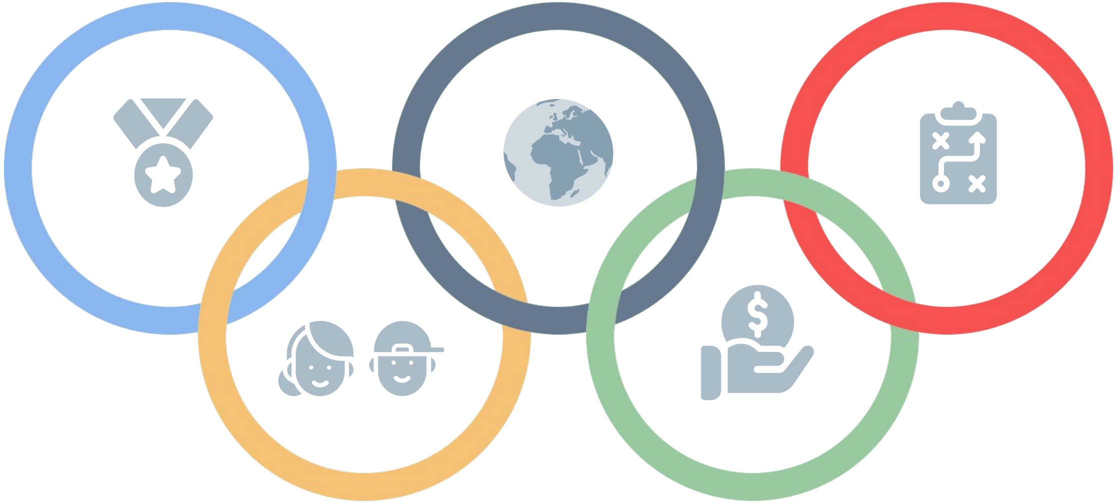
Machine Learning

OLYMPICS
1896 - 2016

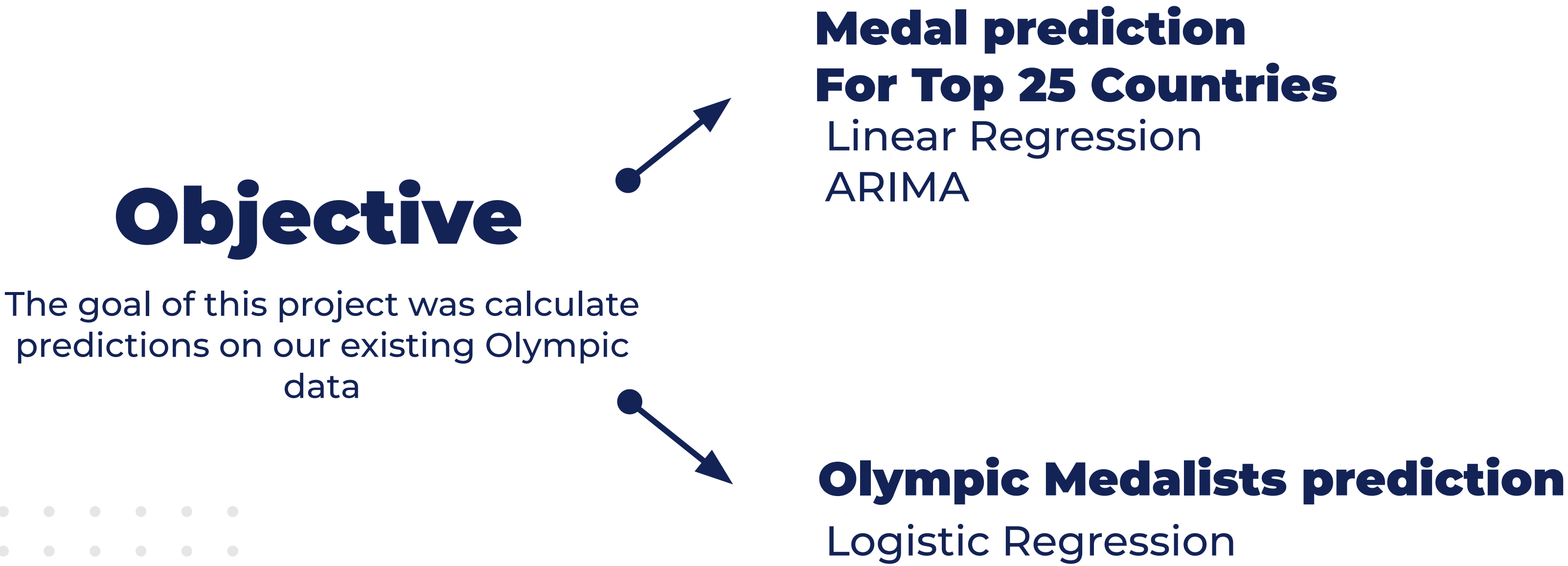
Aastha | Dianne | Duong
Ritika | Swarna



01 **Data Source** - Kaggle Olympics Dataset (Link to dataset)

02 **Tools Used** - Python Pandas, Python Matplotlib, Tableau, HTML/CSS/Bootstrap

03 **ML models** - Linear Regression, ARIMA, Logistic Regression



Objective

The goal of this project was calculate predictions on our existing Olympic data

Medal prediction
For Top 25 Countries

Linear Regression
ARIMA

Olympic Medalists prediction

Logistic Regression

Dataset

Dataset from Kaggle had data for individual athletes (Total Rows: 271,116)

Columns:

- Athlete Name
- Sex
- Age
- Height
- Weight
- Country
- Olympic Year of participation
- Olympic Season (Summer/ Winter)
- Sport
- Event
- Medal (Medal won- Gold, Silver, Bronze or NaN (if medal not won)

Dataset (Data Transformation)

For predicting medals won by a country, data was transformed to show aggregate values country wise for different years

