UCSAS 2024 DATA CHALLENGE REPORT

Nick Grener, Yash Vora, Adam Zabner

Team Name: The Uneven X-Bars (Graduate Division)

OVERALL STRATEGY

Choosing a team of 5 athletes to represent the United States at Paris 2024 is a complex problem. The goal of this project was to use the past performance of athletes to model which team of 5 would maximize the expected medal count of the men's and women's teams. This process brought up interesting questions. Should the team be made up of specialists in individual apparatus or should the United States pick the best all-around athletes? Given the mixture of individual and team events, who is best suited to help the program achieve the goal of winning the most medals? Are there competitive advantages that can be discovered vis a vis opponent countries? While our project did not answer every one of these questions fully, we picked teams that we believe will maximize the number of American medals in 2024.

The complexities of gymnastics and its scoring were new to us, and it took quite a bit of research for us to feel comfortable understanding the competition. After much discussion and background research into the history of international artistic gymnastics, our team decided to build a model that puts simplicity, flexibility, and interpretability over other characteristics. We recognize that sacrifices inevitably have to be made when attempting to solve a problem as multifaceted as the one under consideration. Given that the full slate of international competitors has not yet been determined for Paris, we chose to focus solely on the U.S. athletes and use historical benchmarks in each event to decide which combination of team members provided the best chance at winning medals. Placing ourselves in the hypothetical role of decision-maker for the program, we realized that the decisions that other countries would be making regarding their own teams and lineups would be information that we could only speculate about until the moments immediately preceding the games. Therefore, we zeroed in on choosing the lineups that gave the team the best shot of reaching medal-level scores in the most events possible. By disregarding the recent scores of athletes from other countries, we are neglecting to answer the aforementioned question regarding competitive advantages, but we believe that the simplicity (both conceptual and computational) of our model justifies this loss.

IMPLEMENTATION

We chose to use justified benchmarks to determine who was a medal-caliber athlete in each event. This necessitated some research. The spreadsheet entitled "Scaling Benchmarks" shows the results of that research with citations. Since the structure of international competitions has been in flux over recent decades, some normalization of scores was required. Using only scores from Olympic competitions and World Championships (the highest levels of competition), we determined, at each event & medal type, the multiplicative factor for each 4-year cycle with respect to the current 2022-23 cycle by finding the geometric mean of the ratios of all (arithmetically averaged) medal scores within said cycle with respect to their analogous (arithmetically averaged) medal score in the current cycle. This normalizing factor was then applied to the past medal scores to bring all of the data we were using into present-day units that matched the scoring that will be used in Paris.

Another issue that needed to be addressed before our model could be constructed was data preparation. To ensure that each athlete was attached to the appropriate scores in the dataset, we created a standardized "Full Name" column in the dataset, which eliminated any spaces at the beginning of the "First Name" and "Last Name" fields, converted all letters to lowercase, and concatenated those results. This did not resolve all of the issues with attaching the correct entries to each real-world athlete, though. To ensure accurate extraction of each athlete's full name information, an optimal approach was employed; this involved utilizing the Tesseract OCR engine on the image-based PDFs to verify the presence of full names within the documents. There were a total of 6 athletes in the dataset whose full names were not found on these PDFs. Furthermore, we noted that "fred richard" and "frederick richard" were still listed as unique athletes, even though they represented the same individual. Inconsistencies also arose in certain athlete entries where a middle name was present in one entry but absent in another, leading to other mismatches. After manually correcting these inconsistencies, we conducted a final sanity check- an automated process was set up using custom code to cross-reference each athlete's name on the web. This method ensured the authenticity of every athlete included in our cleaned dataset. Another aspect of data cleaning we conducted was to treat all vault performances as an independent occurrence, so all "VT1" and "VT2" events were re-coded to simply "VT".

