OLYMPICS

Eugenia Chien, Herbert Dennis, James Lee, Lindsey Krempa

Purpose & Inspiration

- Inspired by the popularity of the 2024 Summer Olympics in Paris.
- Exploring the rich Olympic history of athletics on the world stage through a wealth of data.
- Will the future of the Olympic games incorporate predictive analytics to achieve the gold, silver, and bronze?

Research Questions

- What metrics are most important in predicting podium placement?
- What countries are most successful in the Olympic games?
- Is there a formula for success in certain events?

Design Concepts

- Project color palette
  - Inspired by the colors of the Olympic rings
  - Also used gold, silver, bronze colors to represent medal colors

- Tableau Dashboard 1 - Medal Breakdown
  - World Map showing the medal count of each country
  - Bubble Chart presenting the athlete count per sport
  - Stacked horizontal bar chart displaying the medal count of each country

Design Concepts

- Tableau Dashboard 2 - Olympic Basketball
  - Bubble chart representing average height and average weight per team of each Olympic games for both men and women. The larger the bubble, the greater the average height of the team. The darker the bubble, the greater the average weight of the team. The medal results are featured on each country's bubble.
  - Interactive table representing player average age per team of each Olympic games for both men and women. Grouped by medal results.

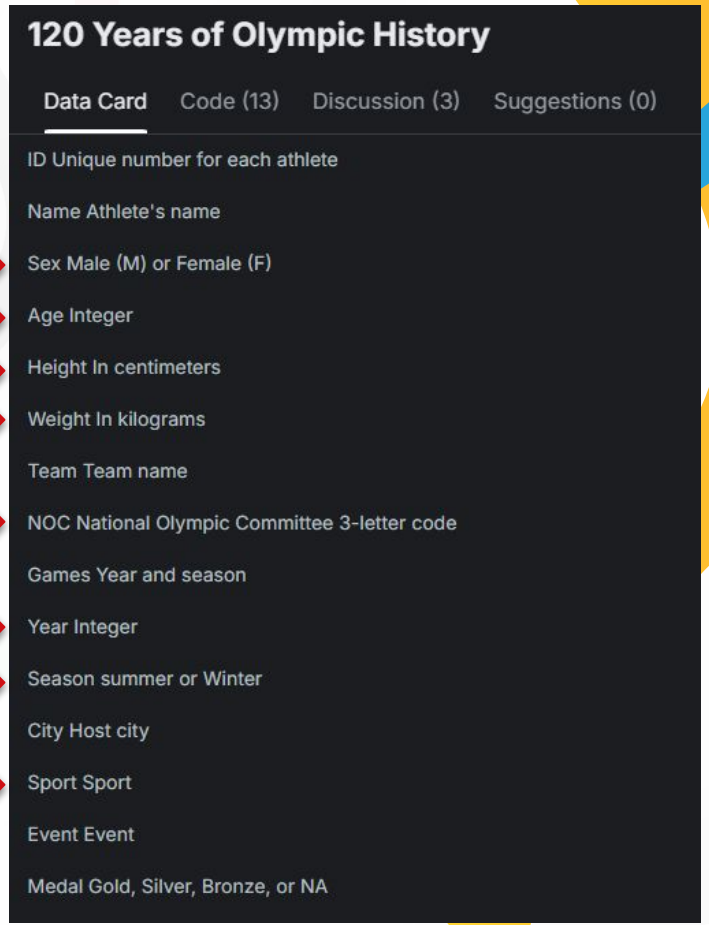
- Design concepts for the web page and ML model
  - Features images from previous Olympic games
  - Colorful visuals that are consistent with theme

Data Engineering

- Over 271k records - focused on 2 Olympic games, 2014 and 2016 - summer and winter, to reduce model size
- Null values for Age, Height, Weight
  - Found median based on Sex and Sport
- Filled in null values for Medal to No Medal
- Dropped unnecessary columns ID, Name, Team, Games, City, Event

