# OlympuSVM: Winning Athlete Recommender Models

Mohamed Al Elew
malelew@ucsd.edu

Rohan Bhargava
r2bharga@ucsd.edu

Simrandeep Singh
sis034@ucsd.edu

## I. IDENTIFY A DATASET

Olympic Athletes and Results:

The dataset includes 271116 rows that correspond to an individual athlete's performance in a specific event. There are 15 columns corresponding to information about the athlete such as age, sex, weight, and height as well as information about their team/nationality. The columns also include information about the event whether they medaled as well as the year and location of that Olympics. The extent of the data is from the 1896 Athens Olympics Games to the 2016 Rio Olympics Games.

The basic statistics of the dataset:

| | Age | Height | Weight |
|---|---|---|---|
| mean | 25.6 | 175.3 | 70.7 |
| std | 6.4 | 10.5 | 14.3 |
| min | 10.0 | 127.0 | 25.0 |
| 25% | 21.0 | 168.0 | 60.0 |
| 50% | 24.0 | 175.0 | 70.0 |
| 75% | 28.0 | 183.0 | 79.0 |
| max | 97.0 | 226.0 | 214.0 |

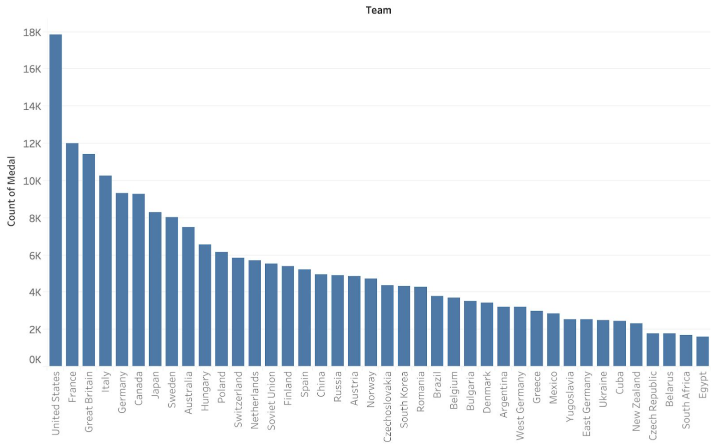
Age is distributed in more or a less a bell curve. Whereas the others are better described by sport as opposed to across the entire Olympic Games.

At a glance, we found several anomalies: years that had no events, countries that surged and disappeared, Games with record low participation, etc..

The Olympic Games over the years has been a venue for not just competitive sports but also geopolitics. There were no Olympic games held throughout World Wars. During the Cold War, the Olympic Games were utilized as a platform to voice opposition against host countries. This information was utilized to focus the extent of our data between the 1948 and 2016 Olympics Games since there haven't been any games canceled due to major catastrophic events.

These observations inspired us to explore whether an athlete's home country has an effect on whether or not they medal. Highly developed countries such as the United States, France, Great Britain, Italy, and Germany have been awarded the most medals. From our initial analysis, it isn't clear whether an athlete competing in their home country has an effect or not in terms of medals awarded.

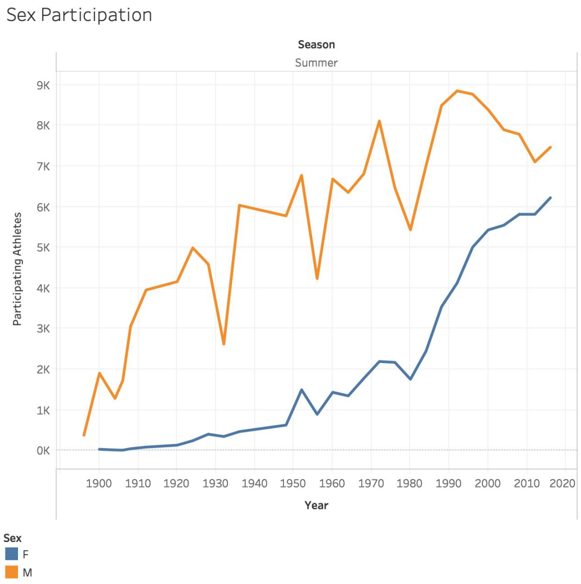
Olympic Team With Over 1500 Medals



Height and weight statistics appear to be clustered by sport. But, from a visual analysis, there seems to be no clear relationship between height and sport and winning a medal or not.

