Project 4: Olympics

Introduction

The Olympic games of Summer 2024 were earlier this year and were met with tons of fanfare. There were many popular events such as track and field, basketball, soccer, gymnastics, and volleyball that drew the attention of many spectators around the world. Many medals were won, but have you thought about what it takes to win an Olympic medal? Yes hard work and dedication of course, but what if we could use predictive technology such as machine learning to gain an edge as well. In our project we do just that by using an Olympic dataset of both the Summer and Winter Olympic games from their inception to 2016. With this wealth of data, our goal is to create interesting and insightful visuals with Tableau, create a predictive model that can reveal how likely an individual is able to medal in a certain event based on physical attributes, and to cohesively present it all in a cool and functional web application.

Data Cleaning & Engineering

The modules we used were pandas, numpy, scikit-learn, matplotlib, and seaborn.

This dataset included 120 years of historical data of the Olympic Games, from Athens 1896 to Rio 2016. Each row represents an individual athlete competing in an event, and includes the athlete's name, sex, age, height, weight, country, and medal, and the event's name, sport, games, year, and city where the game took place.

During the initial data exploration, there were null values for Age, Height, Weight, and Medal. Null values for Medal simply meant the athlete did not win a medal so we filled it in as “No Medal”. For Age, Height, Weight, we filled in the null values with the median grouped by Sex and Sport.

We thought realistically about the type of information that is actually needed to predict whether an athlete would win a medal or not. We wanted to train the model on only the

most essential features. So we dropped the columns ID, Name, Team, Games, City, and Event. The features used to train the model were Sex, Age, Height, Weight, NOC (name of country), Year, Season, and Sport.

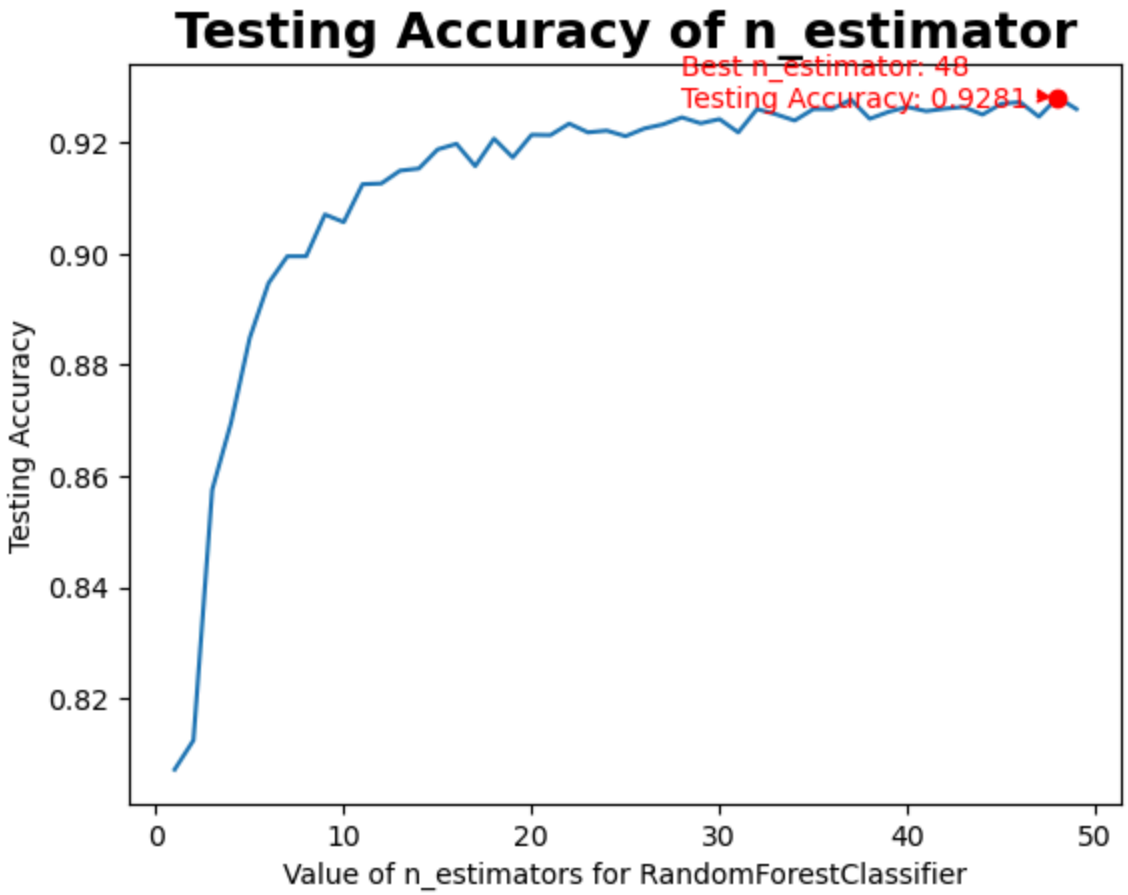
Lastly, we focused on only two Olympic games, 2014 and 2016. The sheer volume of data made the model run incredibly slow and created a very large output file. Focusing on two years still gave the model plenty of records to train on and created a more efficient runtime and manageable output file size.

