Olympics Final Report: Predicting Olympic Medal Counts by Team

Jillian Green, Data Science Student at Brown University

https://github.com/jillian-green/1030_Final_Project

INTRODUCTION

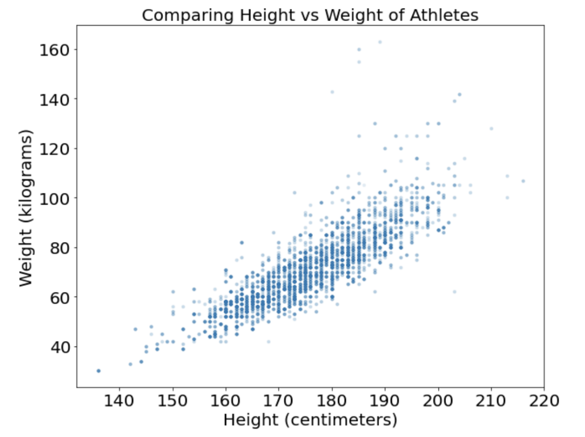
In this report we explore the Olympic Games from Athens 1896 to Rio 2016. Winning olympic medals is important for a country because it promotes job opportunities, increases investments (sponsorships, tourism, revenue) and creates a sense of national pride. In this report we want to predict the number of medals a country will win. Our target variable is “Medal_Won_Corrected”. We created the “Medal_Won_Corrected” target by accounting for single and team events (only counting a team win once rather than per athlete). Because we are trying to predict the number of medals, this is a regression problem.

The original dataset included 271,116 data points, with columns such as athlete name, sex, age, height, weight, etc. After manipulating and cleaning the data, the final dataset used is 20,799 data points with columns Team, Population, Total_Athletes, Female_Athletes, Male_Athletes, Year, Season, and Sport. The olympic dataset was joined to a population dataset to get population by team (country). The olympic dataset also required cleanup. This included but is not limited to: country code names being fixed so they could be joined to the population data, team and single events being accounted for to get “Medal_Won_Corrected”, population data only being available from 1961 onward so the olympic data is filtered to 1961 onward. Missing data was also present in the “Medal” column and “Population” column. To account for missing “Medal” data, empty “medal” columns were converted to “DNW” (did not win). Missing “Population” data was a bit trickier. Russia and Germany had similar scenarios: in the past, each of these countries competed with multiple teams. To solve this problem, all Russian teams were given the same population, the same with Germany. Serbia’s population data did not exist in the population dataset, so this was added manually. The rest of the missing population data accounted for ~1% of the data, with few medal wins, so these rows were removed from the dataset.

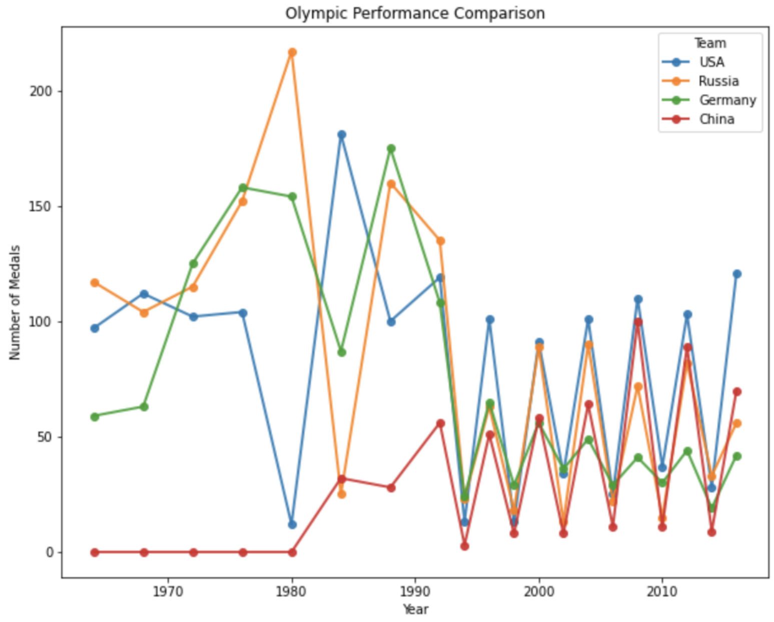
This dataset is from Kaggle, where there are numerous projects and publications about the Olympic dataset. One project by Sagar Chadha was particularly inspiring for this project. Chadha focused on predicting medal tally using gdp and contingent size (correlation to medal tally, 0.622 and 0.703 respectively). Chadha created a regression model and found an r^2 value of 0.75, meaning his variables accounted for 75% of the medal tally variation (Chadha, 2018). Another author, Tomasz Klimek, focused on women’s participation in the Olympics. Klimek found that in 1896 there were 0 female athletes, but since 1900 the number has increased. Klimek also classified sports as “manly” (-1) or “girly” (1) and found that summer has a more noticeable split with men in weight lifting sports and women in swimming, whereas winter includes both men and women skating and skiing. Further, women from Russia, USA, and Germany have acquired the most summer medals, and women from Russia, Finland, and Canada have gained the most winter medals (Klimek, 2018). Both population and gender and features I am interested in exploring.