Next, we explored medal performance by athlete by Olympic Games. We saw that an athlete winning at least one medal at a previous Olympic Games led to it being highly likely that they will win a medal at their next Olympic appearance.

An athlete’s sex may provide useful insights or improve a predictive model. Unfortunately, all sex related explorations communicated the gender disparities in Olympic Games. Women historically have not been allowed to participate at the same rate as men in the Olympic Games. The 2016 Rio de Janeiro Olympic Games were the first where every single participating country sent women athletes including countries such as Saudi Arabia, Qatar, and Brunei.



## II. Identify a Predictive Task

Given this data set there were a few different options that we brainstormed. One of the most frequent forms of predictive task done utilizing Olympic data is predicting how many medals (or gold medals specifically) each country will win. These models often utilize some form of economic and population information as well as past Olympic results. We ultimately decided not to go this route since there was already a lot of modeling out there on this subject and it seemed likely that a large amount of medal variance stemmed from external factors - such as wars and other geo-political or economic forces. On the other hand, we noticed countries that had a high Gross Domestic Product (GDP) were winning a large number of medals, and considered using the number of medals to try and predict a country’s GDP.

Another idea we tinkered around with was predicting which sport or event a player might be participating in based on their country, age, height, weight, and utilizing K-means clustering to determine which cluster would be a best fit.

We ultimately concluded that an exciting way to look at the data would be to utilize it to predict individual Olympic medal winners. This would stand out a little against typical predictions focused on only the countries and could be used to figure out top competitors in events. A lot of other predictors that we saw predicting medalling for competitors relied on outside information such as how well they were doing in other national/international competitions. We wanted to rely only on Olympic history in order to make the model straightforward and reliable for all years.

For processing the data, we utilized the python data manipulation and analysis package Pandas. Pandas allowed us to easily manipulate and view the data utilizing built in functions such as the drop null values and on the fly filtering of the data.

The extent of the data is from the 1948 Olympics to the 2016 Olympic Games, again, the Olympics Games before 1948 had been cancelled multiple times due to World Wars.

We started off our focus on some basic features that were readily available directly from the dataset. This included an athlete’s country of origin and their height, weight, and age. The countries were one hot encoded and the height, weight, and ages are quantitative data.

From there, we decided to also build out some other features based on the fields provided in the data.

First, whether a particular competitor has competed in an Olympics previously as well as how many of each medal each participant won: gold, silver, or bronze. We felt that this feature would have one of the biggest positive impacts on our accuracy since people who have won medals, especially gold, in the past are more likely to win medals in the future.

Another feature that we considered and implemented was looking at the home field advantage - a boolean 1 or 0 on if the athlete was competing in their home country.

Next, a feature that counts the number of each medal an athlete has won across all of the Olympic Games resulted in a high accuracy and precision. Unfortunately, we realized the training set's medal counts included the count for every year including whichever year we were that was being predicted.

To resolve this, we decided that for every year being predicted we would calculate the medal counts for every Olympics post-World War II to the preceding Olympic Games.

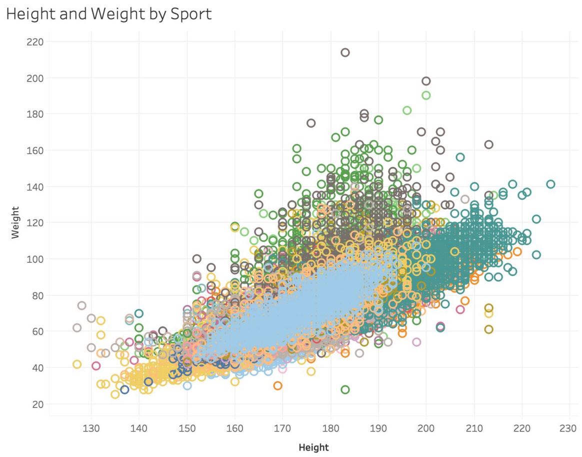
For a baseline model, we simply predict false for every prediction we end up with a fairly high accuracy of around 85% on average per year, since the vast majority of competitors do not medal. Our goal was to beat this baseline which would indicate that the model is accurately predicting medal winners and not just simply guessing the negative result for each prediction.

We assessed the validity of our model by testing it on the most recent available Olympic Games from 2004 to 2016 and utilizing the data before those years to train our model. This will communicate how a model is performing closest to the current date and will likely be able to help us determine a model's performance when predicting the medal results for the 2020 Summer Olympic Games.

