RUNNING THE NUMBERS

TRAINING NEURAL NETWORKS TO PREDICT MARATHON WINNERS

Group 1:

Jair Solano, Nicole Perez,

Kevin Zhang, Scott Kutlick

October 2024

Overview

Over 1 million people finish marathons per year and that doesn't include the people who sign up and don't finish.

Each marathon runner whether elite or average signs up for their own unique purpose, but, ultimately, everyone wants to do their best.

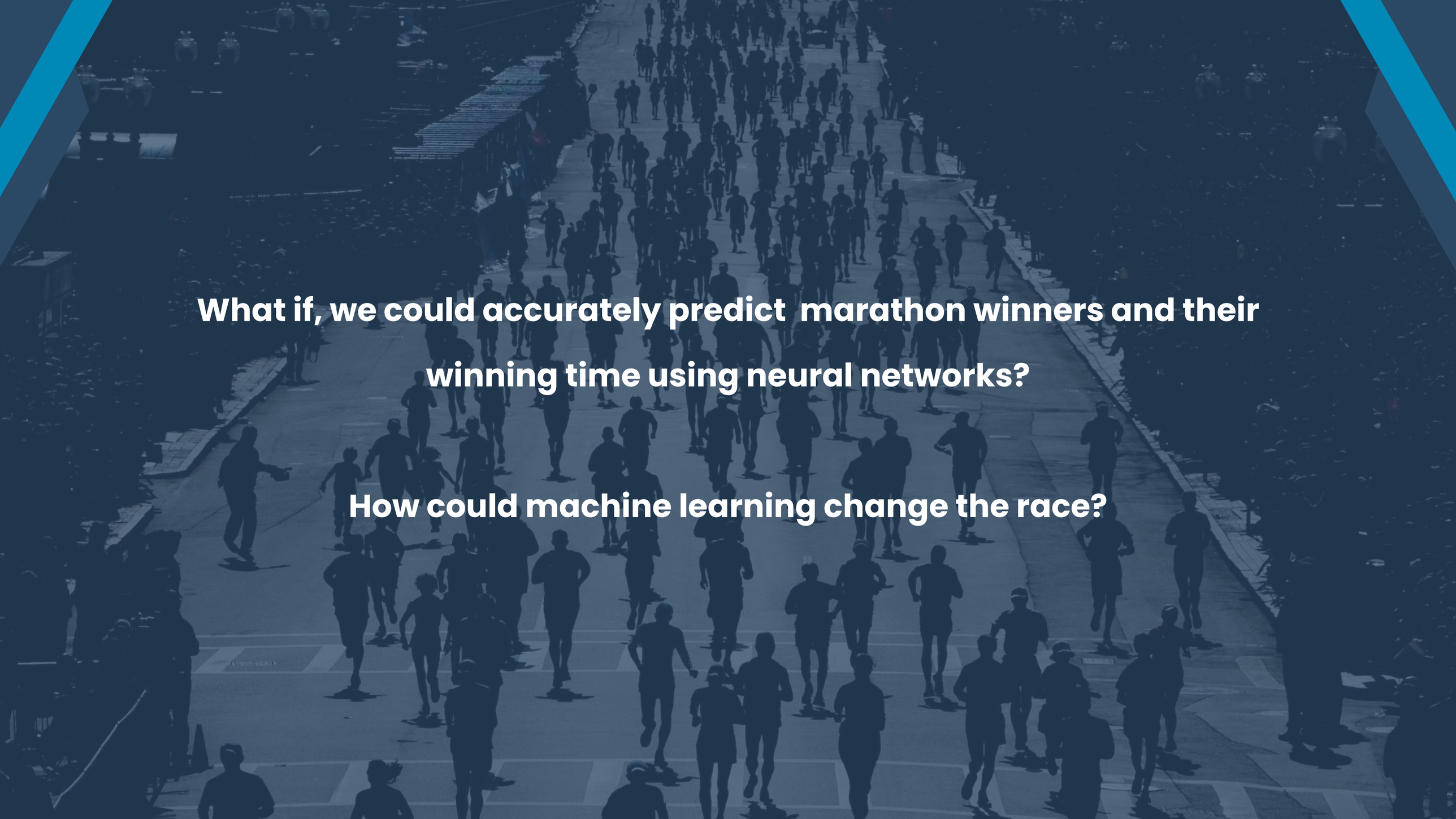
What if machine learning could help them optimize their marathon running experience?