EXPLORATORY DATA ANALYSIS

After performing EDA on the dataset, there are many interesting observations we can make. Each row in the final dataset represents the number of medals and athletes per Team, Year, Season, Sport combination.

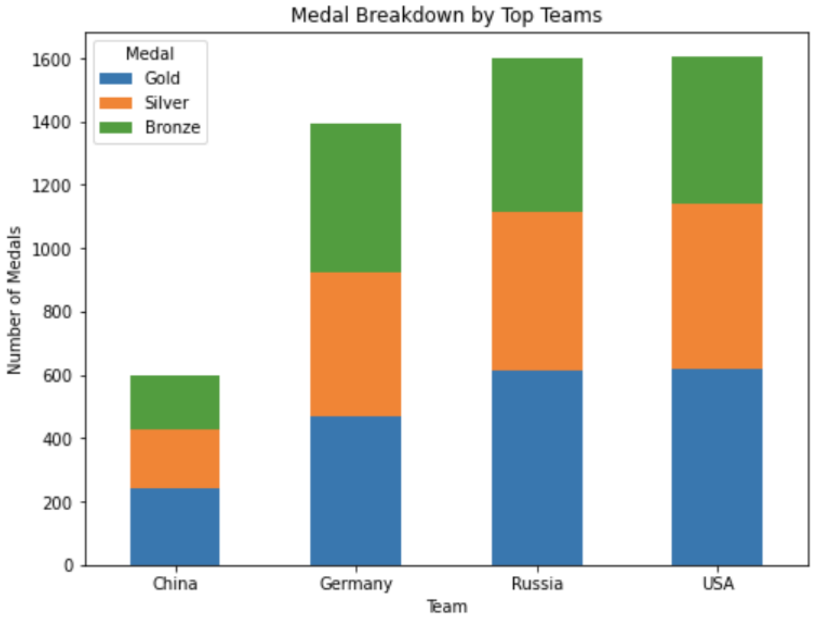


Before manipulating the original Olympic dataset, we had access to other metrics. We were able to see the weight and height of each athlete. The figure to the left compares the height and weight of gold medal winners between 1924 and 2016. We can see that there is a linear relationship between height and weight. Yet, finding average height and weight by team, season, year, sport may not necessarily be a helpful feature. If you think about gymnasts, you'd expect them to have similar body proportions no matter what team they are on.



