IFSC Bouldering World Cup Winner Classification Report: Abridged Version

Tyler Chang

November 2022

Report Summary

In the upcoming Paris 2024 Olympic Games, competitive rock climbing will be an official event. This report documents the development of supervised machine learning models to predict whether a given rock climber should be classified as a contender to win an International Federation of Sport Climbing (IFSC) bouldering[1] world cup based data taken from 2018 and 2019 IFSC bouldering competitions[2]. Using a binary classification structure, a random forest model that is over 95% accurate is constructed. Given the similarity between the IFSC competition structure and that of the Olympic Games, being able to accurately predict the likely winners of IFSC events will help countries select their team members. The models covered here could also be implemented to set betting odds for IFSC or Olympics bouldering events.

This report is an abridged version of my earlier report on the same data set[3].

Outline of the Report

In the following section, **Methods and Analysis**, I cover importing, cleaning, and exploring the data set, as well as discuss my approach to modeling the data.

In the third section, **Results**, I discuss and compare the performances of the individual models. As will be shown, the random forest model outperforms the alternatives in both sets of models with an accuracy rate of approximately 94%.

Finally, in Section 4, **Conclusion**, I provide a brief summary of the report, discuss possible applications for the models and the limitations of the work done here. I also note options for future work and improvements.

Methods and Analysis

Cleaning and Validating the Data Set

The original data set, *boulder_results.csv*, has 5535 rows and 13 columns and contains the following columns:

```
##  [1] "Competition.Title" "Competition.Date"  "FIRST"
##  [4] "LAST"              "Nation"            "StartNr"
##  [7] "Rank"              "Qualification"     "Qualification.1"
## [10] "Qualification.2"   "Semifinal"         "Final"
## [13] "Category"
```

[1]It is assumed that readers know what bouldering is. If you are not familiar with the sport, I recommend visiting: https://www.nytimes.com/2022/04/22/well/move/fitness-bouldering-rock-climbing.html

[2]Source: https://www.kaggle.com/datasets/brkurzawa/ifsc-sport-climbing-competition-results

[3]Available at: https://github.com/Habeus-Crimpus/IFSC/blob/main/IFSC%20Classification%20Report.pdf

The columns' definitions are as follows:

1. *StartNr* is the order in which an athlete climbed in the qualification round.
2. *Rank* is an athletes ranking at a given competition.
3. For each of the three rounds (*Qualification*, *Semifinal*, and *Final*), the format of their score appear as *#T#Z##* and are interpreted as:

- The first number (before the T) is the number of climbs completed ('Tops').
- The second number (after the T and before the Z) is the number of zones reached.
- The final number is a combination of the number of attempts for both tops and zones. If there are only two digits, the first digit is the attempts for tops and the second digit is for zones. If there are three digits, the first digit is for tops and the latter two are for zones. If there are four digits, the first two digits are for tops and the latter two are for zones.

4. *FIRST* and *LAST* are the first and last names of the athletes.
5. *Nation* is the country for which an athlete competes.
6. *Competition.Title* and *Competition.Date* are self-explanatory.

I have made the following modifications to the data set[4].

1. Three columns have been removed since they were either empty or contained the same value for all rows.
2. Missing values were found StartNr and Rank columns. Since the missing values could not be determined, the associated rows have been removed.
3. The table includes columns for multiple rounds of competition. Since not every climber competes in every round, many of the rows had blank spaces. These have been amended to say 0T0Z00, indicating zero tops and zero zones.
4. I divided the data points that are currently formatted as 3T2z89 or similar into new numeric columns so that they could be better analyzed and visualized. This resulted in 16 new columns: 4 for total values and 4 for each of the 3 rounds.
5. The first and last names of the athletes have been combined into a single column and put into title case.
6. The original *Qualification*, *Semifinals*, and *Finals* columns became redundant and were removed.
7. A column *Winner_Contender* was added that marked whether a climber is considered a contender to win a competition. The determination was based on whether a climber completed at least one zone at any competition.

This left a table with 5498 rows and 23 columns, which was then checked for missing or incorrect values.