To build the model itself, we first decided to reduce the computational complexity of our algorithm by filtering out American athletes in the dataset who we did not feel merited serious consideration for inclusion on the team. We sorted the US female athletes by average score and best score in every event, and if a given athlete was not amongst the top 8 average scores or top 8 best scores in any event, we dropped them from the dataset. The idea here was that if a woman's best scores in every event and average scores in every event were never among the top eight in the country's program, they were not strong enough to fill a coveted slot on a 5-member team. (Note that for the men, given the larger number of athletes and larger number of events, we changed the criterion to "top 6", which resulted in a filtered dataset that was still larger than the filtered women's dataset.)

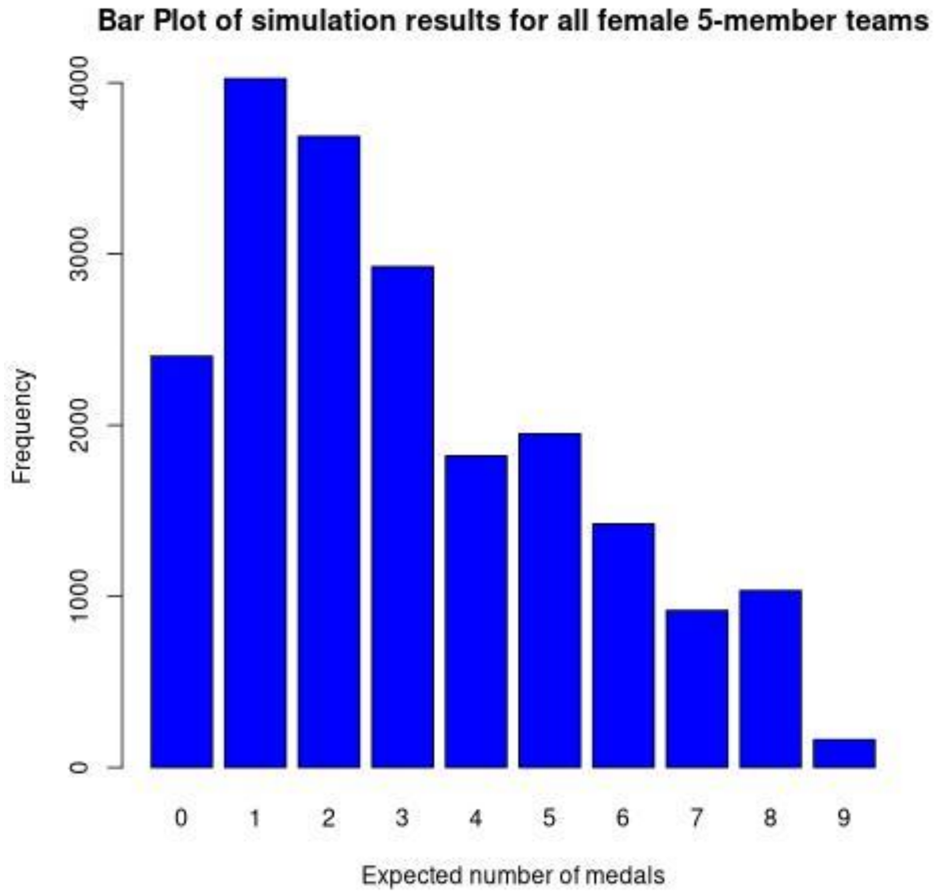
Next, we used the combn() function to create a data frame of all unique 5 name teams that could be formed from this filtered subset of legitimate contenders. For the women, this still resulted in a data frame with approximately 20,000 combinations; the men's data frame was even larger. However, this still represented a considerable savings in the computations described below, as the difference between 20 choose 5 (the approximate number of rows in the filtered dataset) and 50 choose 5 (the approximate number of rows if we had considered all possible 5 name teams using the every name provided in the cleaned data set) is a factor of about 100.