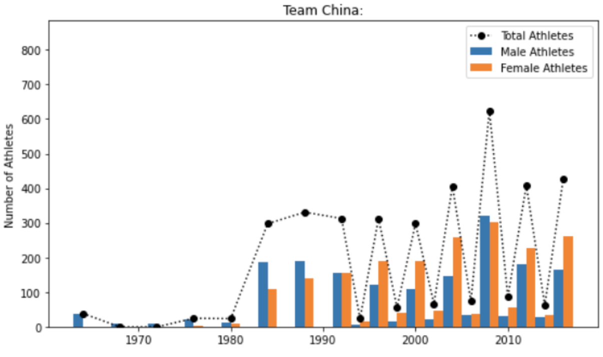
If we want to predict why particular teams are high medal earners, a helpful exercise in EDA is to filter our dataset to better understand the top medal winning countries. We look at the top 4 countries -- USA, Russia,Germany, China.

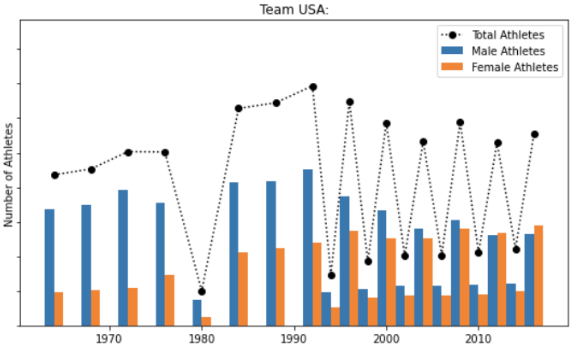
The image to the right breaks down the number of medals won per country. We can see that before the early 1990s, the USA, Russia and Germany dominated. It wasn't until approximately 1984 that China began winning. This may be a result of more participants in the olympics from other countries and better overall talent.

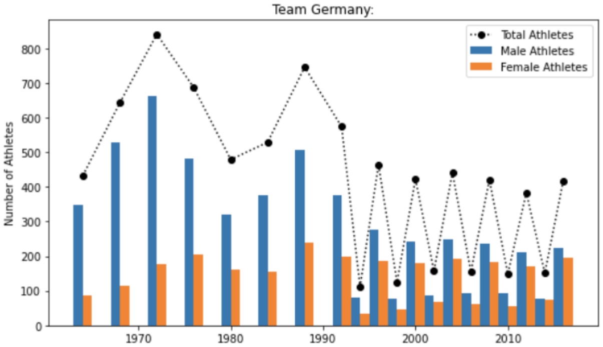


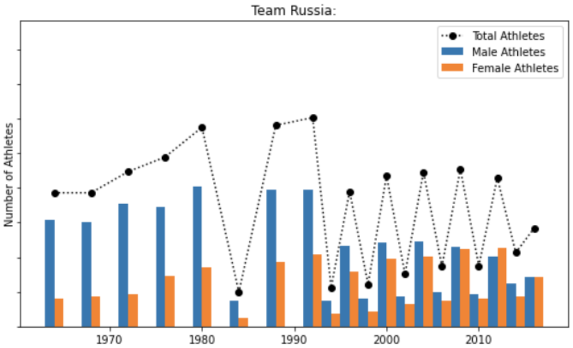
Further, we can look at medal breakdown per country by "type" of medal. We see in the image to the left, a similar number of bronze medals for Germany, Russia and the USA. Yet when looking at gold medals, Russia and the USA have about 150 more than Germany, and about 400 more than China. One possible explanation for this could be that these "powerhouse" teams are all winning medals because they are competing and winning in different events, against other teams.

We can also break down athletes (total, male and female counts) by country. In the image below, we can clearly see that males dominated olympic participation until around the 1990s when the number of female athletes began to increase. The number of male and female athletes may impact medal wins, as some events are gender specific.