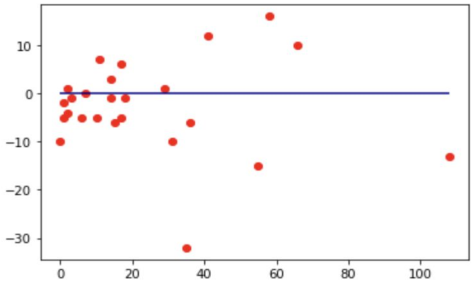
Columns in transformed dataset: (Possible features for ML model- Medal prediction)

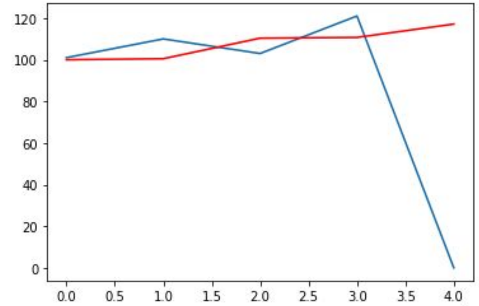
- Year
- Country
- Total medals won
- Gold medal won
- Silver Medals won
- Athletes participated
- Events in which country participated
- Sports in which country participated
- Host (1 if country was host nation for Olympics, 0 otherwise)

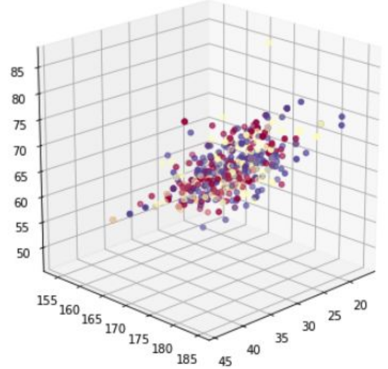
Medal predictions were made for **Summer** Olympics for **25** Countries**.**

Models used

- Linear Regression
- ARIMA Model
- Logistic Regression
- KNN
- Random Forests







Linear Regression Model

Predicted Summer Olympic Gold, Silver, Bronze and Total Medals for the Top **25** countries for the years **2004, 2008, 2012, 2016 and 2020**

Training Set for 2020 Predictions: 'Year' <= 2020

Features: 'Year', ' Host’, 'Athletes', 'Events', ‘Sports’, ‘Athletes/Sport’

R-squared (R2) for Training data:

- Gold: 0.769302651944956
- Silver: 0.7426667887063532
- Bronze: 0.8011844979516292
- Total Medals: 0.8372685664251086

R-squared (R2) for Testing data:

- Gold: 0.7513879821564176
- Silver: 0.7047764327585497
- Bronze: 0.719818759648077
- Total Medals: 0.7704399769927465

Linear Regression Model 1

For USA, data available for years 1896 to 2016

Training Set	'Year' <= 2000
Test Set	'Year' > 2000

Features:
'Year', 'Athletes', 'Event', ' Host'

Prediction **'Total_Medals'**
Mean Squared Error (MSE): 81.60
R-squared (R2): -0.333

Prediction **'Silver'**
Mean Squared Error (MSE): 31.03
R-squared (R2): -0.499

Prediction **'Gold'**
Mean Squared Error (MSE): 28.46
R-squared (R2): -0.138

Prediction **'Bronze'**
Mean Squared Error (MSE): 25.59
R-squared (R2): -0.137

Linear Regression Model 2

For USA, data available for years 1896 to 2016

Training Set	'Year' <= 2000
Test Set	'Year' > 2000

Features:
'Year', 'Athletes per Event', 'Sport', ' Host'

Prediction **'Total_Medals'**
Mean Squared Error (MSE): 48.79
R-squared (R2): 0.202

Prediction **'Silver'**
Mean Squared Error (MSE): 42.33
R-squared (R2): -1.046

Prediction **'Gold'**
Mean Squared Error (MSE): 21.01
R-squared (R2): 0.159

Prediction **'Bronze'**
Mean Squared Error (MSE): 17.67
R-squared (R2): 0.214

Linear Regression Model 3 (Bad)

For USA, data available for years 1896 to 2016

Training Set	'Year' <= 2000
Test Set	'Year' > 2000

Features:
'Year', 'Athletes per Event', 'Participation Event/ Total Events', ' Host'

Prediction **'Total_Medals'**
Mean Squared Error (MSE): 544.95
R-squared (R2): -7.90

Prediction **'Silver'**
Mean Squared Error (MSE): 130.43
R-squared (R2): -5.305

Prediction **'Gold'**
Mean Squared Error (MSE): 142.45
R-squared (R2): -4.69

Prediction **'Bronze'**
Mean Squared Error (MSE): 12.67
R-squared (R2): 0.437

Linear Regression Model 4 (Best)

For USA, data available for years 1896 to 2016

Training Set	'Year' <= 2000
Test Set	'Year' > 2000

Features:
'Year', 'Athletes per Event', ' Host'

Prediction **‘Total_Medals’**
Mean Squared Error (MSE): 38.02
R-squared (R2): 0.378

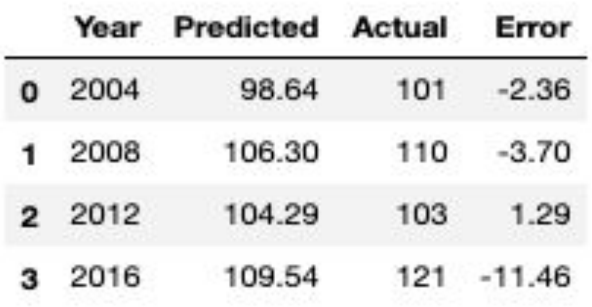
Prediction **‘Silver’**
Mean Squared Error (MSE): 34.85
R-squared (R2): -0.68