Machine Learning Model (Genie)

Because of the nature of a competitive sporting event, where there are far fewer athletes that make it to the podium than the total number of competitors, our data was unbalanced. Those who won medals comprised 15% of the dataset and non-medal winners were the other 85%. We had to use SMOTE to over-sample the data in order to balance it.

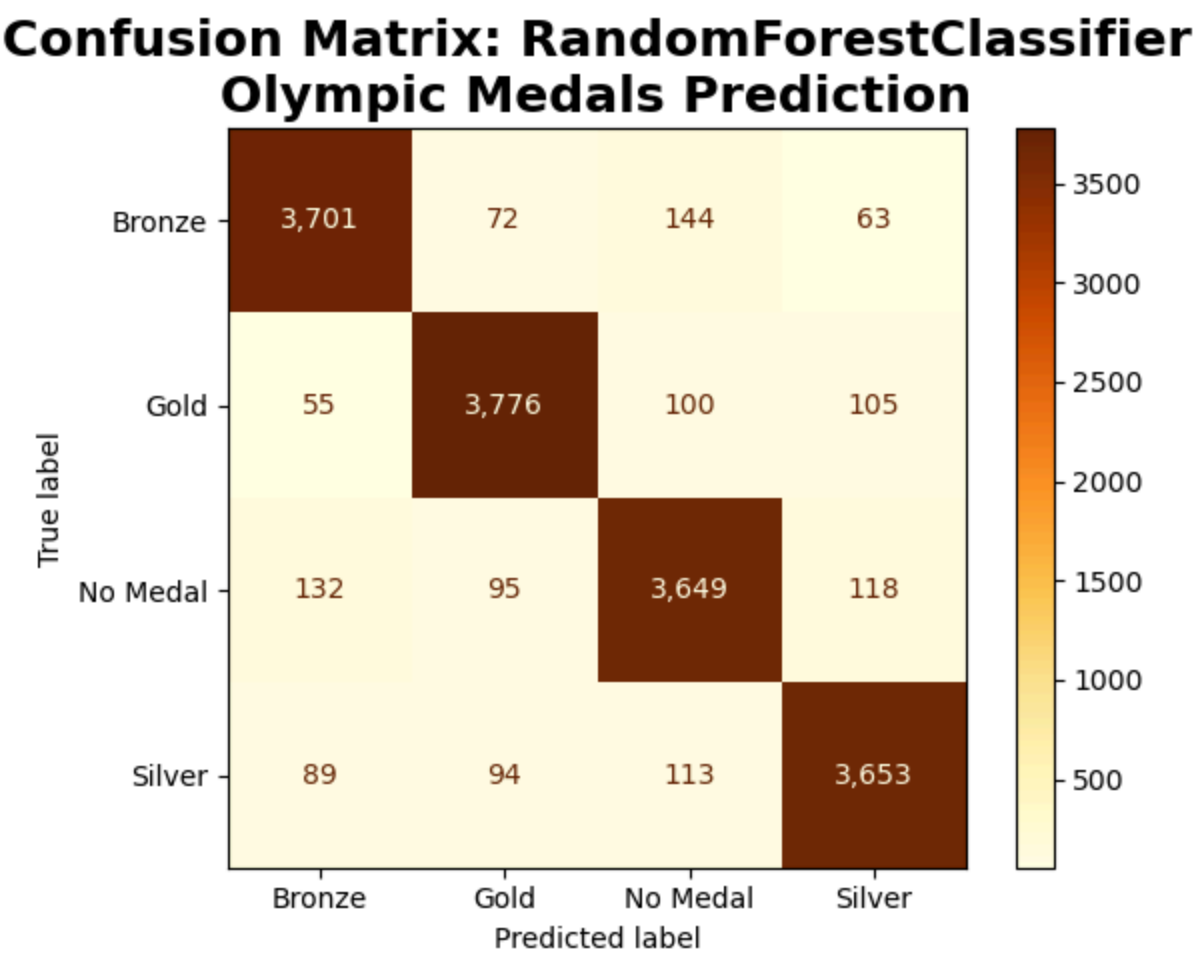
The model we decided to use was Random Forest Classifier because we had a multi-classification prediction to make: Will the athlete win gold, silver, bronze, or no medal? One of the parameters of this model is the n-estimator, or the number of decision trees in the forest. Usually the higher the number, the higher the accuracy but will require considerable more computing power at diminishing returns.

