# Final Report

**Team name:** YSAL

**Captain:** Min Sung Choi (alstjd122067@yonsei.ac.kr)

**Team members:** Hee Yeon Jeon (heeyeon930@yonsei.ac.kr), Hwa Yun Song (hwayun.song@yonsei.ackr)

## 1. Data

Data were collected from major domestic and international gymnastics competitions from the seasons leading up to the 2020 Tokyo and 2024 Paris Olympics and provided to contestants. Additional data was collected from the following source and used to clarify, standardize the location of competitions and nationality of athletes.

https://www.olympiandatabase.com/index.php?id=1670&L=1#google_vignette

## 2. Data Pre-Processing

Data pre-processing took the following path:

### Data Cleaning

The follow procedures were applied in data cleaning to address missing values and outliers.

- Athlete Names: Data with missing values or erroneous entries in names of athletes was corrected
- Penalty: Rows consisting of a penalty greater than 1 were deleted
- Score: Rows consisting of in either E_Score or D_Score were deleted
- Round Names: some scores were recorded repeatedly due to different round names, so round names were organized.
- Apparatus Names(VT, HB): uniform terminology to facilitate smoother analysis
- Addition of full name column(ex: SIMONE Biles)

### Host Advantage Test

Since aesthetics is an important criterion in judging artistic gymnastics, we presumed that home advantage could cause difference in scores.

- Using Nationality data, we revised terms in columns ‘Location’ and ‘Country’ so they could be compared
- The t-test findings show that there is a difference in the average scores of athletes from the host country and those from non-host countries
- Based on this result, we incorporated the variable 'Host' (1 for athletes from the host country and 0 otherwise) in the XGBoost model for score prediction.

### Weighted Mean

In order to reflect the recent form, we calculated the weighted average of D and E Scores for each athlete and apparatus, giving more weight to competitions closer to the Paris Olympics.

### Rank Contribution

Using the calculation method of WAR (Wins Above Replacement), we determined the contribution to D and E Ranks.

> For instance, if four athletes achieved 1st, 2nd, 3rd, 4th, 5th, 6th with the socre of 6, 5, 4, 3, 2, 1, respectively,
>
> the athlete in 1st place would have a contribution of 6-1/5 + 6-2/4 + 6-3/3 + 6-4/2 + 6-5/1,
>
> the one in 2nd place would have a contribution of 6-1/5 + 6-2/4 + 6-3/3 + 6-4/2,
>
> the one in 3rd place would have a contribution of 6-1/5 + 6-2/4 + 6-3/3,
>
> the one in 4th place would have a contribution of 6-1/5 + 6-2/4,
>
> the one in 5th place would have a contribution of 6-1/5,
>
> and the athlete in 6th place would have a contribution of 0.

## 3. Ridge Regression

To determine the contribution of athletes’ medal count for each competition, the regression approach method was employed. The total number of medals obtained by the United States in each competition would be the dependent variable, while the participation status of athletes (denoted as 0 or 1) was considered as the independent variable for conducting regression analysis. However, due to a

significantly lower number of data points (number of competitions) compared to the number of variables (number of U.S. athletes participating in one or more competitions), we experienced the issue of instability in the regression coefficients. We decided to use ridge regression in order to solve this problem.

- Creating the Design Matrix:

  The design matrix consisted of the names of U.S. athletes in each column while the last column showed the total medal count for each competition. A design matrix was generated where each row represented a competition, and a binary value (1 or 0) indicated whether a particular athlete participated in that competition. The total medal count for the corresponding competition was appended to the last column.

- The 10-fold cross-validation method was employed to select the optimal ridge regression parameter(lambda).

Through ridge regression, regression coefficients (beta) were obtained for each athlete, representing how many additional medals the United States could potentially secure if and when the particular athlete participated in the competition.

# 4. Score Prediction

## XGBoost

We utilized the XGBoost model to predict scores for athletes in the Paris Olympics. The D Score exhibits relatively low variability for each athlete, depending on the attempted skills. In contrast, the E Score reflects how well an athlete performed in the actual competition, resulting in significantly greater variability. Due to these distinct characteristics of D Score and E Score, we trained separate models for each.

The XGBoost model was selected because it showed the best performance among Randomforest, Multi Layer Perceptron method.