Analysis and Visualization

The average number of tops is 2.34 and the median is 1 out of a possible 12, meaning that the vast majority of climbers do not usually top most of the problems. The typical climber also does not reach the majority of zones , with the mean and median being 3.495 and 3, respectively. With the measures of center being quite low for both tops and zones, it follows that a good deal of climbers are separated in the ranks by their number of attempts. So, it is helpful to instead consider the ratio of average attempts to average successes for both tops and zones.

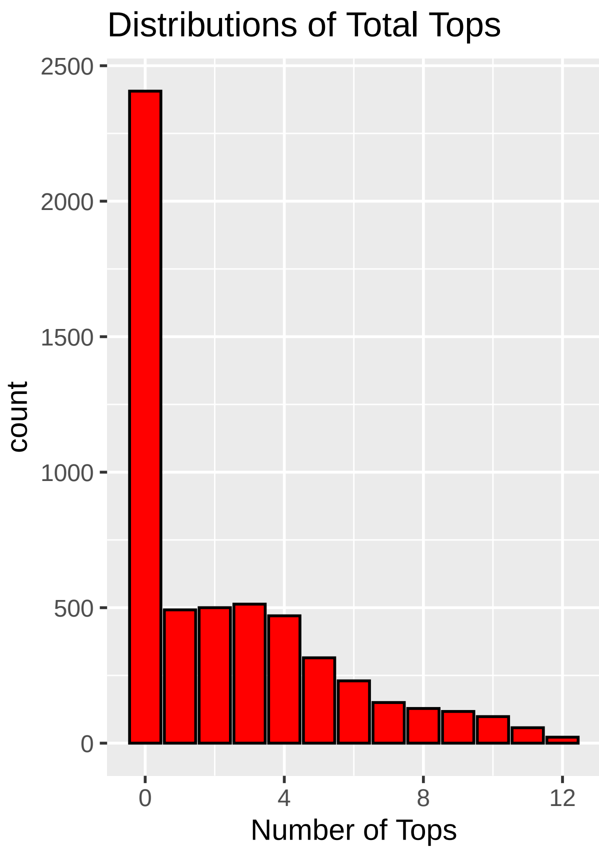
Category	Ratio
Overall Attempts to Top	1.822217
Overall Attempts to Zone	2.042306
Qualification Attempts to Top	1.711736
Qualification Attempts to Zone	1.926624
Semifinal Attempts to Top	2.338950
Semifinal Attempts to Zone	2.541230

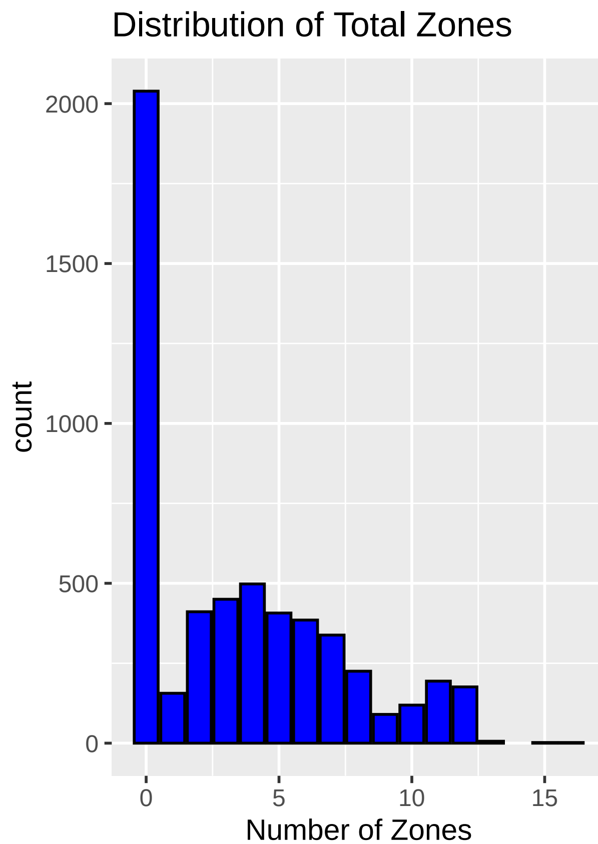
[4]For full details of the cleaning process, please see https://github.com/Habeus-Crimpus/IFSC/blob/main/IFSC%20Classification%20Report.pdf