## III. Describe your Model

Our final model is a set of support vector machines (SVMs) one for each individual sport and filters for their features: height, weight, age, team, whether the athlete has competed before, and how many of each medal the athlete has previously won in that sport.

Although our early explorations of the data didn't clearly communicate whether or not height and weight would be predictive of medal wins, including the measures in our features increased the accuracy by approximately one percent. Since height and weight is clustered by sport, training separate SVMs for each sport will decrease the noise added by the highly varied measures.



Our initial model was a single SVM that was trained on the features from age, height, weight, repeat attendance, team for all entries between 1948 to 2000.

The initial model did worse than the baseline model leading to the quick replacement of repeat Olympic appearance for whether or not an athlete has ever medaled. Whether or not an athlete has medaled is a better predictor of whether or not they will medal in the future, and the predictive task needs to optimize for determining winners since given any athlete and event there's approximately an 80 percent chance they did not medal.

Our performance improved after replacing repeat appearance with previous medal victory. Most of our accuracy scores either did as well or slightly better than the baseline.

The model did especially well with team sports with a precision of near or equal to one. Each individual on a team sport has their own entry in the data, so team sports has a larger training set of the number of medal types times the team size. The model probably does significantly better for team sport medal predictions since the set is larger and is trained with more positive results.

We focused on improving the predictions for a few sports to better understand what the model is doing and determine how to best improve individual event prediction. One sport we decided to focus on was gymnastics, because it is a large subset of the data and our accuracy for gymnastics was an abysmal 18 percent.

A confusion matrix of the predictions for each sport showed that our model was outputting the baseline results for most sports. This result is completely understandable given the distribution of medaled athletes to non-medaled athletes motivated us to focus the model on medal winners: three medal winners per individual sports and the corresponding number of medal winners for team sports.

The next iteration of the model implements multiple SVMs and a decision function similar to homework four. There is a SVM for each sport trained on its corresponding feature set: height, weight, age, team, and medal counts.

This model resulted in significant improvements. Previously, the majority of sport medal predictions had a precision score of zero, where as utilizing multiple SVM resulted in non-zero precisions scores for almost all sports. The model was no longer making a few poor positive guesses and defaulting to the negative prediction. This is thanks to there guaranteed positive predictions for each sport, and a subset of those predictions being true positives.

The model was consistently beating the baseline accuracy and had a non-zero precision scores. To try and improve the prediction, there were a two features we inserted to attempt to improve the model: repeat appearances and home-field advantage.

Although repeat appearances was removed from the feature vector earlier, it was re-inserted as an exploration of whether it was actually hurting the model's performance. The reinclusion of the repeat appearance feature increased accuracy marginally, approximately one percent. It turned out that the medal count feature was a strong predictor for winning a medal, but not exactly a perfect replacement for repeat appearance.

The home field advantage was a phenomenon we were excited to include in our mode, but it was reserved for after we built a solid model since we expected it have a marginal effect on the predictions. We assumed that home field advantage may be predictive of an athlete medaling. This feature, however, ended up inducing more noise and had a negligible effect on our accuracy and precision score. We believe that although the home field advantage is a true phenomenon, it did not prove to be helpful for our model. The home country is already measured in the team feature and it is likely could confounding the model to double count it without the home field games inducing a significant advantage.

**The Underperforming Model**

We've spoken to extent about the SVM, and the many iterations that it went through, but we also trained a logistic regression alongside the SVM; its low technical debt made it easy to keep in our for loops since the SVM was the only thing taking much time.

However, the logistic regression performed worse than the SVM in every category, sport, and event except for one. In general predictions from the logistic regression had a lower precision which caused its overall accuracy to drop significantly while the overall accuracy of the SVM dropped only a little since its higher precision made up for it.

Accuracy:

| Sport | Baseline | Log. Regression | SVM |
|---|---|---|---|
| Ice Hockey | .69 | .50 | .78 |
| Sync. Swimming | .75 | .73 | .84 |
| Gymnastics | .91 | .82 | .90 |
| Overall | .86 | .75 | .84 |

Precision:

| Sport | Baseline | Log. Regression | SVM |
|---|---|---|---|
| Ice Hockey | .0 | .16 | .64 |
| Sync. Swimming | .0 | .25 | .68 |
| Gymnastics | .0 | .03 | .36 |
| Overall | .0 | .09 | .39 |

## **IV.** Describe Literature

We used an existing Kaggle dataset that was scraped from www.sports-reference.com, a tracking website for different sports statistic. The dataset was scraped to provide an easy to use interface to perform analysis on the history of the Olympic games. In particular, the author wanted to give people an opportunity to ask questions about “the participation and performance of women, different nations, and different sports and events.” Similar datasets also cover the history of the olympics and changes over the years but they tend to focus on countries and their attributes instead of individual athletes.