We included the following variables for score prediction: Gender, Country, Apparatus, Round, Weighted_Average_D, RankContribution, Host, Beta.

We then split the data into training and testing sets for both D Score and E Score, with random selection.

The prediction results showed a relatively high accuracy with a rmse=0.278 for D Score and rmse=0.60 for E Score.

# 5. Identify Best 5 Players

1. Team Medals

   After predicting the scores for each apparatus(for each athlete), the top 10 scoring athletes for each apparatus were nominated for participating in the 2024 Paris Olympics. All possible combinations of 5 athletes were made into a list, and each combination(team) was made into a dataframe which consists of scores of each apparatus for each athlete.

   The top 3 scores within the team for each apparatus were then added, and the sum of the corresponding scores for every apparatus became the team score. A higher team score indicates a higher possibility of winning the team medal.

   Woman

| | p1 | p2 | p3 | p4 | p5 | Score |
|---|---|---|---|---|---|---|
| 0 | BILES Simone | CAREY Jade | JONES Shilese | LEE Sunisa | LINCOLN Kaliya | 173.842777 |
| 0 | BILES Simone | CAREY Jade | DICELLO Kayla | JONES Shilese | LEE Sunisa | 173.551659 |
| 0 | BILES Simone | DICELLO Kayla | JONES Shilese | LEE Sunisa | LINCOLN Kaliya | 173.512294 |
| 0 | BILES Simone | CHILES Jordan | JONES Shilese | LEE Sunisa | LINCOLN Kaliya | 173.507164 |
| 0 | BILES Simone | CAREY Jade | JONES Shilese | LEE Sunisa | MCCLAIN Konnor | 173.483410 |

   Man

| | p1 | p2 | p3 | p4 | p5 | Score |
|---|---|---|---|---|---|---|
| 0 | HONG Asher | MALONE Brody | WISKUS Shane | YOUNG Khoi | PHILLIPS Curran | 257.149549 |
| 0 | HONG Asher | RICHARD Frederick Nathaniel | WISKUS Shane | YOUNG Khoi | PHILLIPS Curran | 256.936239 |
| 0 | MALONE Brody | WHITTENBURG Donnell | WISKUS Shane | YOUNG Khoi | PHILLIPS Curran | 256.738977 |
| 0 | LOOS Riley | MALONE Brody | WISKUS Shane | YOUNG Khoi | PHILLIPS Curran | 256.648462 |
| 0 | HONG Asher | MALONE Brody | RICHARD Frederick Nathaniel | YOUNG Khoi | PHILLIPS Curran | 256.622599 |

2. Individual Medals

   Individual medal prediction was based on the predicted scores. The top 3 scoring athletes for each apparatus were each given a medal.

| Athlete | Round | Apparatus | Score_Prediction | Medal_Prediction |
|---|---|---|---|---|
| CAREY Jade | AAfinal | BB | 13.513639 | 0 |
| CAREY Jade | final | BB | 12.898303 | 0 |
| CAREY Jade | AAfinal | FX | 14.020990 | 0 |
| CAREY Jade | final | FX | 13.946219 | 0 |
| CAREY Jade | AAfinal | UB | 13.112485 | 0 |
| CAREY Jade | AAfinal | VT | 14.438005 | 0 |
| CAREY Jade | final | VT | 14.550428 | 1 |
| LINCOLN Kaliya | AAfinal | BB | 13.351761 | 0 |
| LINCOLN Kaliya | final | BB | 13.200818 | 0 |
| LINCOLN Kaliya | AAfinal | FX | 14.178653 | 0 |
| LINCOLN Kaliya | final | FX | 14.280256 | 0 |
| LINCOLN Kaliya | AAfinal | UB | 13.193085 | 0 |
| JONES Shilese | AAfinal | BB | 13.731583 | 0 |
| JONES Shilese | final | BB | 13.446031 | 0 |
| JONES Shilese | AAfinal | FX | 13.636383 | 0 |
| JONES Shilese | final | FX | 13.785485 | 0 |
| JONES Shilese | AAfinal | UB | 14.903877 | 1 |
| JONES Shilese | final | UB | 14.480194 | 1 |
| JONES Shilese | AAfinal | VT | 14.176430 | 0 |
| BILES Simone | AAfinal | BB | 14.419993 | 0 |
| BILES Simone | final | BB | 14.193394 | 0 |
| BILES Simone | AAfinal | FX | 14.892567 | 1 |
| BILES Simone | final | FX | 14.545826 | 1 |
| BILES Simone | AAfinal | UB | 14.033108 | 0 |
| BILES Simone | final | UB | 14.317711 | 0 |
| BILES Simone | AAfinal | VT | 15.420900 | 1 |
| BILES Simone | final | VT | 13.999699 | 0 |
| LEE Sunisa | AAfinal | BB | 13.869656 | 0 |
| LEE Sunisa | final | BB | 13.841568 | 0 |
| LEE Sunisa | AAfinal | FX | 13.371685 | 0 |
| LEE Sunisa | AAfinal | UB | 14.652035 | 1 |
| LEE Sunisa | final | UB | 14.872469 | 1 |
| LEE Sunisa | AAfinal | VT | 14.562347 | 1 |