Category	Ratio
Final Attempts to Top	2.261506
Final Attempts to Zone	2.271400

The qualification round is typically the easiest round and the semifinals is usually the hardest round. In the finals, the number of attempts needed to reach zone is very close to the number of attempts needed to top a problem, suggesting that the impact of zones on placement in the finals is likely less than in earlier rounds.

The vast majority of climbers did not top even a single problem and almost none topped all 12 problems, as shown by the plot below.





Approximately 43.8% of climbers never topped a problem and only 0.4% topped all the problems. There are also clusters around 1 to 4 and 7 to 9 tops. This is to be expected since those completing 1 to 4 problems are likely those who made it to semifinals but not finals, and those scoring 7 to 9 tops being those who qualified for finals.

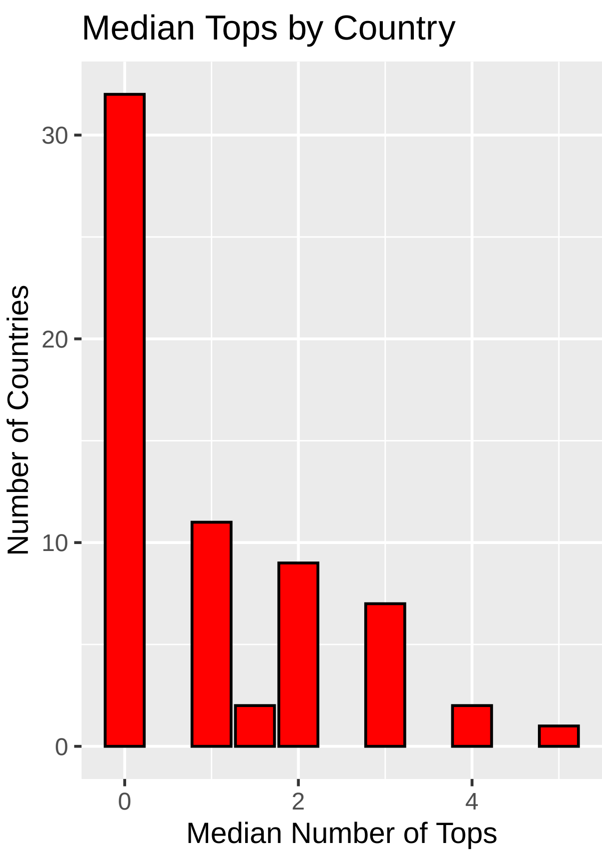
For zones, 0 is again the most common number at 37.1% and 16, the maximum number, being the least common at 0.0364%. This implies that the probability that a given climber has a perfect performance, meaning that all tops and zones are completed, are exceedingly low at approx. 0.0146%. Interestingly, 4 is the second most common number of zones at 9.06% and those with only a single top are less common than those with 0, 4, 3, 2, 5, 6, 7, 8, 11, or 12 zones. This suggests that reaching a high number of zones does not necessarily translate into a high number of tops.