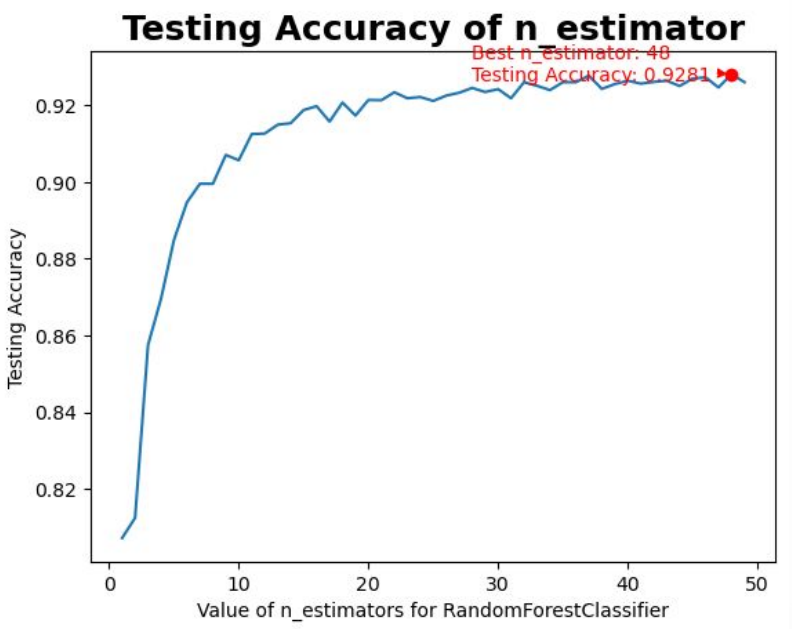
Features to include in model:

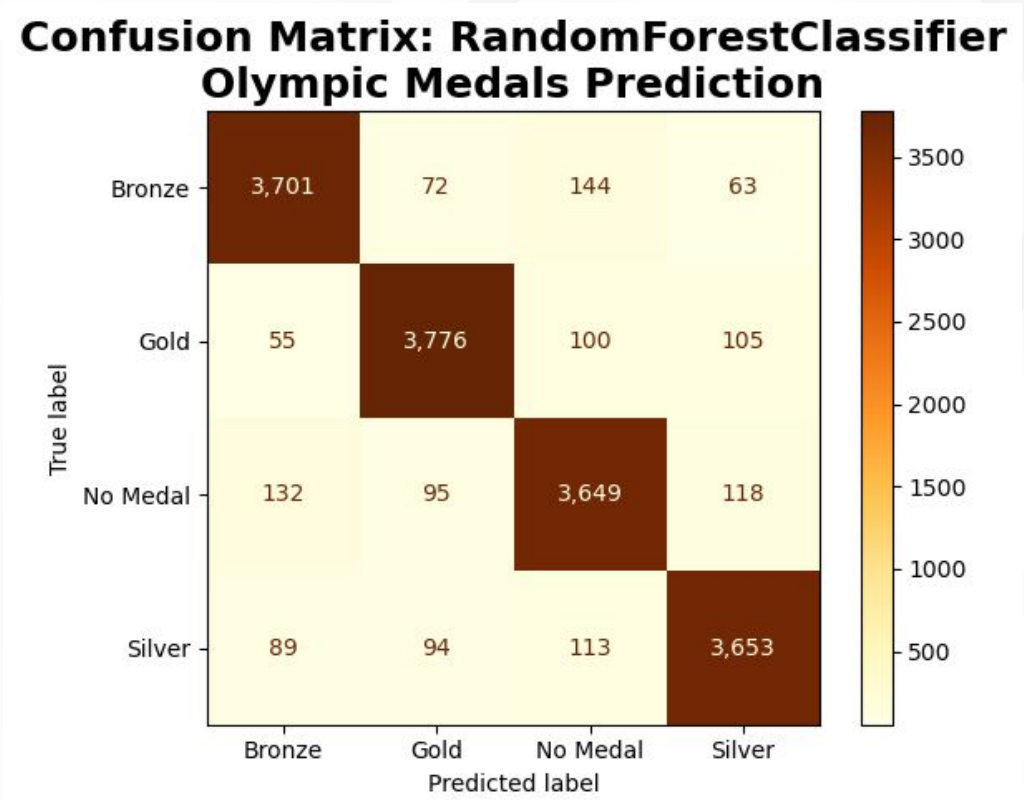
Sex
Age
Height
Weight
Name of Country
Year
Season
Sport



Machine Learning

- Because we multi-classification, we used Random Forest Classifier
- Had to find the optimal n-estimator (decision tree) : 48
- Had to use SMOTE to over-sample because of the unbalanced data
    - Medals were only 15% of the data





Overall accuracy of 93%

Live Demo

https://eugchien.pythonanywhere.com/

Conclusions

- We found that the country that an athlete is competing from was the most important feature in our predictive model.
- We saw that the United States had the highest amount of medals by a significant margin.
- We learned that Track & Field, Gymnastics, and Swimming had the most amount of competitors.

Limitations & Biases

- Dataset was too large - each Olympic game had huge amounts of data which limited the diversification that the model can train on

- There were a lot more Summer sports than Winter sports

Future Work

- Possible features to include
  - Experience (how many years in the sport, or how many Olympics competed in)
  - Injuries reported
- Gather information on coaches' experience as well
- Gather information on host country's climate for outdoor events.
- Cross-reference with census data demographics

Q & A