The core of our program is a "medal estimator" function that takes as arguments the names of athletes, their average score in every event, and the benchmarks for every level of medal in each event. First, this function fills out a hypothetical lineup of which four gymnasts will compete in qualifying in each event based on their sorted historical average scores in each event. Next, because we are making the assumption that our best guess for each athlete's performance in Paris will be their historical average, the function passes on the top two performers in each event to the finals and use if-else statements to check which medals (if any) they could be expected to win by comparing their averages to the benchmarks. We chose to by-pass the checking of scores against qualifying benchmarks in each event, because what we are really interested in is whether the athletes were going to reach the medal benchmarks, and there was no harm in passing somebody whose score did not reach the qualifying benchmark on to the finals- in any case, a check against some benchmark that the athlete failed to reach was going to be needed, and the coding was much easier to do all the checking against reference marks in the finals.

Finally, this function sums up the entire lineup's expected scores in all qualifying events to use as a tiebreaker should teams tie for the number of medal benchmarks that were met. This was deemed to be the best simple measure of a given combination's overall quality as a potential representative team.

RESULTS

A clear-cut top team choice did not emerge for either the women's or the men's teams, but our model was able to identify a preferred team based on the tie-breaker mentioned above. On the women's side, quite a few teams were projected to earn nine total medals as seen in the bar plot below:

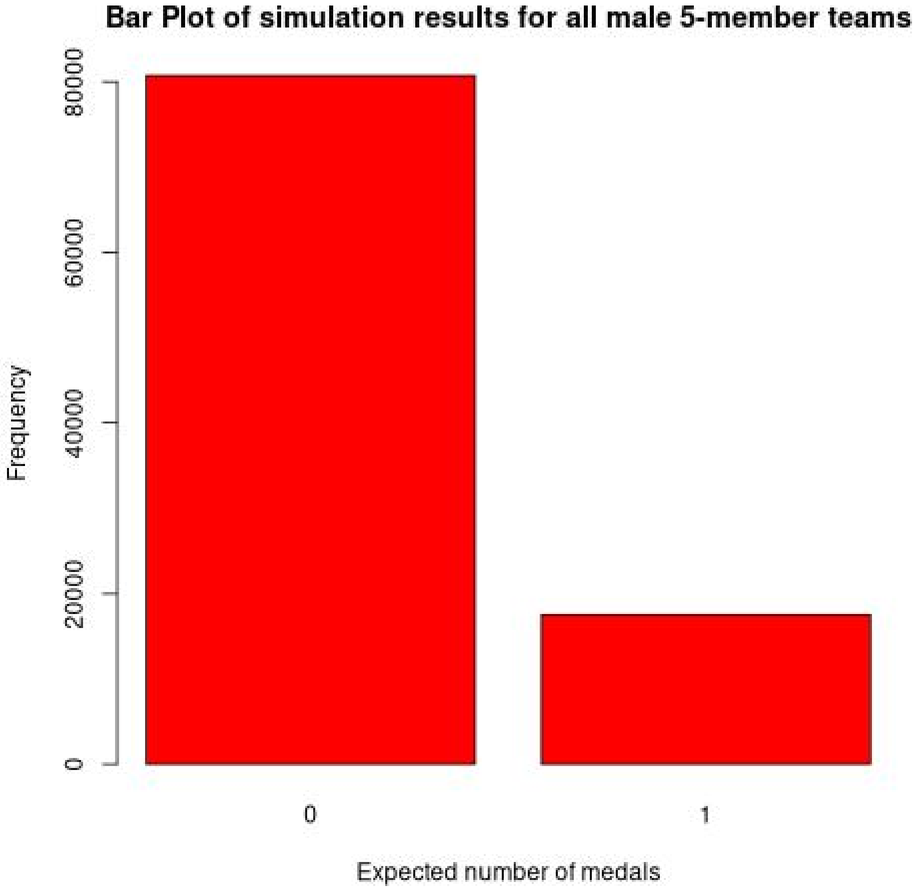


The team with the best projected overall score consists of Simone Biles, Shilese Jones, Konnor McClain, Leanne Wong, and Kaliya Lincoln. This represents a team made up of strong all-around athletes rather than apparatus specialists. As seen in the table below, replacing Leanne Wong with Jade Carey may be preferable if the goal is to earn more higher-value medals, as the resulting team would be projected to win more silver medals according to our model.