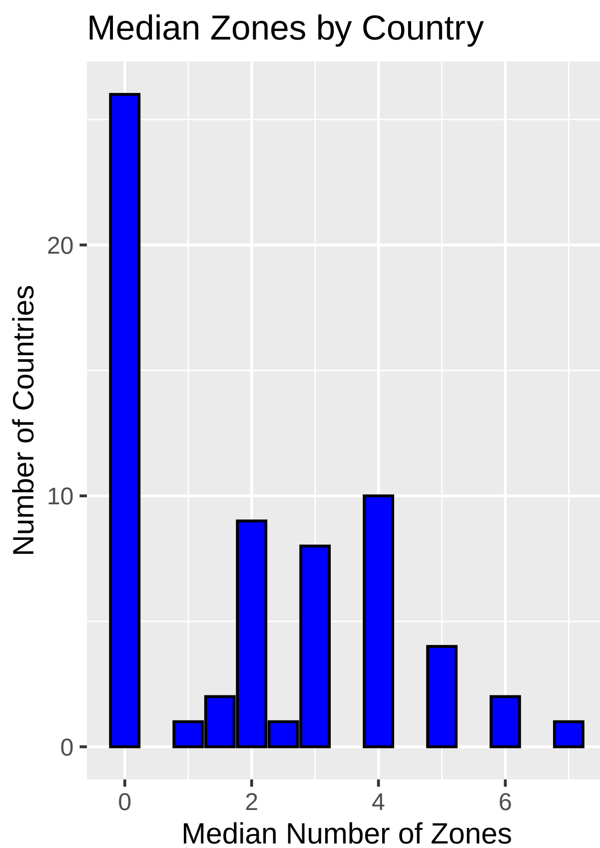
Not all countries have equal access to climbing facilities, natural formations, or financing for national-level teams. As a result, there may be nation-specific biases. To see this, we can look at the average and median numbers of tops and zones are for each country. To see how the two measures of center compare, consider:

```
## # A tibble: 6 x 5
```

```
##   Nation avg_tops avg_zones med_tops med_zones
##   <chr>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 FRA        4.89      6.47        5         7
## 2 SLO        4.06      5.53        4         6
## 3 AUT        4.02      5.52        4         6
## 4 ITA        3.65      5.09        3         5
## 5 BUL        3.56      5.12        3         4
## 6 LUX        3.38      4.23        2         5
```

We see that French, Slovenian, and Australian athletes are the only groups who averaged 4 or more tops per competition. Notably, France, the nation withe highest average tops, has a significant lead of approximately 0.83 over Slovenia, the second highest. This same trend and ordering holds for zones, where France again displays a significant advantage over even their closest rival. The averages and medians do tell somewhat varying stories, though neither provides a consistently higher or lower metric than the other.