| Athlete | Round | Apparatus | Score_Prediction | Medal_Prediction |
|---|---|---|---|---|
| HONG Asher | AAfinal | FX | 14.062750 | 0 |
| HONG Asher | AAfinal | HB | 12.530790 | 0 |
| HONG Asher | AAfinal | PB | 14.290291 | 0 |
| HONG Asher | final | PB | 14.041542 | 0 |
| HONG Asher | AAfinal | PH | 12.706124 | 0 |
| HONG Asher | final | PH | 12.962471 | 0 |
| HONG Asher | AAfinal | SR | 14.072449 | 0 |
| HONG Asher | final | SR | 13.879915 | 0 |
| HONG Asher | AAfinal | VT | 14.178295 | 0 |
| HONG Asher | final | VT | 14.869057 | 1 |
| MALONE Brody | AAfinal | FX | 13.443098 | 0 |
| MALONE Brody | AAfinal | HB | 13.307528 | 0 |
| MALONE Brody | final | HB | 14.337938 | 0 |
| MALONE Brody | AAfinal | PB | 14.031466 | 0 |
| MALONE Brody | final | PB | 14.203604 | 0 |
| MALONE Brody | AAfinal | PH | 13.676336 | 0 |
| MALONE Brody | final | PH | 13.666485 | 0 |
| MALONE Brody | AAfinal | SR | 13.662375 | 0 |
| MALONE Brody | final | SR | 13.950905 | 0 |
| MALONE Brody | AAfinal | VT | 14.412266 | 0 |
| PHILLIPS Curran | AAfinal | HB | 14.172194 | 0 |
| PHILLIPS Curran | final | HB | 13.992470 | 0 |
| PHILLIPS Curran | AAfinal | PB | 14.299543 | 0 |
| PHILLIPS Curran | final | PB | 15.132078 | 1 |
| PHILLIPS Curran | AAfinal | VT | 14.625566 | 1 |
| YOUNG Khoi | AAfinal | FX | 14.447371 | 0 |
| YOUNG Khoi | AAfinal | HB | 12.322660 | 0 |
| YOUNG Khoi | AAfinal | PB | 14.746084 | 1 |
| YOUNG Khoi | AAfinal | PH | 13.602676 | 0 |
| YOUNG Khoi | final | PH | 13.674330 | 0 |
| YOUNG Khoi | AAfinal | SR | 13.290311 | 0 |
| YOUNG Khoi | AAfinal | VT | 14.972307 | 1 |
| YOUNG Khoi | final | VT | 15.114654 | 1 |
| WISKUS Shane | AAfinal | FX | 14.017761 | 0 |
| WISKUS Shane | AAfinal | HB | 14.041380 | 0 |
| WISKUS Shane | AAfinal | PB | 14.633082 | 1 |
| WISKUS Shane | AAfinal | PH | 13.508825 | 0 |
| WISKUS Shane | AAfinal | SR | 14.066788 | 0 |
| WISKUS Shane | AAfinal | VT | 14.409405 | 0 |

## Final Result

bases on the above result,

**woman team**:

Simon Biles, Jade Carey, Shilese Jones, Sunisa Lee, Kaliya Lincoln with 7 expected medals

**man team**:

Asher Hong, Brody Malone, Shane Wiskus, Khoi Young, Curran Philips with 6 expected medals