Predicting winners in individual events is a very commonly studied problem. The Economist discusses an Elo based system for head to head events, where an Elo rating is calculated for each participant based on their previous results.[1] This Elo rating helps provide a baseline for the expected results. However, one problem with the Elo rating is that it only functions well in head to head events and doesn’t work for group events such as many track and field, swimming, and winter olympics events. In the future, we would want to adjust our model to take into account the Elo rating for each participant as well - this would provide a much better determination of their skill level than just height, weight, and age.

Another big related area of study isn’t based on individual events, but instead the total medal counts for countries overall. These models tend to be much more successful since they provide a holistic overview and have more room for error. One model, developed by researchers at Dartmouth, is focused on predicting medal counts based on four economic factors: “A country’s population, its comparative level of wealth, its performance in previous Olympic Games, and whether it is hosting that year’s Olympic Games.”[2] From the articles, it appears that these models typically rely on statistical methods, with assigning a weight to each category manually.

One common pain point raised across many reports is that the winter games are notoriously more difficult to predict because so many countries that typically do well at the summer ones don’t do well at the winter ones. Combined with the smaller amount of events and many less head to head events, the medal counts

FiveThirtyEight, a data driven news site, did some research before the 2016 Rio de Janeiro

[1] T, J. “How to Predict Winners at the Winter Olympics.” *The Economist*, The Economist Newspaper, 13 Feb. 2018, www.economist.com/game-theory/2018/02/13/how-to-predict-winners-at-the-winter-olympics.

[2] Swanson, Ana. “How to Predict Olympic Results before the Games Even Start.” *The Washington Post*, WP Company, 22 July 2016, www.washingtonpost.com/news/wonk/wp/2016/07/22/how-to-predict-olympic-results-before-the-games-even-start/?noredirect=on&amp;utm_term=.d912d97daeae.

Olympic Games into the "home field advantage", where the the hosting country will win more medals than they typically do.[3] However, this advantage is sometimes fickle, such as in the Rio olympics, where Brazil faced a lot of public backlash for the corruption and money spent on the games and people predicted it would have less of a home field advantage. All in all, there are a lot of different predictions that people have made about the olympics however our approach is still a little bit different since we hope to be able to predict medal winners without any previous knowledge about their skills.

## **V.** Results and Conclusions

From the start, we knew that our goal was to beat the baseline accuracy while increasing our precision score. Since the baseline we were testing against predicted  "no medal" for every single competitor, the baseline precision rate would be 0. Thus, by successfully predicting some medal winners our model would be able to beat that precision.

Model Performance: the final SVM model had overall accuracy 84 percent and precision of 39 percent. The individual sport SVM accuracy varied from the high 80s to the low 60s, and the precision varied from from zero to one.

In the future, I think we would want to improve our analysis by augmenting this dataset with Elo ratings. These ratings would provide a better sense of an athletes skill level than just height, weight, and age. Since Elo ratings are based off of athlete performance in recent events, it is easier to keep up to date and take into account recent injuries and other factors that might limit and athletes performance.

Another way to improve the model is to focus on one event at a time and include all geopolitical factors that might be related to the countries competing. One example of this is in the 2016 Rio Olympics where the entirety of the Russian track and field team was banned from competing. This might lead to unlikely competitors doing surprisingly well in these events and we might want to account for that in our model.

We also want focus more on the home field advantage. Although our naive implementation of the home field advantage did not end up working very well, we want to modify our model to implement a more accurate way to implementing this home field advantage with comparing a ratio of more medals won with the advantage vs without.

Overall, our model performed well enough that we could make impressive prediction at a viewing party, but would not feel confident enough that we'd but money on the line.

[3] Pettigrew, Stephen, and Danyel Reiche. "Is There Home-Field Advantage At The Olympics?" *FiveThirtyEight*, FiveThirtyEight, 10 Aug. 2016, fivethirtyeight.com/features/is-there-home-field-advantage-at-the-olympics/.