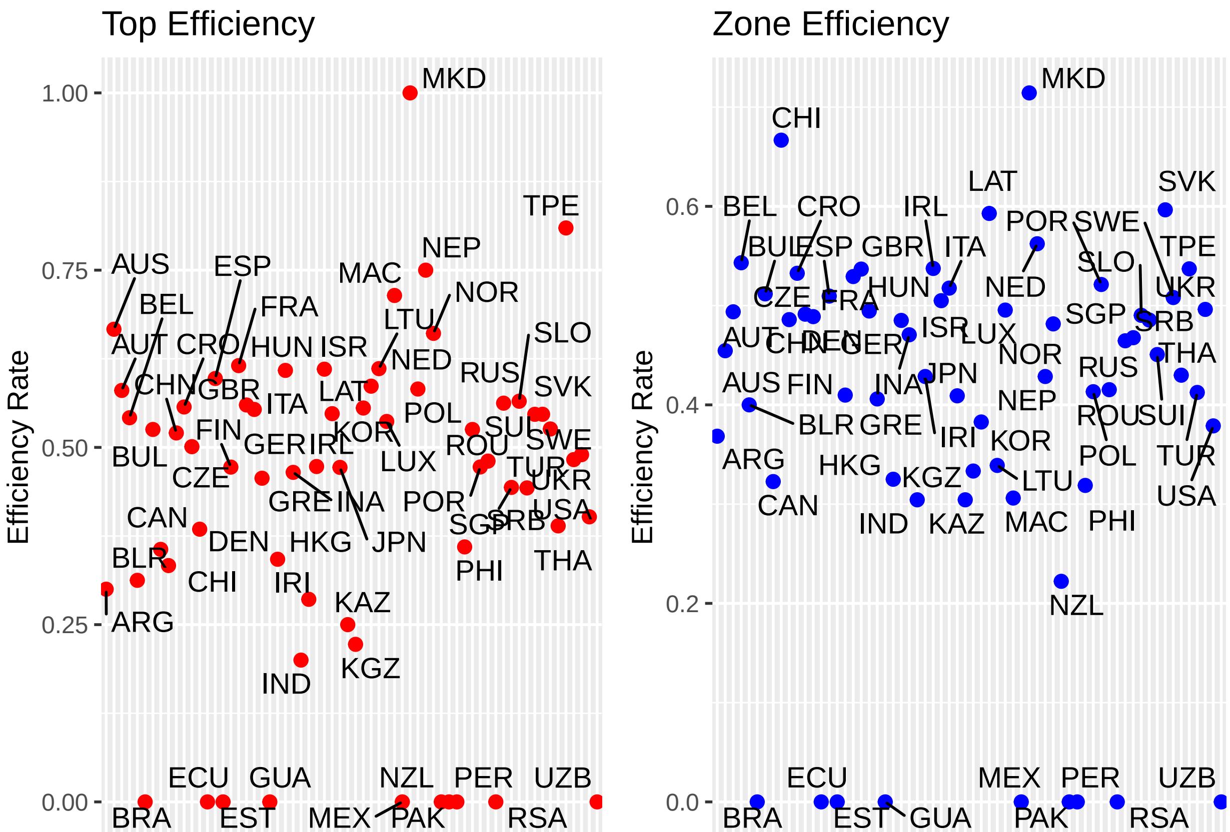




Another measurement of each country's typical performance is efficiency, which gives insight into how often each nation's athletes succeeded in getting a zone or top.

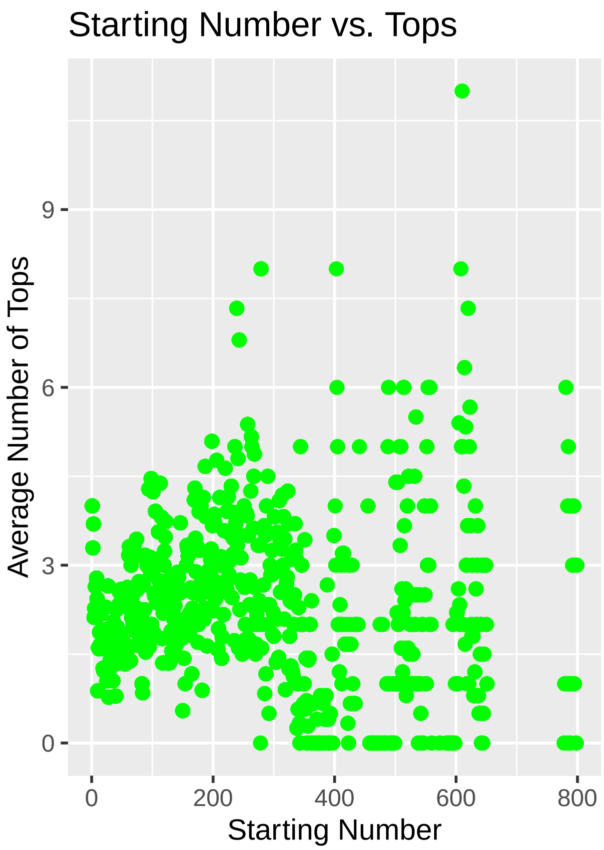
```
## # A tibble: 6 x 3
##   Nation top_eff zone_eff
##   <chr>    <dbl>    <dbl>
## 1 MKD      1        0.714
## 2 TPE      0.810    0.537
## 3 NEP      0.75     0.429
## 4 MAC      0.714    0.306
## 5 AUS      0.667    0.455
## 6 NOR      0.661    0.482
```

