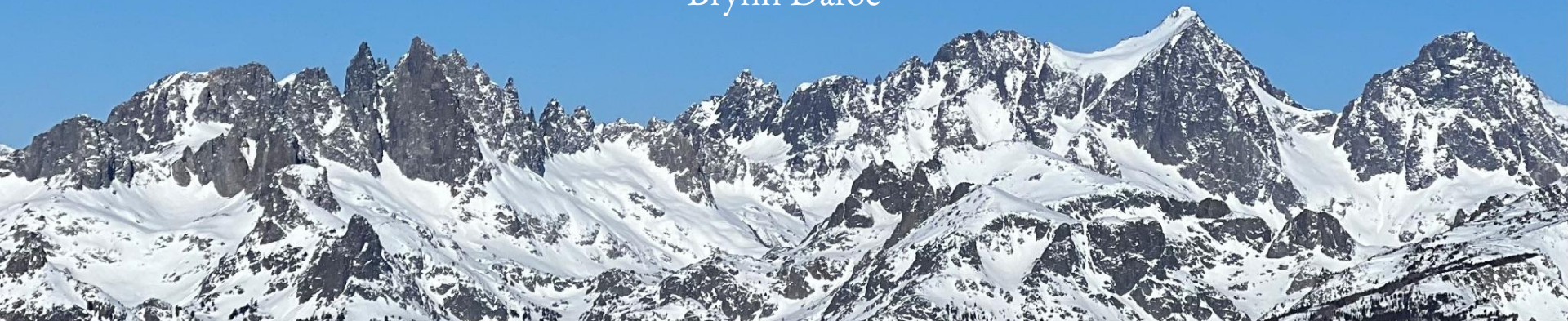
2026 Italy Winter Olympics: Predicting the Probability of Making the Top 5 in Men's and Women's Singles Moguls

Brynn Dafoe



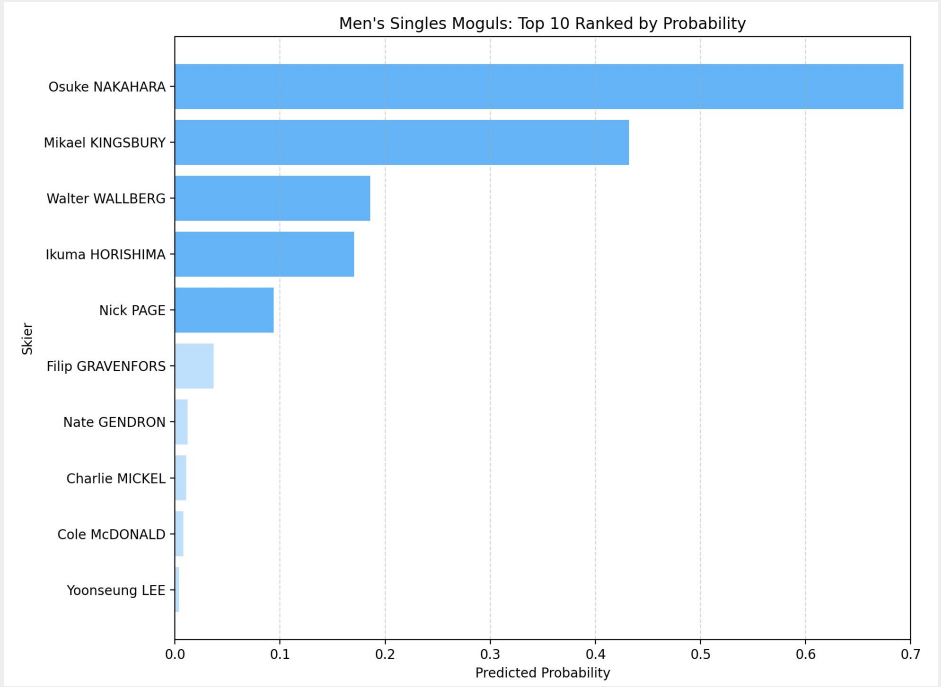
Introduction:

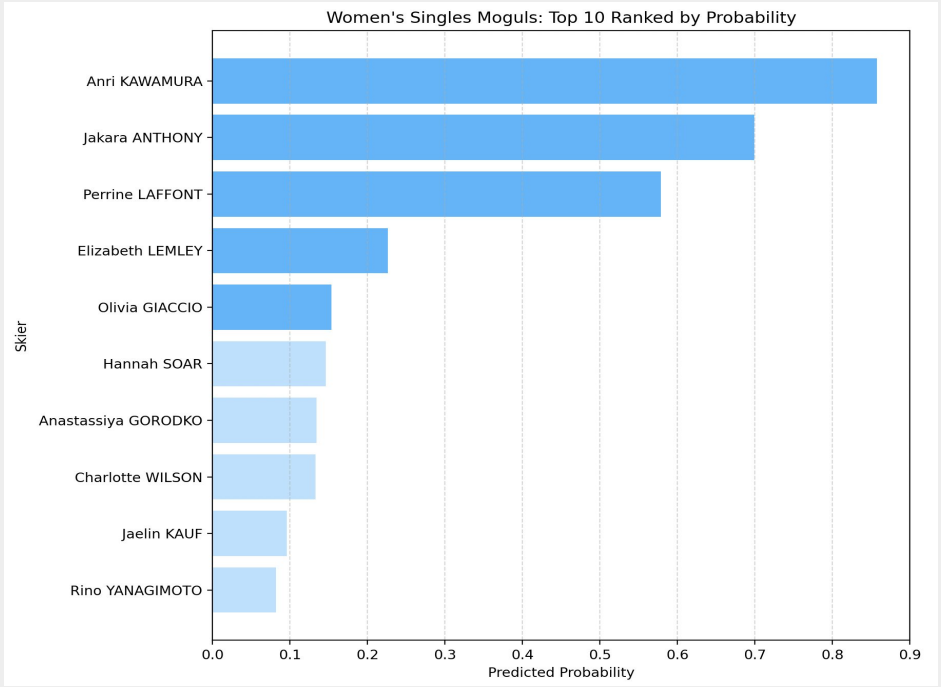
- This project uses logistic regression to predict the probabilities of making the top 5 in men's and women's singles moguls for the 2026 Winter Olympics. The model was trained on two previous Winter Olympic cycles: 2018 Pyeongchang and 2022 Beijing. Each cycle used the four seasons of World Cup events leading up to the Olympics as well as athlete biographies. The goal of the model was to assign probabilities to each athlete based on how likely they were to make the top 5, and also to see how different variables influence the outcome of a skier's performance at the Olympics by looking at the feature's weight.
- Mogul skiing is a discipline of skiing in which a skier skis down a course of bumps (moguls) and two jumps, the scoring of their performance being: 60% turns, 20% jumps, and 20% speed.

Data Sources:

Data Source:	Description:	Approach:	Size:
• FIS World Cup Singles Moguls Results	• Results from World Cup events from the 2015 season to the 2025 season. • Collected: Rank, FIS Code, Name, Nation, Birth Year, Final Score (out of 100), Time Points (out of 20), Air Points (out of 20), and Turn Points (out of 60).	• Used pdfplumber to scrape the data since the score sheets were only available via downloadable pdf files.	• Raw: 4678 • Cleaned: 3342
• Olympic Singles Mogul Results	• Results from the 2018 Pyeongchang Olympics and 2022 Beijing Olympics. • Collected: Rank, Name, and Country.	• Used json and regular expression to web scrape the data due to the data being in a json embedded portion of the html.	• Raw: 120 • Cleaned: 118
• FIS Athlete Biographies	• Data from the each athlete's biography page on the FIS website. • Collected: Name, FIS Code, Birth Year, Age, and Gender.	• Used BeautifulSoup to web scrape the data.	• Raw: 370 • Cleaned: 370