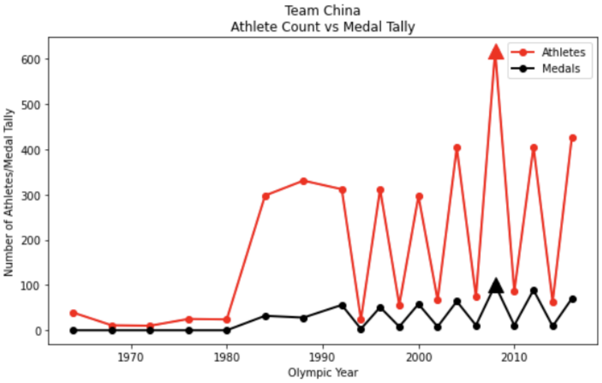


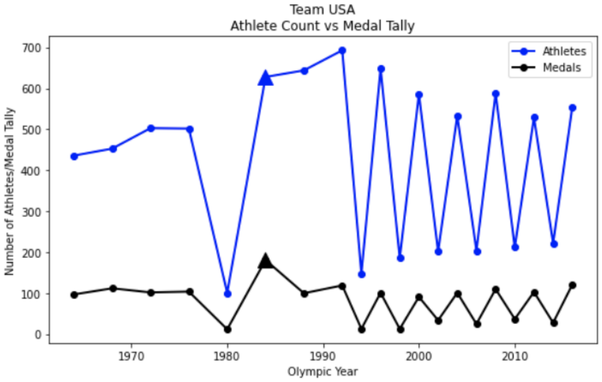


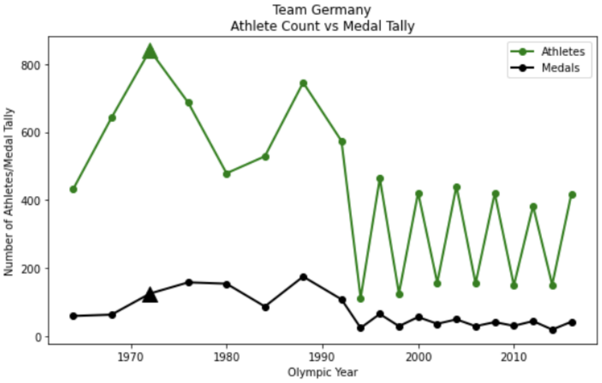


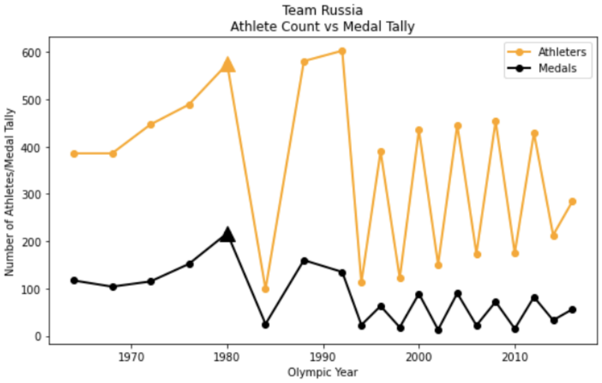


Lastly, we can compare the number of athletes to the number of medals won by year. Interestingly enough we see peaks of athletes and medals won align. This further suggests that the number of athletes may correlate to medals won. This makes sense, the more athletes you have, the more chances you have to win a medal.