- **The Facts | Considerations**
- **The Question + Theory**
- **The Process + Results**
- **The Insights + Conclusion**

Marathon Facts | Considerations

- **Marathons are 26.2 miles**
- Men’s Average Time: 4:30h
- Women’s Average Time: 4:56h

- **Current Marathon Records**
  - **Chicago Marathon**
    - Men's: 2:00:35 — Kelvin Kiptum
    - Women's: 2:09:56 by Ruth Chepngetich



What if, we could accurately predict marathon winners and their winning time using neural networks?

How could machine learning change the race?

Yes, we can.
We Did.
Here’s what we learned.

Based on comparison graphs and calculated fields of data we observed accuracy close to **1.**

Due to the output being non-binary, we looked at the predication output as a way to gage our prediction accuracy.

Beyond this number, what we observed were the trends and comparisons of predicted time and actual time as well as drew some powerful conclusions for the data we plotted and graphed.

Data that could ultimately change the race.

Our Process

Getting the Winning Numbers and More

Preprocess the data (ETL breakdown)

- Install keras-tuner — import dependencies — read in csv
- Determine the data types - change the marathon finish time from object to seconds using pd.to_datetime()
- Converted categorical variables into indicator variables using pd.get_dummies()
- Separate the data into target and features
- Split the data using train_test_split()

Choosing a Neural Network

- First we graphed the X_train data to determine what the data looks like, this will help in choosing an activation function.
- Looked at the complexity of the data and determined that it may require simpler mathematical computations
- Considered the sparsity of the data once plot
- Finally experimented with different activation functions and density.
- Optimizing the Neural Network
- After initial results we looked at ways to make the model more efficient
- Set 'year' column into 'decade' and converted to indicator variable using get_dummies()
- Used tuner to auto optimization and determine what is the best model

Choosing the correct

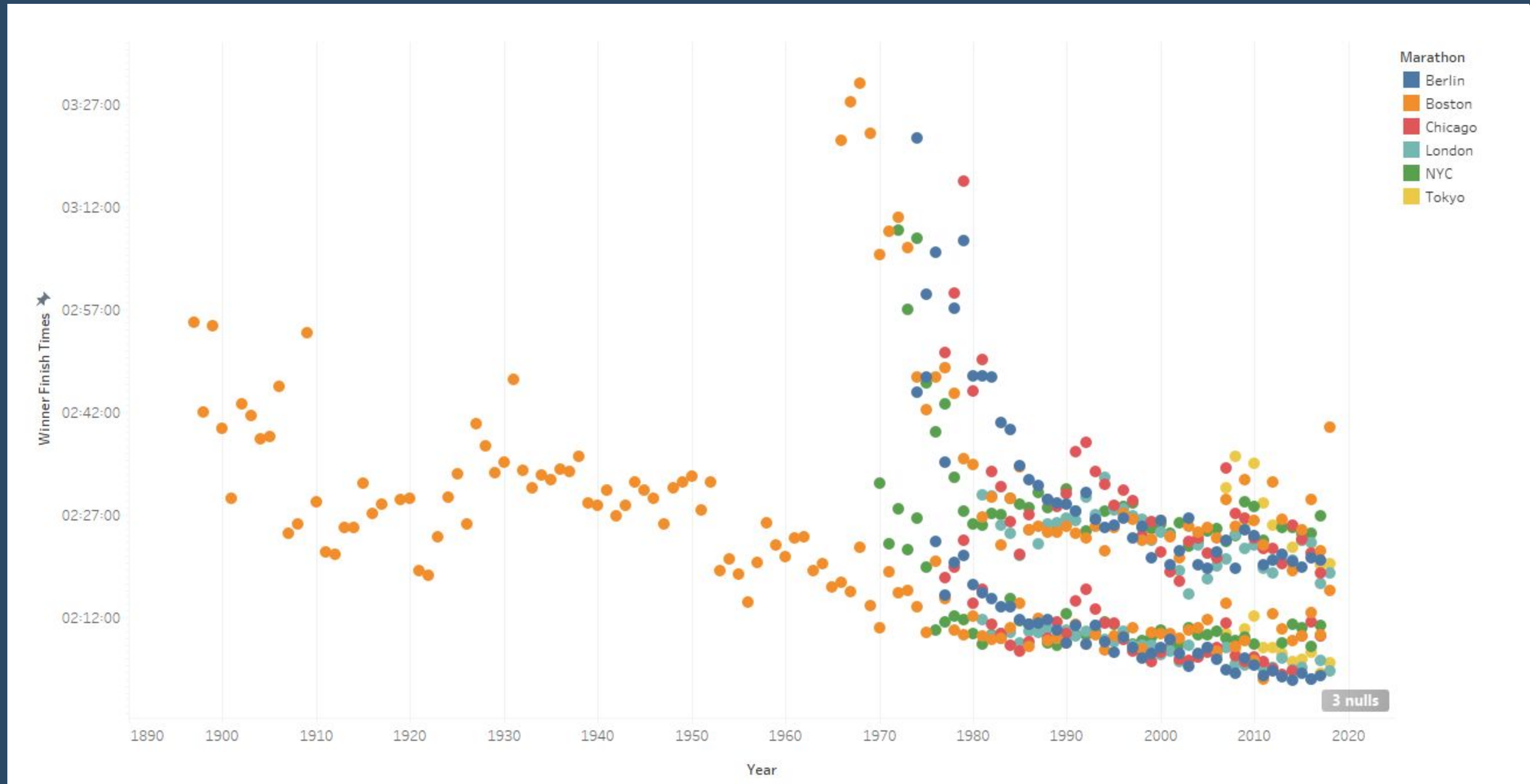
- We added the prediction from out Neural Network Models to the dataframe to compare