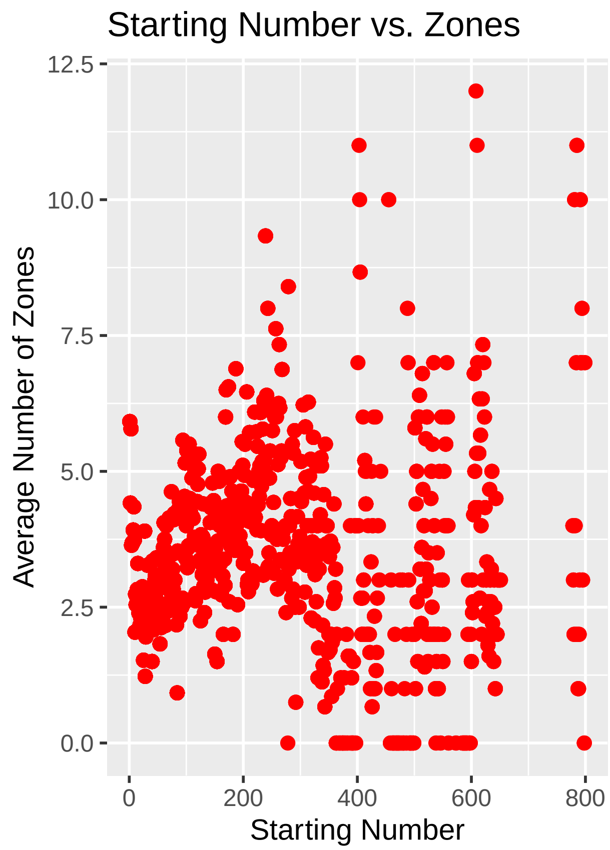
North Macedonia (MKD), Taiwan (TPE), Nepal (NEP), and Macao (MAC) all are more efficient than France, Norway, Austria, and other countries that placed much higher in terms of mean and median numbers of tops and zones. This is likely due to the current efficiency calculus not weighting efficiency by how many tops or zones a country typically receives. To see the whole distributions, consider the following plots:



With a small number of exceptions (Brazil, Ecuador, Guatemala, Pakistan, Uzbekistan, Mexico, Peru, South Africa, Estonia, and North Macedonia), the efficiency rates of most countries for both tops and zones is between 0.3 and 0.6. Notably, the countries that had the highest average and median tops and zones were not the most efficient countries; France, Austria, and Slovenia all landed around the middle for both efficiency plots.

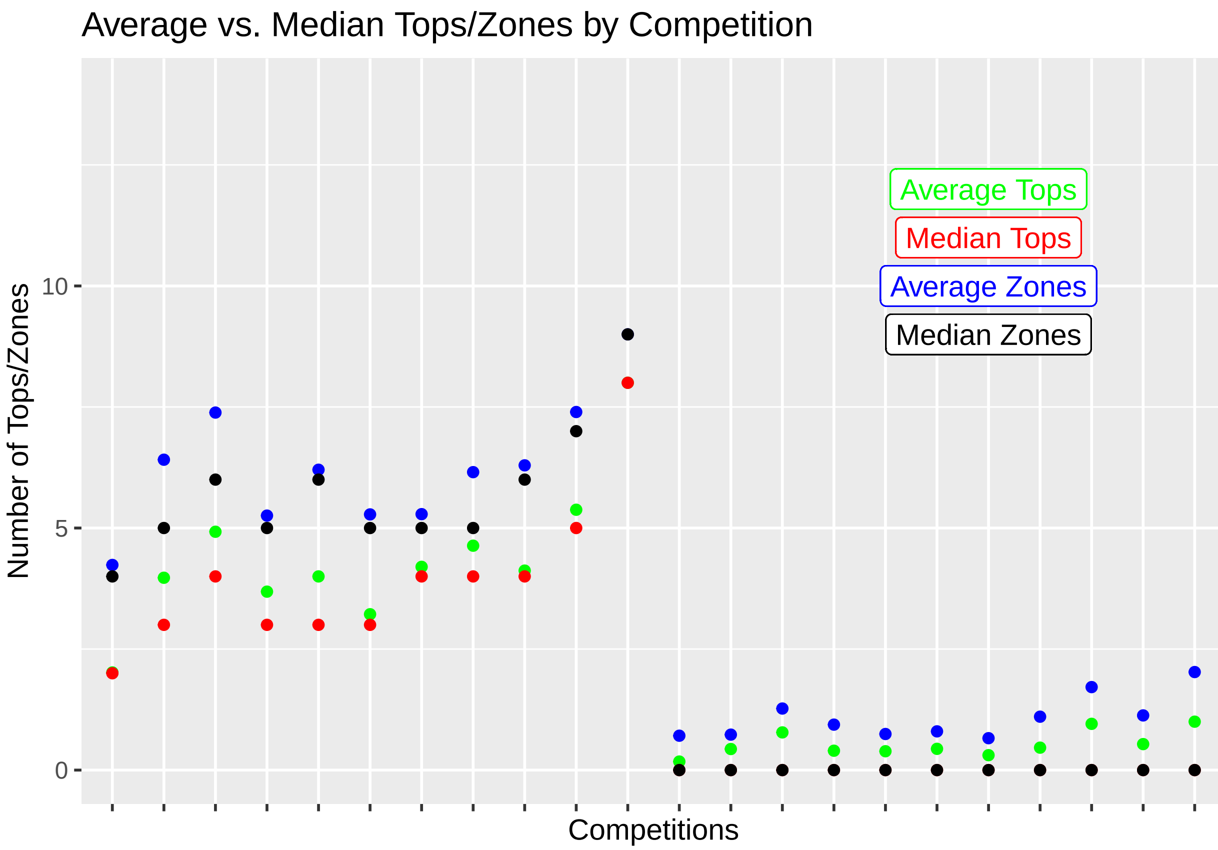
Given the number of climbers at any given competition far exceeds the number of problems in any round, there is up to a several hour gap between when the first athletes starts climbing and the final athlete begins. This, alongside potential fatigue/stress from waiting, makes an athlete's starting number a potential source of bias. To see whether this is the case, consider the following two plots:





In both plots, the best performing athlete had a starting number of approximately 600. For tops, the best performer is fairly isolated from all others. For zones, there are several other athletes who performed somewhat comparably who started anywhere from approximately 400th to 800th. Overall, those who climbed at the very beginning did worse than those who climbed later on.

The timing of the competitions may also have an effect on climber's performance. Since no two competitions have overlapping dates, this is the same as examining whether particular events had an effect on climbers' performances.



This plot[5] shows that the typical performance of the athletes is not even close to being uniform across competitions. The first half of the tournaments have notably higher mean and median numbers of tops and zones than any of those in the latter half, though this might be explained by some of the competitions being at the junior level (under 19 years of age) and others being at the senior level (17 years or older)[6].

Overview of the Modeling Process

The goal of these models is to predict whether a climber is an actual contender to win an IFSC competition. All of the models used here will be classification models based around the *Winner_Contender* column in the *bouldering* table[7].

Importantly, at no point should any model attempt to predict zone-performance based on tops. While zones are a requirement for tops, the reverse is not true. In any practical context, no model could predict information about zones based on tops prior to the athletes attempting the zones that are meant to be predicted. As such, the following potential predictors have be excluded:

1. *Name*: Given the similar trends between zones and tops, little additional information is gained by its inclusion.
2. *Rank*: This refers to the climbers' ranks post-competition and so, does not inform on how a climber will perform at said competition.
3. *Date*: This is functionally identify to *Competition* and is therefore redundant.
4. *Nation*: While there was a difference in the performances of nations, the data gleaned from this column is too susceptible to being influenced by outliers[8]

[5]For some of the competitions, the median tops looks like they are missing. They are not; since the median was 0, they are overlapping the median number of zones and are therefore obscured.

[6]Transitioning to senior level world cups becomes an option when someone turns 17 but is not mandatory until one turns 19.

[7]For the full details of the model development, please see https://github.com/Habeus-Crimpus/IFSC/blob/main/IFSC%20Classification%20Report.pdf

[8]There are outlier climbers in IFSC competitions. For example, Janja Garnbret, a Slovenian climber, wins over 80% of the competitions she participates in and there are some who have competed over 100 times without ever qualifying for semifinals.

5. *Competition*: This is being removed for two reasons: (1) Only 22 competitions are included in the data table and given the 1500+ competitors, this might lead to over-grouping of the data, and (2) I have insufficient information to determine whether the variation observed in the relevant plot is due to climate, location, or divisional (junior vs. senior competition) reasons.