Prediction **‘Gold’**
Mean Squared Error (MSE): 20.63
R-squared (R2): 0.17

Prediction **‘Bronze’**
Mean Squared Error (MSE): 15.47
R-squared (R2): 0.312

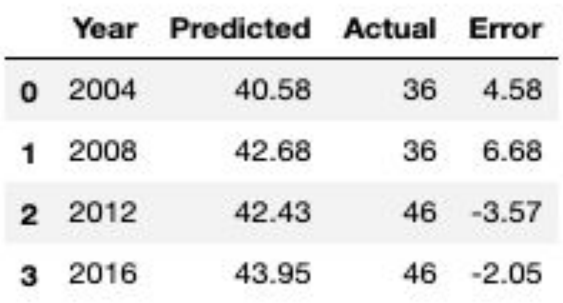
Predictions Model 4

Prediction **'Total_Medals'**



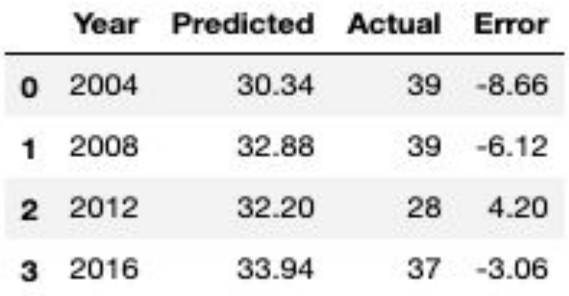
	Year	Predicted	Actual	Error
0	2004	98.64	101	-2.36
1	2008	106.30	110	-3.70
2	2012	104.29	103	1.29
3	2016	109.54	121	-11.46

Prediction **'Gold'**



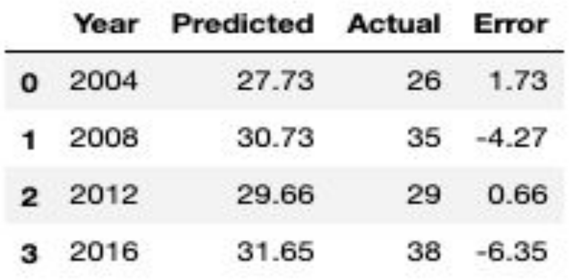
	Year	Predicted	Actual	Error
0	2004	40.58	36	4.58
1	2008	42.68	36	6.68
2	2012	42.43	46	-3.57
3	2016	43.95	46	-2.05

Prediction **'Silver'**



	Year	Predicted	Actual	Error
0	2004	30.34	39	-8.66
1	2008	32.88	39	-6.12
2	2012	32.20	28	4.20
3	2016	33.94	37	-3.06

Prediction **'Bronze'**



	Year	Predicted	Actual	Error
0	2004	27.73	26	1.73
1	2008	30.73	35	-4.27
2	2012	29.66	29	0.66
3	2016	31.65	38	-6.35

Model 4 (Changing Train/ Test Set)

Prediction **‘Total_Medals’**

Train Set 'Year' <= 1988
Test Set 'Year' > 1988
MSE: 589.62
R2 : -6.906

Train Set 'Year' <= 1992
Test Set 'Year' > 1992
MSE: 709.34
R2 : -7.32

Train Set 'Year' <= 1996
Test Set 'Year' > 1996
MSE: 45.43
R2 : 0.542

Train Set 'Year' <= 2000
Test Set 'Year' > 2000
MSE: 38.02
R2 : 0.378

Prediction for 2020: 95.77

Train Set 'Year' <= 2004
Test Set 'Year' > 2004
MSE: 45.04
R2 : 0.179

Prediction for 2020: 96.21

Train Set 'Year' <= 2008
Test Set 'Year' > 2008
MSE: 57.09
R2 : 0.29

Prediction for 2020: 96.75

Total Medal Prediction for 2020: **97.85**
(using data till 2016 in training set and information about U.S. participation in 2020)

ARIMA

An ARIMA model is a class of statistical models for analyzing and forecasting time series data.

ARIMA is an acronym that stands for **AutoRegressive Integrated Moving Average**.