We decided to test a range of n-estimators from 1 through 50 to find the best accuracy. Below is a plot of our test. The best n-estimator to use was 48 for our model.

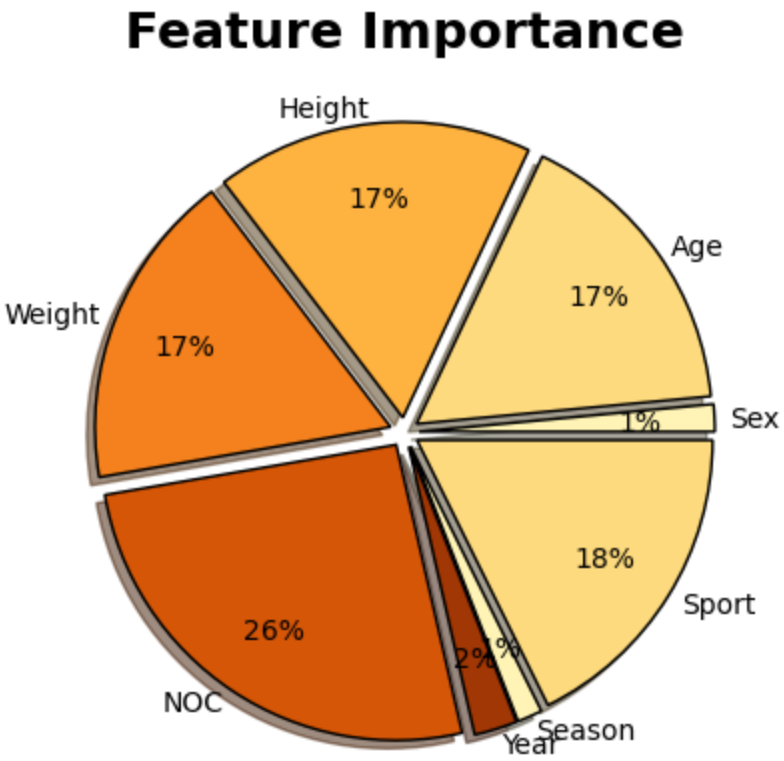


After running the model, this is the resulting classification report and confusion matrix.

	precision	recall	f1-score	support
Bronze	0.93	0.93	0.93	3980
Gold	0.94	0.93	0.93	4036
No Medal	0.91	0.92	0.91	3994
Silver	0.93	0.93	0.93	3949
accuracy			0.93	15959
macro avg	0.93	0.93	0.93	15959
weighted avg	0.93	0.93	0.93	15959

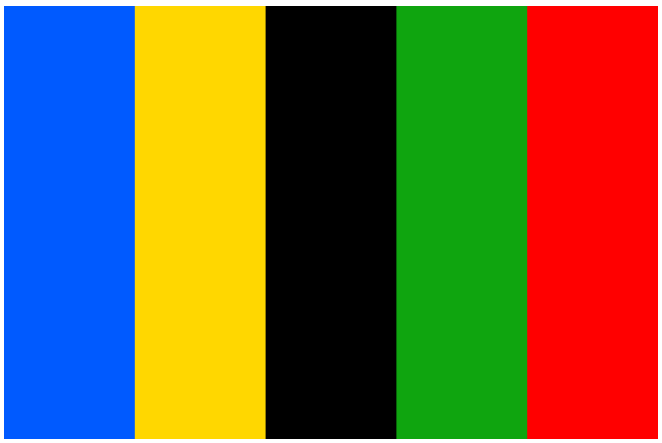


Then we were able to calculate the feature importance. Based on the pie chart below, the Name of Country is the most important feature for predicting medals, followed by Sport, Height, Weight, and Age.



Color Design Considerations

Our main color palette for the project was the color of the Olympic rings, as shown below. The Olympic rings signifies the unity of the five continents and the Olympic athletes of various countries coming together. These bright, distinct colors are displayed throughout our web page and Tableau visualizations. We also used gold, silver, and bronze colors to represent the different medal types in certain visualizations.