1	2	3	4	5	Gold	Silver	Bronze	Total Score
Simone Biles	Shilese Jones	Konnor McClain	Leanne Wong	Kaliya Lincoln	5	1	3	226.3442
Simone Biles	Shilese Jones	Konnor McClain	Skye Blakely	Kaliya Lincoln	5	1	3	226.2204
Simone Biles	Shilese Jones	Konnor McClain	Jordan Chiles	Kaliya Lincoln	5	1	3	225.7087
Simone Biles	Shilese Jones	Konnor McClain	Jade Carey	Kaliya Lincoln	5	2	2	225.5260
Simone Biles	Shilese Jones	Konnor McClain	Kayla Dicello	Kaliya Lincoln	5	1	3	225.4358
Simone Biles	Konnor McClain	Leanne Wong	Skye Blakely	Kaliya Lincoln	5	1	3	225.3929
Simone Biles	Shilese Jones	Konnor McClain	Nola Matthews	Kaliya Lincoln	5	1	3	224.9971
Simone Biles	Konnor McClain	Jordan Chiles	Skye Blakely	Kaliya Lincoln	5	1	3	224.7574
Simone Biles	Shilese Jones	Konnor McClain	Kaliya Lincoln	Zoe Miller	5	1	3	224.7372

Simone Biles	Shilese Jones	Konnor McClain	Tiana Sumanasekera	Kaliya Lincoln	5	1	3	224.6733

On the men’s side, the selection of a team is based on the combination the provides the best chance to medal in the team all-around, as no individual competitor reached the bronze benchmark in any particular event, although a few athletes did come close (notably, Khoi Young in vault, Stephen Nedoroscik in pommel horse, and Paul Juda in floor exercises). The combination of athletes that achieved the highest overall total score is Brody Malone, Paul Juda, Vitaliy Guimaraes, Yul Moldauer, and Colt Walker, although the difference between the overall score for the top team and the tenth-best team is comparatively small, as seen in the table below:



1	2	3	4	5	Gold	Silver	Bronze	Total Score
Brody Malone	Paul Juda	Vitaliy Guimaraes	Yul Moldauer	Colt Walker	0	0	1	338.3997
Brody Malone	Paul Juda	Vitaliy Guimaraes	Colt Walker	Asher Hong	0	0	1	338.2530
Brody Malone	Paul Juda	Vitaliy Guimaraes	Colt Walker	Khoi Young	0	0	1	337.8693
Brody Malone	Paul Juda	Vitaliy Guimaraes	Shane Wiskus	Colt Walker	0	0	1	337.8458

Brody Malone	Paul Juda	Yul Moldauer	Colt Walker	Khoi Young	0	0	1	337.7072
Brody Malone	Paul Juda	Colt Walker	Khoi Young	Asher Hong	0	0	1	337.5605
Brody Malone	Vitaliy Guimaraes	Yul Moldauer	Colt Walker	Khoi Young	0	0	1	337.4773
Brody Malone	Paul Juda	Vitaliy Guimaraes	Colt Walker	Stephen Nedoroscik	0	0	1	337.3679
Brody Malone	Paul Juda	Shane Wiskus	Yul Moldauer	Colt Walker	0	0	1	337.2860
Brody Malone	Vitaliy Guimaraes	Shane Wiskus	Yul Moldauer	Colt Walker	0	0	1	337.2748