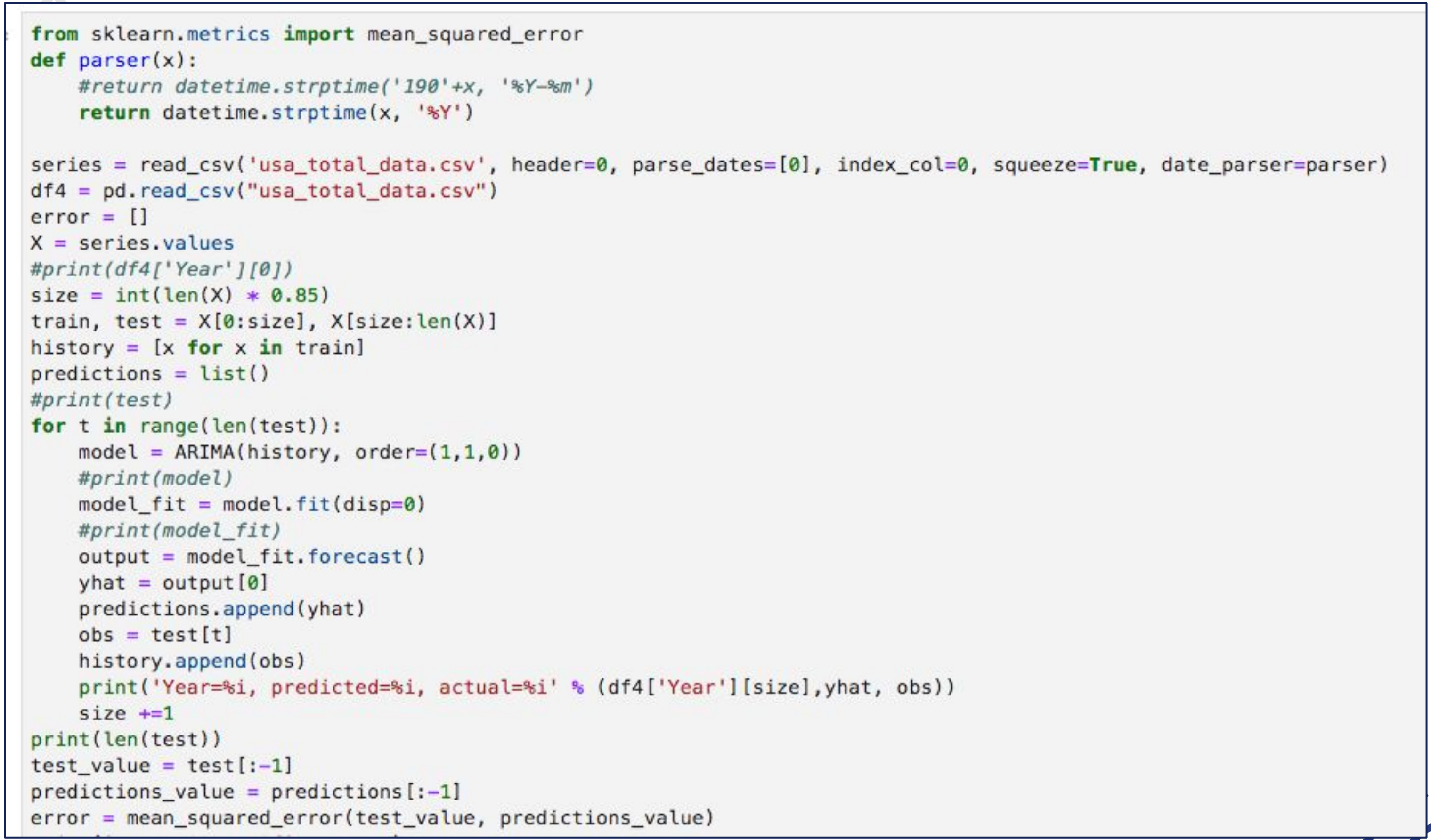
This acronym is descriptive, capturing the key aspects of the model itself.
Briefly, they are:

- **AR**: *Autoregression*. A model that uses the dependent relationship between an observation and some number of lagged observations.

- **I**: *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.

- **MA**: *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

ARIMA



```
from sklearn.metrics import mean_squared_error
def parser(x):
    #return datetime.strptime('190'+x, '%Y-%m')
    return datetime.strptime(x, '%Y')

series = read_csv('usa_total_data.csv', header=0, parse_dates=[0], index_col=0, squeeze=True, date_parser=parser)
df4 = pd.read_csv("usa_total_data.csv")
error = []
X = series.values
#print(df4['Year'][0])
size = int(len(X) * 0.85)
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]
predictions = list()
#print(test)
for t in range(len(test)):
    model = ARIMA(history, order=(1,1,0))
    #print(model)
    model_fit = model.fit(disp=0)
    #print(model_fit)
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    print('Year=%i, predicted=%i, actual=%i' % (df4['Year'][size],yhat, obs))
    size +=1
print(len(test))
test_value = test[:-1]
predictions_value = predictions[:-1]
error = mean_squared_error(test_value, predictions_value)
```

ARIMA Predictions

- ARIMA, short for ‘AutoRegressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values (ref: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)

- We used Gold, Silver, Bronze and total medals won by USA for summer olympics between 1896 to 2016 to predict 2020 medals count for USA.