Observations/Graphs

- Auto optimization using tuner for hyperparameters did not yield a model that could be used to predict the time

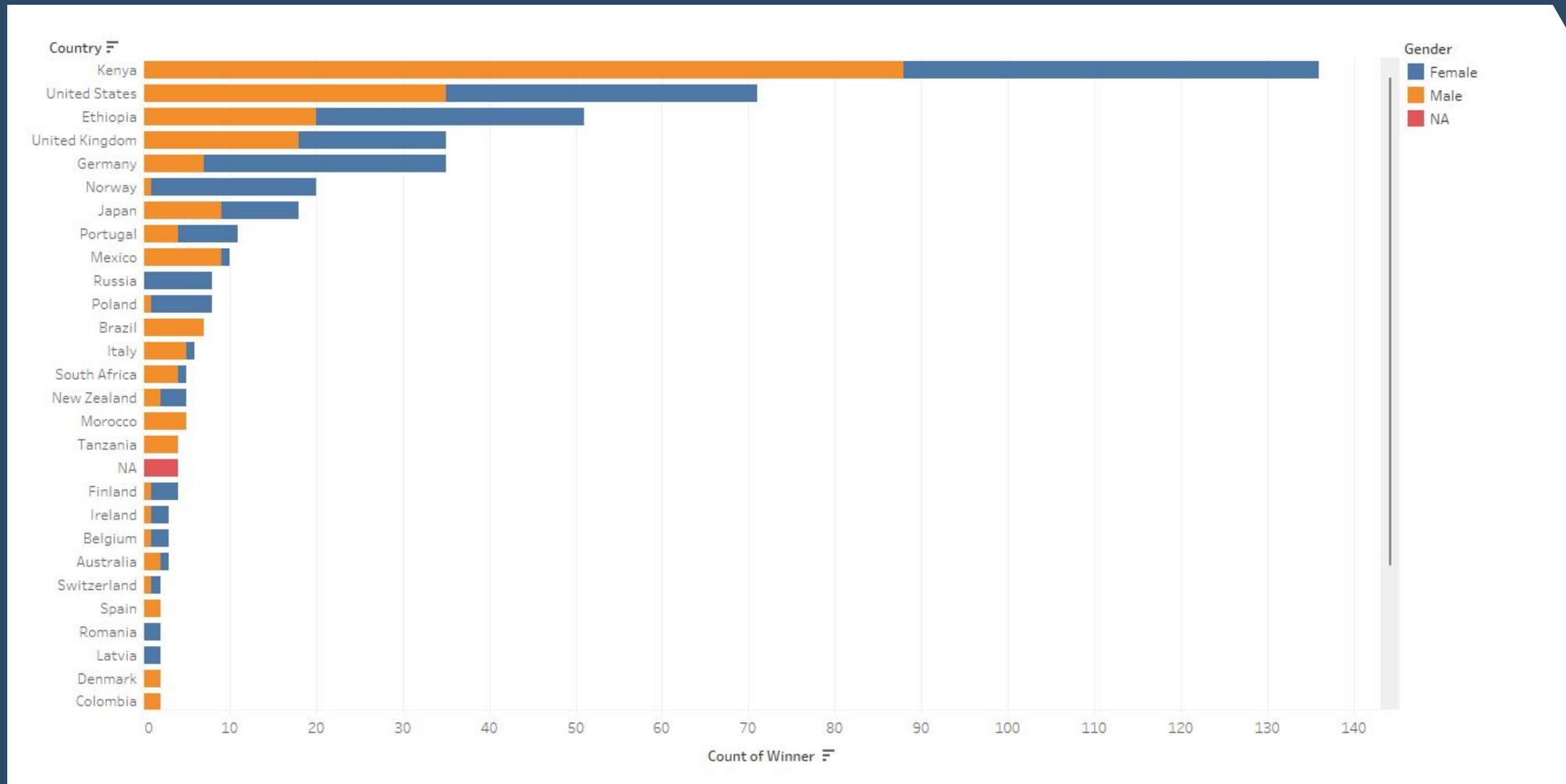
Where We Started - Raw Data

Here you see the raw data from the csv file on a scatter plot using Tableau



Where We Started - Analyzing the Data

Analyze the data to determine an approach for our mission.