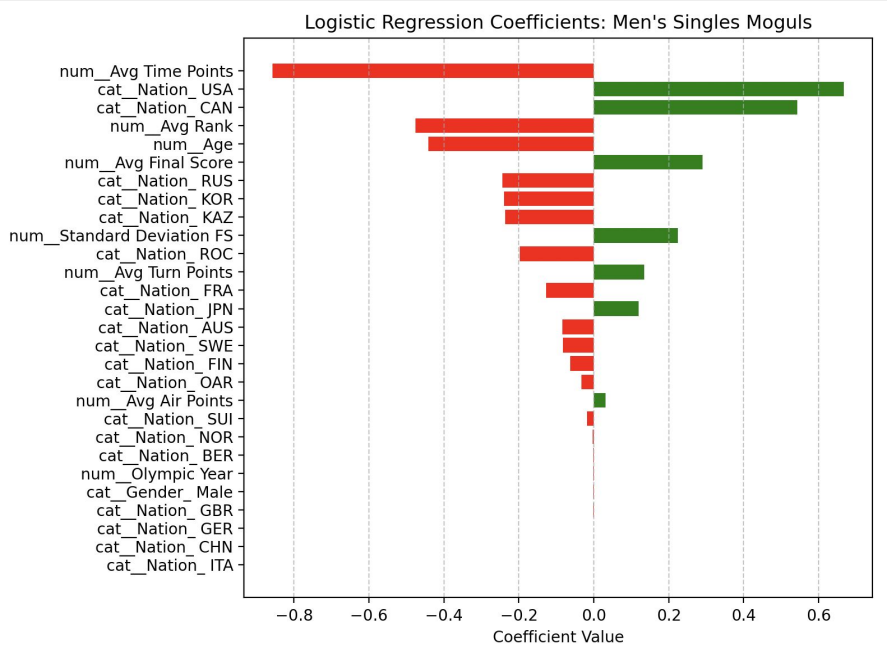
Summary of Results:

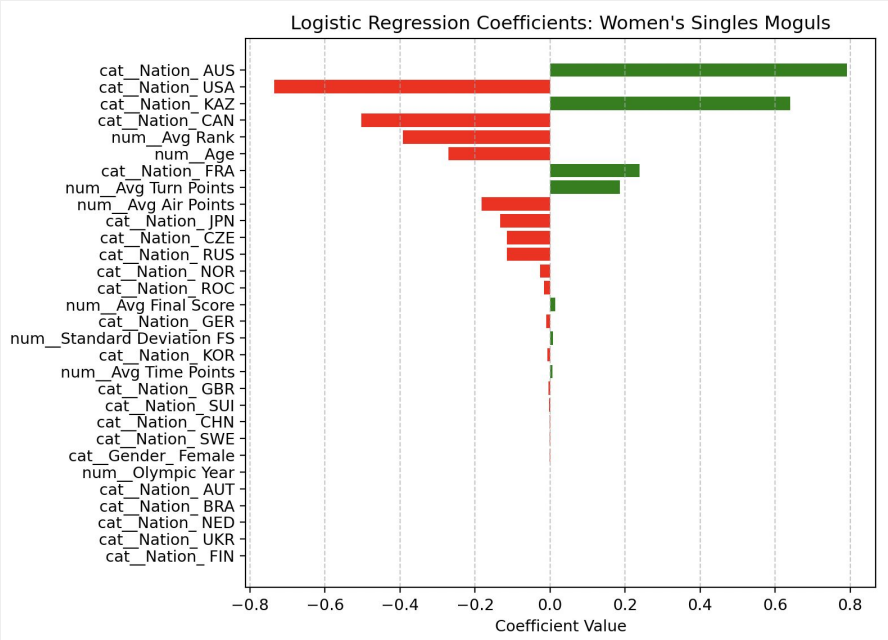




- These two charts show the top 10 men and top 10 women ranked by highest probability based on whether or not they will make the top 5 or not. Longer bars represent a stronger probability of making the top 5. To note: the skiers are ranked solely on the probability of making the top 5. It does not indicate their actual placement (like first, second, third, etc.).