- Used 85-15 training and test model to predict medal count for 2004 onwards.

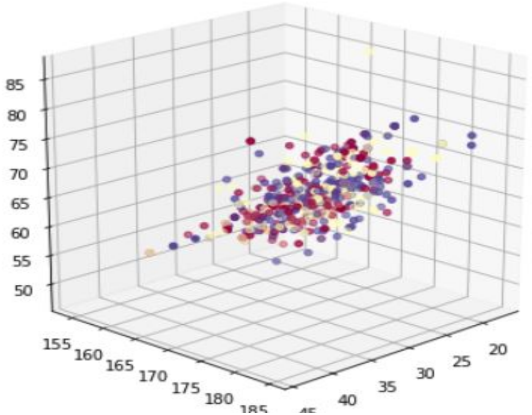
- Compared prediction and actual values (2004-2016) to fine tune ARIMA model by changing dpq values (used 1,1,0) (ref: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/)

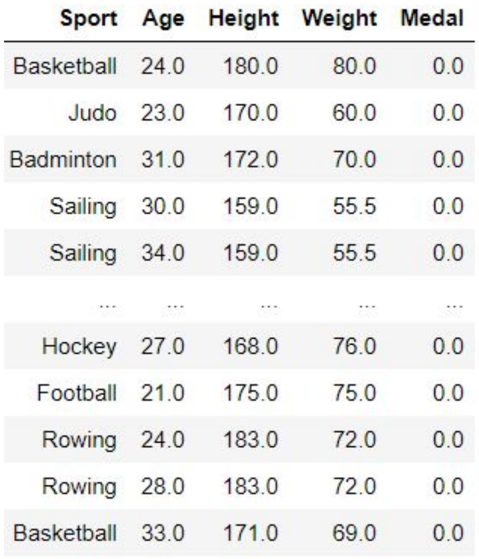
- The MSE value for 2004-2016 predictions are Gold 29.77, Silver - 26.13, Bronze - 21.09 and total - 62.82.

- There are few outliers in gold (1904 - 76 and 1984 - 82). If we replace these values with average gold medal won, predictions for 2004 onwards had better accuracy.

Logistic Regression

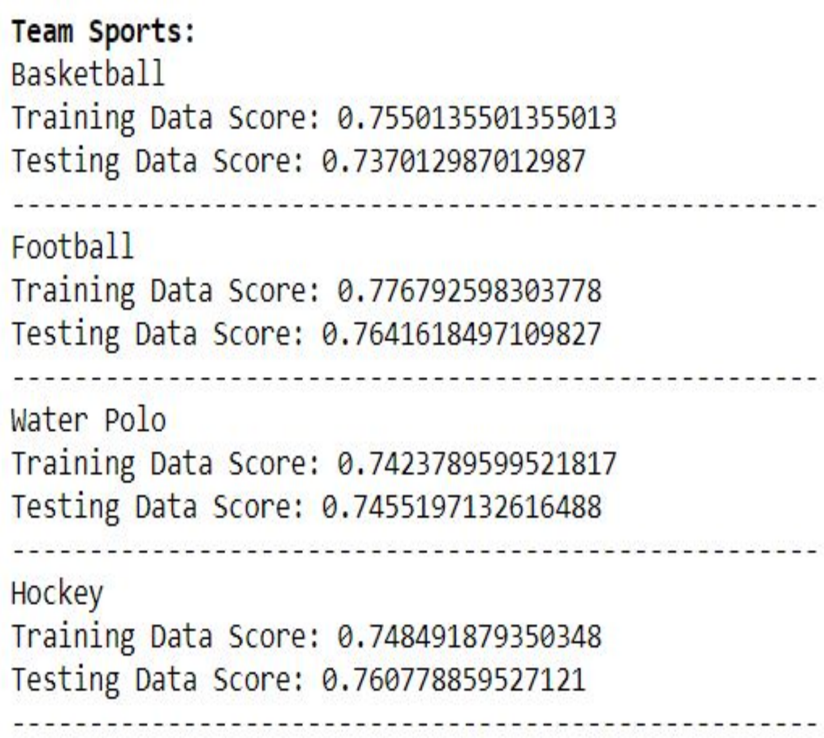
- In sports, the difference between success and failure is having the right body to suit that particular sport.
- Logistic regression is basically a supervised classification algorithm.
- We use the athlete body composition data to predict whether a particular athlete going to win medal in the next Olympic.
- Used 75-25 training and test data of the 1896 - 2016 Olympic.



