybe just observationally and anecdotally examining the events]

Computing a Weighted Average for Gymnastics Data

$$\text{Weighted Average} = \frac{\sum_{i=1}^{n} w_i \cdot s_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

where:

- n is the total number of scores,

- w_i is the weight assigned to the i-th score, and

- s_i is the actual score of the gymnast for the i-th competition.

In this case, you can set the weights based on whether the score is from a final round(1.5) or not(1), plus the recency of the score, where very recent (< 6 months), somewhat recent (6 - 18 months), not recent (18+ months), are weighted by a multiplier of 2, 1.5, and 1, respectively. The explanation of these weights are offered in the paragraph below. For example, if $w_i = 3.5$ for a final round score that is very recent, and $w_i = 2$ for a non-final round score that is not recent, the formula becomes:

$$\text{Weighted Average} = \frac{\sum_{i=1}^{n} (w_i \cdot s_i)}{\sum_{i=1}^{n} w_i} \tag{2}$$

This will give more importance to the scores from the final round, and scores that are more recent, in the calculation of the average.

Weighted Average Score Distributions

For a final round performance, we multiply the score earned by 1.5. The reasoning behind this weighting is that we believe a heavier weight should be placed on athletes with a history of performing better under pressure in the final rounds, since the goal of this study is to achieve the most medals, which are only awarded to the top three performers in each event on an individual level. We then classify a performance by how recent the competition was, compared to January 1st, 2024, into 3 categories: very recent (< 6 months), somewhat recent (6 - 18 months), not recent (18+ months). These three groups are weighted by a multiplier of 2, 1.5, and 1, respectively. It's important to take into account if a gymnast is at a peak performance era and simultaneously look out for younger gymnasts on the rise, which is why we incorporated this step into our selection process. We then generate each USA gymnast's distribution assuming a normal distribution of performances. This assumption is justified, as when we look at the top 10 gymnasts regardless of country across apparatuses, we see they individually have a performance score distribution that is approximately normally distributed. To obtain a probability of medaling, we find a threshold of a weighted average score that it usually takes to medal on that apparatus from compiling past results. This threshold is adjusted higher or lower based on who we predict other top countries might send.

Results

Discussion

Appendix

Works Cited

Gunston, Jo. "Judging the Judges – How Statistical Analysis Evaluates Fairness And . . . " Olympics.Com, 19 Oct. 2023, olympics.com/en/news/how-statistical-analysis-evaluates-fairness-accuracy-gymnastics. Accessed 06 Nov. 2023.