MODEL STRENGTHS

We feel that the most attractive qualities of our model are its simplicity and flexibility. For the input data, we chose to use the entire dataset of scores from prominent competitions over the last two years that we were provided with. However, if a user decided that scores from only the past, say, 6 months were important to consider, a simple filter of the input data would provide an updated recommendation in a very short amount of time. Or, if a coach had access to scores from practice/training sessions that they felt were a reliable reflection of each athlete's expected score in each event, they could simply feed the model a list of all the athlete + event + expected score data from their training sessions, and the model would give a clear suggestion for the competitors to send to Paris. By relying on fixed, justified thresholds that need to be reached in defining "medal-worthy" status, there is no need to track what competing programs are doing- it's simply a matter of letting the model find the optimal combination of performances within the team to reach the most thresholds possible.

A related strength of our approach is the model's universality. That is, it could be used equally effectively by any country's gymnastics program to determine their own optimal team, as long as their program had trustworthy input data regarding their own athletes' performance in each event. We feel that the model embodies the common coach's mantra of not worrying about who the competition is- success in athletics is first and foremost a matter of putting your best foot forward and maximizing your own performance.

One other advantage of our model is its interpretability- a layperson can understand the idea of an average and how that can be used to compare people in a given domain, and we'd like to believe that anybody with a minimal amount of programming experience could read through our code and associated comments and understand how the model is working. In our experience, we have too often encountered "original" or "clever" techniques that the designer was at a complete loss to justify, let alone clearly explain to an interested party.

CRITIQUES AND REPLIES

The most obvious critique we could make of our model was that it lacked validation. Though we were provided with all of the results from the 2021 games in Tokyo, we were not given the input data from all of the major competitions in the cycle leading up to those games. Obtaining that data and ensuring its completeness and accuracy was deemed too time-consuming of a task; instead, we chose to use the past performance of the US team as a check on the legitimacy of our model. Going back through 1988 (the 1984 games were anomalous with many strong gymnastics programs boycotting for political reasons), the US women's team has averaged 4.9

medals at Olympic competitions, with a maximum of 9 medals in 2016, while the men’s team has averaged 1.2 medals, with their maximum being 3 medals (also in 2016). So while our model may be slightly sanguine in its predictions for the women’s team in the upcoming games, the predictions are not unreasonable when compared to recent results at the same level.

One could also argue that our approach oversimplifies the data by flattening each athlete’s historical results in a particular event to a single value. We seriously considered setting up a large number of simulations when running the medal estimation function on each lineup of 5 athletes: in this scenario, we would bootstrap from each athlete’s vector of past scores on each iteration, and then average out the medals earned for the lineup over all of the simulations. While this approach may have allowed us to separate teams’ potential performances more clearly, we decided it was not worth the vast increase in program run time that this would have required. The paucity of data for many athletes in many events (in examining our master dataset, it was uncommon to find athletes who had more than a half-dozen scores in an event) also caused us to question the need to bootstrap.

Related to this, our tie-breaking procedure feels somewhat arbitrary, and when examined closely, does not offer clear separation between the top handful of options on either side. We would hope that subjective considerations of experts who are working closely with the athletes in question could be used to select the best team based on qualitative factors that our model - as well as any comparable model - would have a difficult time including quantitatively, foremost of which is the health of the athletes in question.

One more weakness of our model worth acknowledging is that we did not factor athlete fatigue into our medal estimation procedure for a given lineup; we simply put the best performer (as measured by historical average score in the event) into the lineup in every situation. Truthfully, we did not feel experienced enough in the realm of high-level athletic performance to estimate the marginal effect of fatigue for those athletes who were being called on to participate in a large number of events.

CONCLUSION

Our team enjoyed the complexity of this challenge and we each individually feel that we learned something about sports analytics, a domain that none of us had any prior experience in. The process of researching the changes that have taken place in the sport of artistic gymnastics over the years gave us an appreciation for the evolution of the competition, and increased our enthusiasm for following the storylines of this program in the coming months. Thank you for providing an interesting data challenge, and for your thoughtful consideration of our entry.