METHODS

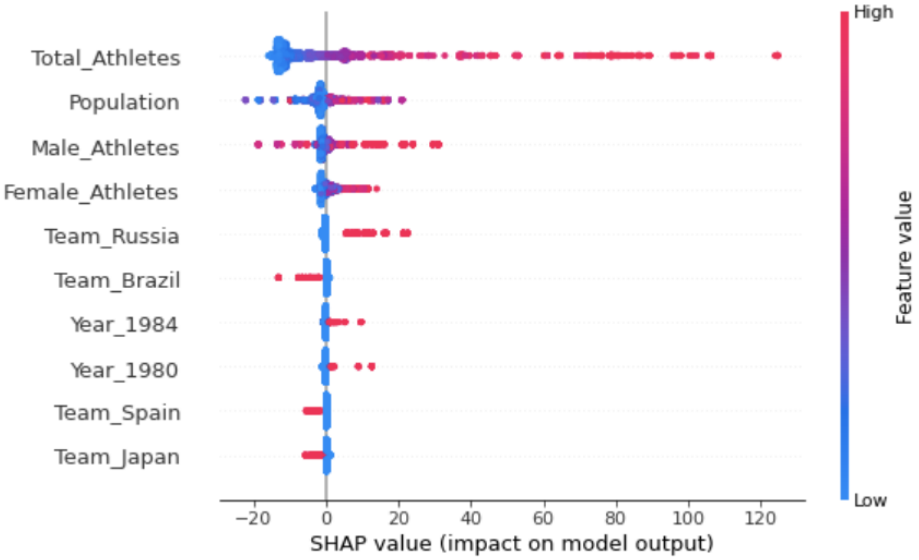
In this problem our goal is to predict the number of medals a team wins per event and year. The splitting strategy used is TimeSeriesSplit since we want to split data based on a team's past performance. In other words, to predict how many medals the USA will win in the next olympics, we'd want to know their past performance. Once we split the data, we preprocesses the data with the following feature breakdown:

- OneHotEncoder (Categorical, No rank): Team, Year, Sport, Season
- StandardScaler (Continuous, No boundary): Population, Total_Athletes, Total_Athletes, Male_Athletes

We use GridSearchCV to help fit the data and find the best parameters and test scores per model. The models we use are Lasso (linear regression with l1), Ridge (linear regression with l2), and KNeighborsRegressor. For Lasso and Ridge we tune the alpha parameter (both use alpha values [1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2]), and find the best alphas to be 0.01 and 10, respectively. For KKNeighborsRegressor we tune weights and n_neighbors. The weight values we try are ["uniform","distance"], and the n_neighbor values are [1,25,50,75,100], and the best parameters are n_neighbors=1, weights=uniform. To help decide which regression model performs best, we use the evaluation metric of mean squared error. MSE tells us how close a regression line is to a set of points by taking the distance between points and the regression line (the "errors") and squaring. We determine which model is best by finding the one with the lowest mean of test scores. To account for uncertainties, we looped through 10 random states. The model with the lowest average test scores was KKNeighborsRegressor.

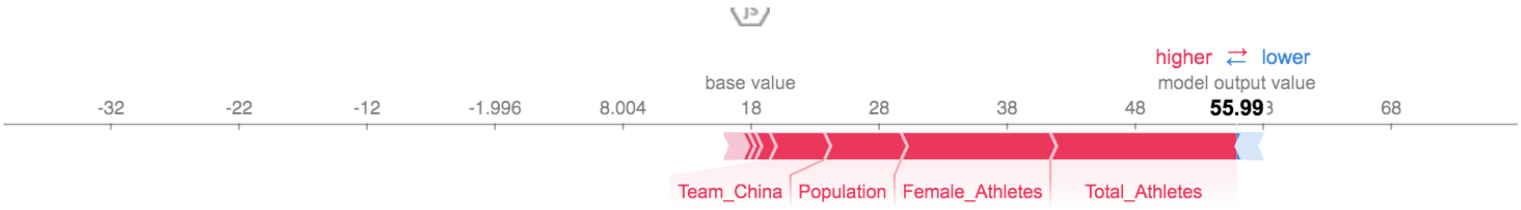
RESULTS

We found the baseline score to be ~ 28.001. Our Lasso model is 8.1826 standard deviations above the baseline, Ridge model is about 8.0664 standard deviations above the baseline, KKNeighborsRegressor was 30.656.