I employ here *K-Nearest Neighbors (KNN)*, *Random Forest (rf)*, and *Logistic Regression (glm)* models.

Results

The following table summarizes the model performances.

Table 2: Final Model Comparison

	Model	Accuracy	Sensitivity	Specificity
Accuracy…1	K Nearest Neighbors (KNN)	0.8935396	0.9618240	0.4807692
Accuracy…2	Random Forest (rf)	0.9563239	0.9830329	0.7948718
Accuracy…3	Logistic Regression (glm)	0.9162875	0.9692471	0.5961538
Accuracy…4	Random Forest [manually adjusted]	0.9481347	0.9756098	0.7820513

The sensitivity of the models remains quite high, with the lowest rating still being above 96%, but the specificity ratings never exceeding 79%. This is an acceptable trade-off, since were those who are actual contenders to be deemed not to be, their nation's team might exclude them and thereby lessen their chances of winning. If, however, someone who is not a contender were to be selected for a team, it is less likely but not impossible that they could still win; their inclusion would not necessarily entail a poorer result.

The goal of these predictions is to tell whether an athlete should be classified as a contender to win a competition. The value of such predictions is best judged on accuracy and sensitivity. Accuracy works well as a baseline metric since relying on an accurate selection mechanic for team building will allow for the makeup of said team to be mostly, if not wholly, composed of qualified persons. Sensitivity is also important here since a person who is incorrectly classified as a non-contender despite their being a contender could deprive teams of their optimal members. Low specificity, on the other hand, could increase the likelihood of including non-contender on a nation's team but as noted before, being a non-contender does not mean that said person has a zero probability of winning; it just means that it is unlikely. That said, higher specificity can only serve to benefit a team-selection process. Since the random forest models are both more accurate than the logistic regression model, the latter is not the best model. Noting that the specificity ratings for both random forest models are approximately equal and their respective accuracy are separated by less than 0.8%, the slightly improved sensitivity of the second model gives significant support for it as the best model.

A random forest model provides the most accurate and sensitive predictions, making it the preferred model for predicting contenders to win an IFSC bouldering world cup.

4: Recommendations, Applications, and Concluding Remarks

IFSC bouldering world cup winning contenders can be predicted quite well using a random forest model. For national-level team selection committees, inclusion of model-driven selection criteria into the team selection process might help to alleviate the unpredictability of the single-day trials that are currently used to determine team composition. These same models can be applied in a gambling setting, as the ability to accurately predict whether a particular climber has a non-trivial chance of winning could be used to set odds for betting on IFSC bouldering events.

There are, however, some notable limitations to the models developed here. A binary classifier was used, meaning that a given climber either was or was not a contender. By expanding the number of categories to include groups such as strong contender, contender, weak contender, and non-contender, team selection committees would be more able to differentiate between candidates who are 'shoo-ins' and those who should be considered despite not performing quite as well as the very best. As to the implications for the Paris 2024

Olympic Games, bouldering is (as of the writing of this report) to be paired with lead climbing, which has a different scoring system and a somewhat differing skillset required. To better adapt the models developed here to the Olympic format, data regarding athelete performances at lead climbing events would need be evaluated and modeled in conjunction with the bouldering data.

Should further work be done on the data set as included here, it would be most impactful, if possible, to develop models that predicted actual placement in a finals based on competition data as it comes in. Since many athletes experience physical injuries during competitions, being able to determine based on qualification and semifinal performances the likelihood of winning could help to advise athletes who are considering whether risking further injury is worthwhile.

Final Notes

The original data set, available at https://www.kaggle.com/datasets/brkurzawa/ifsc-sport-climbing-competition-results, was created by Brett Kurzawa and uses data scraped from ifsc-climbing.org.

Some webpages and books were consulted in the making of this report. They include:

1. ifsc-climbing.org
2. https://www.kaggle.com/datasets/brkurzawa/ifsc-sport-climbing-competition-results
3. R for Everyone: Advanced Analytics and Graphics, Second Edition (Jared P. Lander)
4. https://rafalab.github.io/dsbook/introduction-to-machine-learning.html#evaluation-metrics