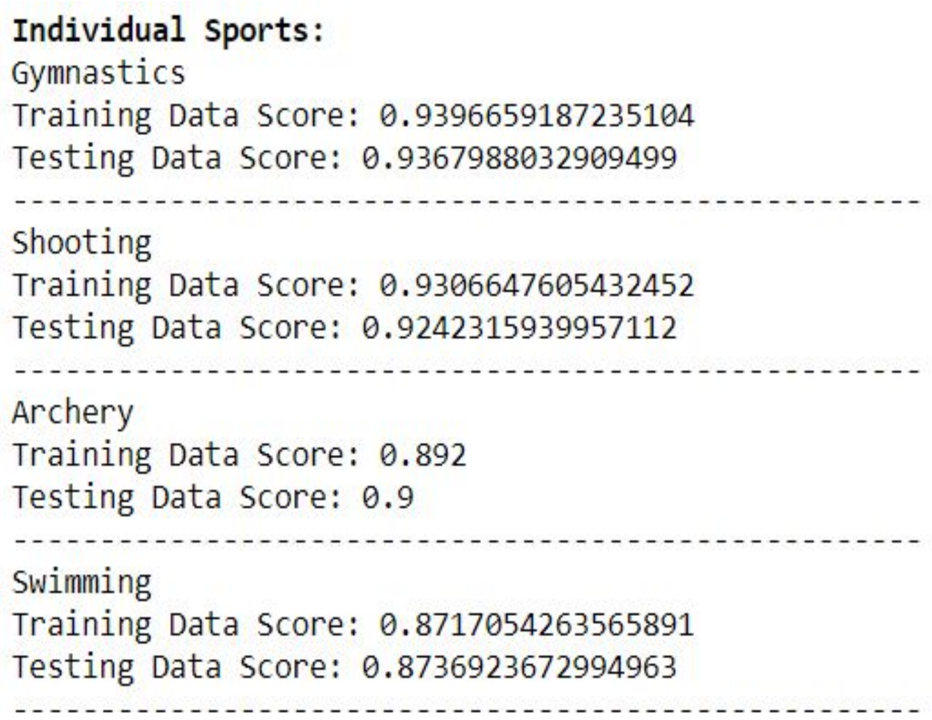


Sport	Age	Height	Weight	Medal
Basketball	24.0	180.0	80.0	0.0
Judo	23.0	170.0	60.0	0.0
Badminton	31.0	172.0	70.0	0.0
Sailing	30.0	159.0	55.5	0.0
Sailing	34.0	159.0	55.5	0.0
...	...	...	...	...
Hockey	27.0	168.0	76.0	0.0
Football	21.0	175.0	75.0	0.0
Rowing	24.0	183.0	72.0	0.0
Rowing	28.0	183.0	72.0	0.0
Basketball	33.0	171.0	69.0	0.0