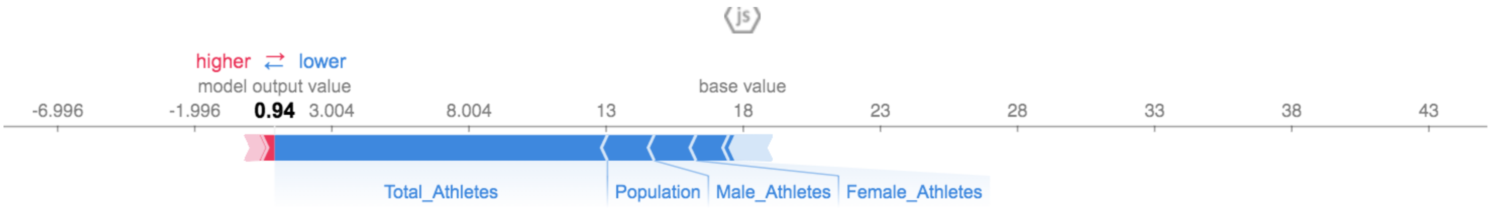


To understand our model better, we can look at feature importance. We do this by training an XGBoost model, and then using SHAP to plot the most and least important features. As seen in the image to the left, Male_Athletes, Total_Athletes and Population have the highest impact on model output. Season wise, we see Summer has a larger impact that Winter. We also see that year and team have a lesser impact on model output.

In the force plots below, the red arrows show feature effects that drive the prediction higher, whereas the blue arrows show ones that drive it lower. The arrow size is the magnitude corresponding to the feature's effect. The image below represents the point where index=0. At this point, we see that Total_Athletes and Female_Athletes have the largest positive affect, whereas Team_Russia has the largest negative affect. Similar to the plot above, years are not the most impactful features.



If we look at the point with index=200 below, we see a different behavior, where Team_Sudan has the largest positive impact, and Total_Athletes and Population have the largest negative affect.



In conclusion, Total_Athletes, Population, and Male_Athletes have the largest impacts. This suggests that population size and the number of males competing on your olympic team will impact how many medals a team wins. For example at index=0, more females and more overall athletes will likely result in more medals. It was surprising to see Female_Athletes as a positive impact over Male_Athletes. It was also surprising that Russia, one of the top medaling teams, has a negative impact at index=0.

OUTLOOK

There are always ways in which a model can be improved. More features and more data is always a good start. More features, such as gdp or athlete salaries, may help predict medal winning teams. As mentioned earlier, Chadha found that gdp and contingent size correlate. Further, athletes with higher salaries may be enticed to train harder and may have access to better equipment and funding. More data will likely help improve model performance. The more we know, the more knowledge we have that we can use to better predict. In this Olympic medal winning scenario, it may be helpful to have more data distinguishing countries. Beyond population, what do high medal winning countries have that countries with few wins don't have? What do both countries have in common that must not affect how many medals a country wins? There are many questions and paths we can explore here.

As mentioned above, we combined all Russian teams to be one (same for Germany). This could have given Russia or Germany too many wins, thus affecting our model's predictions. To improve the model, we may want to consider keeping all Russian and German teams separate and using data beyond population if we aren't able to find a population split.

We could have used other techniques in our model as well. For instance adding more parameters to the models or increasing the parameter range may result in better parameters. We could have also increased the number of n_splits in the TimeSeriesSplit method. Further, we could have tested a different evaluation metric to see if results were better. Finally, we could have also explored other regression models such as random forest regressor.

Overall, models can always be improved and we must continue to check on them and update them over time to account for changes in the data or new information.

REFERENCES

Dataset: 120 years of Olympic history: athletes and results, Most recent update by rgriffin, (2018) (https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results)

Countries Population, most recent update by Alexander Shakhov (2017) (https://www.kaggle.com/centurion1986/countries-population)

Marco Giuseppe de Pinto, (2018) Let's discover more about the Olympic Games! (https://www.kaggle.com/marcogdepinto/let-s-discover-more-about-the-olympic-games)

Sagar Chadha, (2018) Olympics Data- Cleaning, Exploration, Prediction (https://www.kaggle.com/chadalee/olympics-data-cleaning-exploration-prediction)

Tomasz Klimek, (2018) Women of Olympic Games (https://www.kaggle.com/kl13mk/women-of-olympic-games)

Link to github repository: https://github.com/jillian-green/1030_Final_Project