Dashboard Design Concepts - Part 1

Our first Tableau dashboard focuses specifically on medal breakdown. The main visualization of this dashboard is a world map that displays the medal count of each country. The size and color of each color signify the medal count. The larger the circle, the more medals the country has. The Olympic ring color palette was used, with blue representing lower medal count and red meaning higher medal count. This visualization can be filtered by sport and medal type (gold, silver, bronze, no medal).

The United States has the most overall gold, silver and bronze medals, but it is interesting to see which countries have higher medal counts for specific sports. For example, Japan and Cuba tie for the most overall medals in baseball, both with a total of 112. In comparison, the United States has 88 total medals for baseball. Another example is that Great Britain and France both have high overall medal counts for cycling, with 582 and 679 total medals respectively. The United States, on the other hand, has 523 overall medals for cycling.

The second visualization of this dashboard is a bubble chart showing the athlete count per sport. Again, this visualization uses the Olympic ring color palette. Red symbolizes higher athlete count, while blue is for lower athlete count. This visualization can be filtered by sex (M/F) and season (winter/summer). The overall most populous sport for both male and female, and regardless of season, was athletics. This includes track and field and other running and jumping competitions. Athletics has the highest athlete count specifically for females (11,666) with swimming coming in second (9,850). Again, athletics has the most male athletes (26,598), and gymnastics has the second most male athletes (17,578).

The last visualization for this dashboard is a stacked horizontal bar chart that shows the medal count per team (country). Each section of the bars represent the different medal types, with the gold, silver, and bronze colors differentiating the medal counts. The athlete count of each medal type is displayed when hovering over the bars for each

country. The United States has the most gold (2,474), silver (1,512), and bronze (1,233) medals compared to every other country.

Dashboard Design Concepts - Part 2

The Olympic Basketball Tableau dashboard dives into the average physical attributes of each Olympic basketball team over the course of the Summer Olympic games and how they reflect on the success of obtaining medals. These physical attributes include the team's average height, weight, and age. You are able to filter the year of the games you would like to research and the sex of the team. The dashboard contains two visuals. The first being a bubble chart that highlights a team's average height and weight during a certain tournament. A team's average height is reflected on the size of the bubble and the average weight is reflected in the darkness of the bubble. With both large and dark bubbles reflecting a greater height and weight of the team. With this visual you are also able to see the name of the country the bubble represents as well as their medal results at the end of the tournament. The average height and weight is also listed when you hover over a bubble. Based on this visual you can see how much of a factor physical attributes played a part in any particular game for both men and women.

The next visual in the dashboard is an interactive table that takes a look at the average age of a basketball team. Often in basketball there is a certain sweet spot of age when it comes to physical ability and adequate basketball IQ. Younger players can have a high physical and athletic ceiling with a lower basketball IQ and older players may be in a physical and athletic decline but have a vast knowledge of the game. Since the 90's there has been more of a focus on finding that balance when crafting Olympic basketball teams. By using the interactive table you are able to see how age played a factor in winning medals compared to the rest of the tournament field. The countries, their medal results, and their average ages are displayed on the table.

Bias/Limitations

The Olympic dataset used did have its fair share of bias and limitations. The first being its size. As many can imagine, using a dataset with data dating back to 1896 to 2016 can make for a very large file. Especially when you can draw down to the individual athlete with their various data points. This in turn made the diversification of the model more limited. There were also a lot more summer sports than winter sports in the data set, which can also skew how the model was trained.

Future Work

In the future it would be beneficial to add more features to the model such as an athlete's years of experience in a sport, how many Olympics have they competed in previously, and amount and types of injuries they have sustained in the past. This would provide improvements to the models predictability on medal attainment. In terms of how Olympic teams perform it would be helpful if features based on the team's coaches experience were implemented as well. In almost all sports, coaching plays a huge factor in the success of a team. For outdoor events, having data on the climate of the host country and how it could affect performance would paint a better picture into how an athlete would perform. Especially if their country of origin has a significantly different climate. Lastly, having the ability to cross-reference the Olympic data with census data from the different countries could provide some insight on how huge of a role does population play on a country's Olympic success.

Conclusions/Reflection

There were many takeaways that came from the data and the predictive model. Within the predictive model the athlete's country proved to be the most important feature. Especially when pulling from countries with great Olympic success. In the Tableau visuals we can see that the United States had the highest amount of total medals over the course of all Olympic games and is historically dominant in the sport of basketball. The most athlete-dense sports are track & field, gymnastics, and swimming. Both the model and the visuals are very feature heavy and can make a lot of interesting predictions and showcases that are all cohesively packaged in the web application.