Summary of Results:

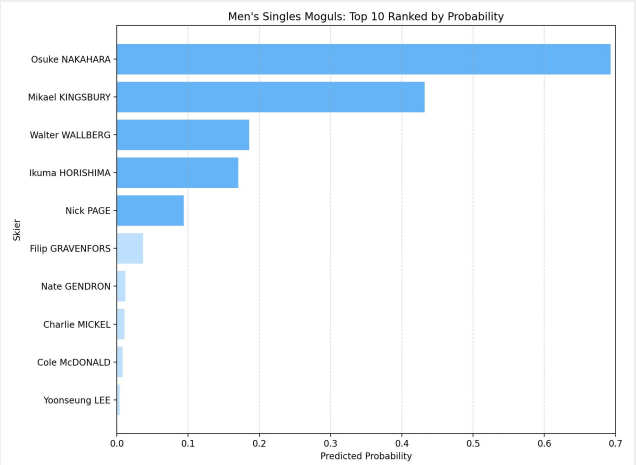


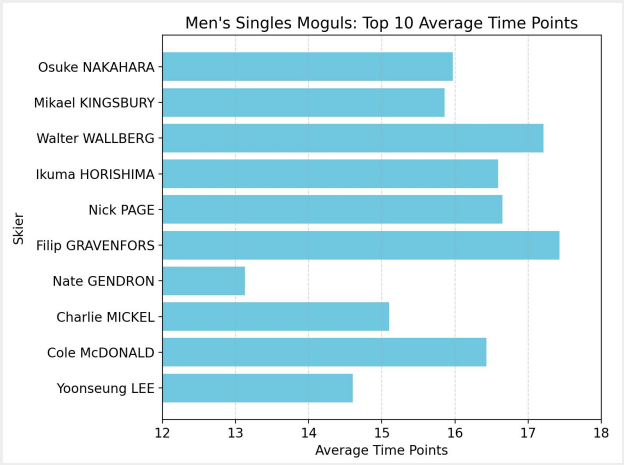


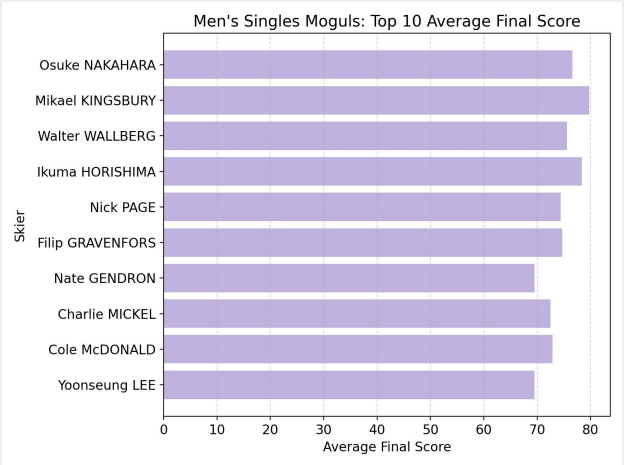
- The strongest numeric predictor for men is time points. It has a negative coefficient (about -0.85). Though the coefficient is negative, the actual relationship between average time points and the probability of making the top 5 is positive. Logistic regression assigns coefficients based on the effect the feature has while holding all other features constant. Because the model estimates coefficients while considering all features at the same time, correlations between features can influence the coefficient sign. Despite the coefficient sign being negative, the large magnitude indicates that it is an important predictor for predicting the probability of making the top 5.

- The strongest numeric predictor for women is rank. It has a negative coefficient (about -0.40). The negative sign here does match the expected relationship: the higher the average rank the lower the probability is of making the top 5.

Summary of Results:

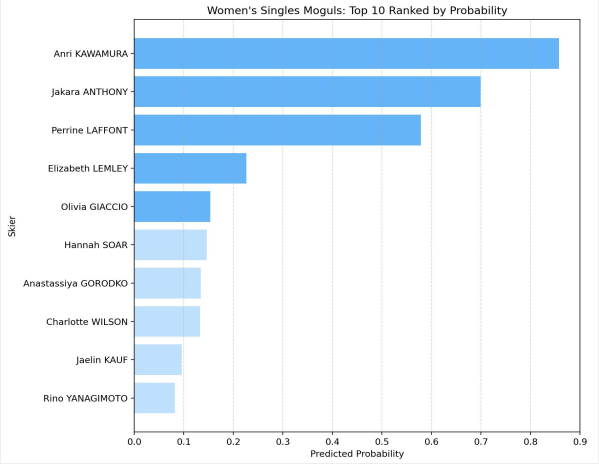


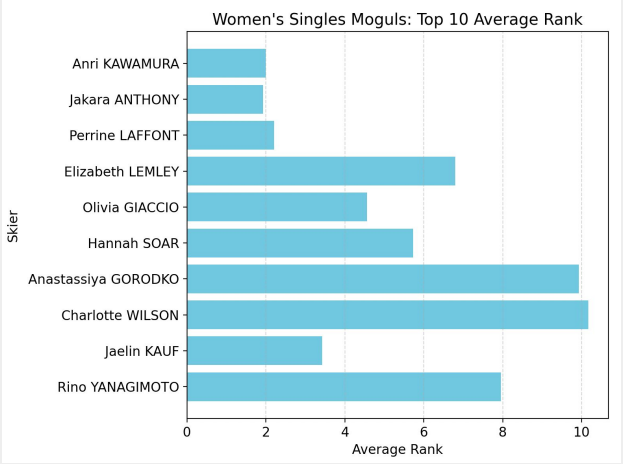


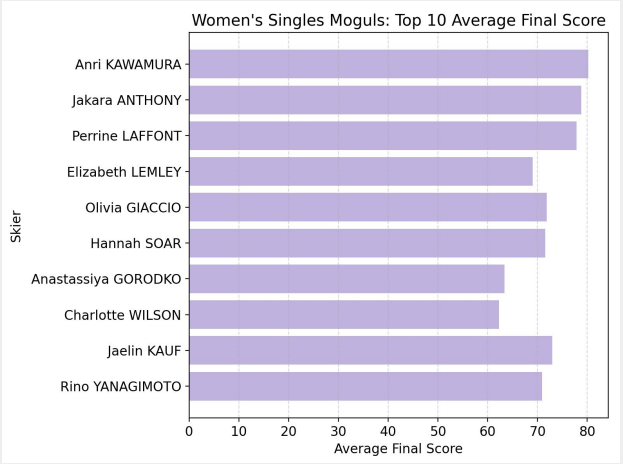


- The first chart shows the top 10 skiers ranked by probability as seen before, the second chart shows the average time points of those same skiers (the strongest numeric predictor for men), and the last chart shows the average final scores of those same skiers (the intuitive predictor). Although the average time points chart has some variability in it, it shows the same pattern as the probabilities chart, showing that: higher average time points (faster runs) corresponds with higher predicted probabilities. The average final score chart shows a more consistent decreasing pattern (though slight), showing that: high average final scores increases the probability of making the top 5.
- Average time points is the strongest numeric predictor for men, but the differences in time between these skiers is small. The skiers going to the Olympics are extremely elite, and their performances (turn wise and air wise) are very similar. Often it comes down to just a 1 or 2 second difference in run time to determine who wins. These small differences in time points makes the difference in determining the probability of making the top 5.

Summary of Results:







- The first chart shows the top 10 skiers ranked by probability as seen before, the second chart shows the average rank of those same skiers (the strongest numeric predictor for women), and the last chart shows the average final scores of those same skiers (the intuitive predictor). The average rank chart shows a more distinct pattern compared to the men's time points chart. This pattern matches the probabilities chart in that a higher average rank decreases the probability of making the top 5.

Challenges:

- The score sheets used to collect data on World Cup results had varying forms. 2015 to 2016 had a very different form compared to 2017 onwards. 2017 to 2020 had non-uniform errors throughout the data, but the results for 2021 onwards were (mostly) consistent and correct. My code could not read the format for 2015-2016 correctly so I made the csv files for these seasons manually rather than try to edit the code.
- All of my sources had data stored in different places. My initial challenge was identifying where the data I wanted was stored in each source. Each of my programs that collect data (World Cup Results, Olympic Results, and Athlete Data) have completely different scraping methods due to this.

Thank You!