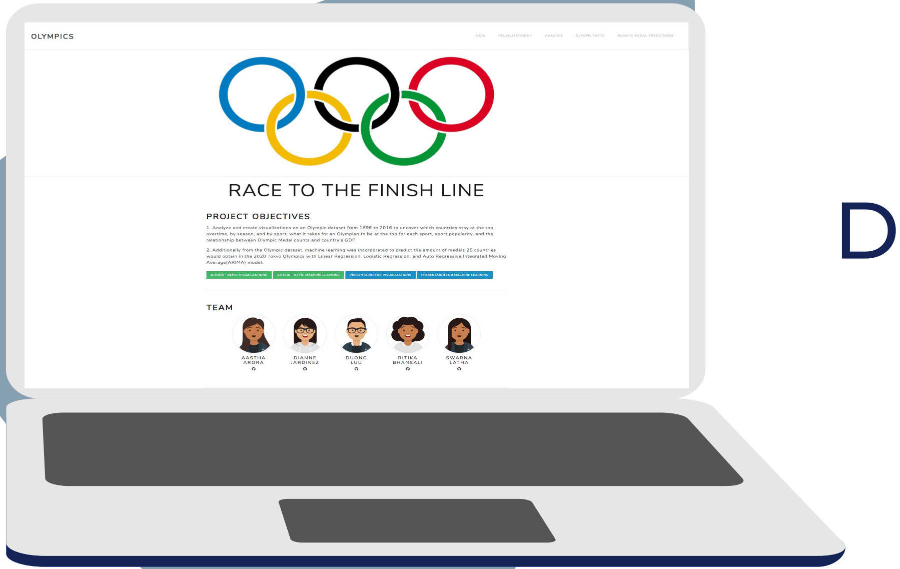
Logistic Regression Predictions



```
Team Sports:
Basketball
Training Data Score: 0.7550135501355013
Testing Data Score: 0.737012987012987
------------------------------------------------------
Football
Training Data Score: 0.776792598303778
Testing Data Score: 0.7641618497109827
------------------------------------------------------
Water Polo
Training Data Score: 0.7423789599521817
Testing Data Score: 0.7455197132616488
------------------------------------------------------
Hockey
Training Data Score: 0.748491879350348
Testing Data Score: 0.760778859527121
------------------------------------------------------
```



```
Individual Sports:
Gymnastics
Training Data Score: 0.9396659187235104
Testing Data Score: 0.9367988032909499
------------------------------------------------------
Shooting
Training Data Score: 0.9306647605432452
Testing Data Score: 0.9242315939957112
------------------------------------------------------
Archery
Training Data Score: 0.892
Testing Data Score: 0.9
------------------------------------------------------
Swimming
Training Data Score: 0.8717054263565891
Testing Data Score: 0.8736923672994963
------------------------------------------------------
```



RACE TO THE FINISH LINE

PROJECT OBJECTIVES

1. Analyze and create visualizations on an Olympic dataset from 1896 to 2016 to uncover which countries stay at the top overtime, by season, and by sport; what it takes for an Olympian to be at the top for each sport, sport popularity, and the relationship between Olympic Medal counts and country's GDP.

2. Additionally from the Olympic dataset, machine learning was incorporated to predict the amount of medals 25 countries would obtain in the 2020 Tokyo Olympics with Linear Regression, Logistic Regression, and Auto Regressive Integrated Moving Average(ARIMA) model.

GITHUB - REPO: VISUALIZATIONS
GITHUB - REPO: MACHINE LEARNING
PRESENTAION FOR VISUALIZATIONS
PRESENTAION FOR MACHINE LEARNING

TEAM

AASTHA ARORA

DIANNE JARDINEZ

DUONG LUU

RITIKA BHANSALI

SWARNA LATHA

DEMONSTRATION **TIME**

ARIMA Vs Linear Regression Vs Logistic Regression

The choice between ARIMA and regression for times series models comes down to a few issues:

- ARIMA generally requires at least 50 data points but > 100 is preferred.
- It is also a rather complex model to estimate and the reliability between experts in determining the right model is very low.
- It is also limited to a single series, unless more complex models are pieced together.
- On the other hand, regression models require as few as 4 observations, the model specification and estimation are much more straightforward, and multiple series can be estimated within the same model.

Limitations

- Linear Regression Model
    - The model which gave the best results for USA could not be used to make predictions for other countries.
    - MSE was high while using the same model for other countries
- ARIMA
    - Could only use one feature to predict
- Logistic Regression
    - Do not have high accuracy rate for the team sports.

Questions?