Target | Features

The target is the winning time for the marathons and the features are outlined below.

Additional features could be weather, race-day conditions, fatigue index, runner biometrics, and training data.

Country

Using the country feature we can determine which countries have the highest times and most likely to win.

Year

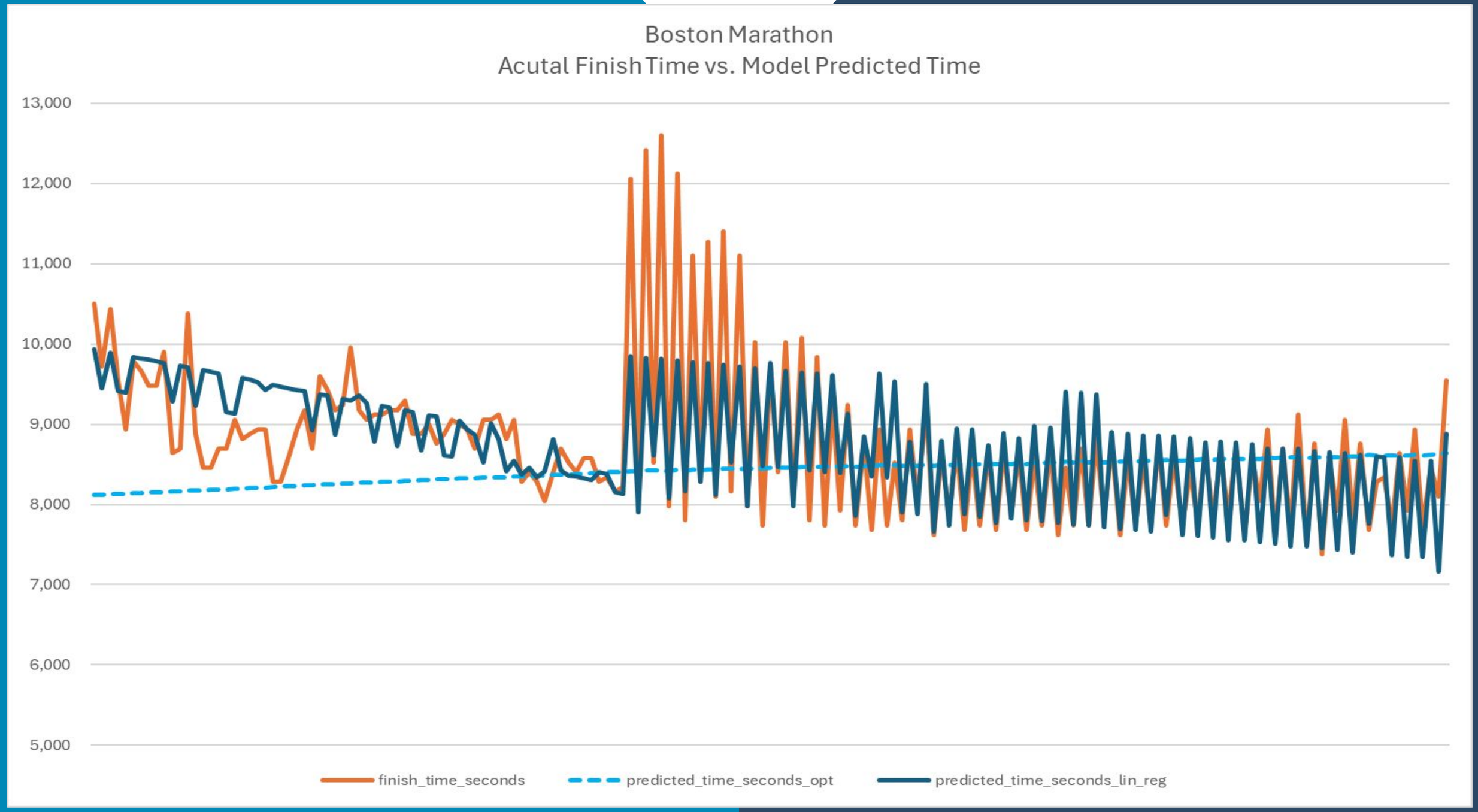
Every year varies based on multiple factors including resources for runners and teh conditions around teh race.

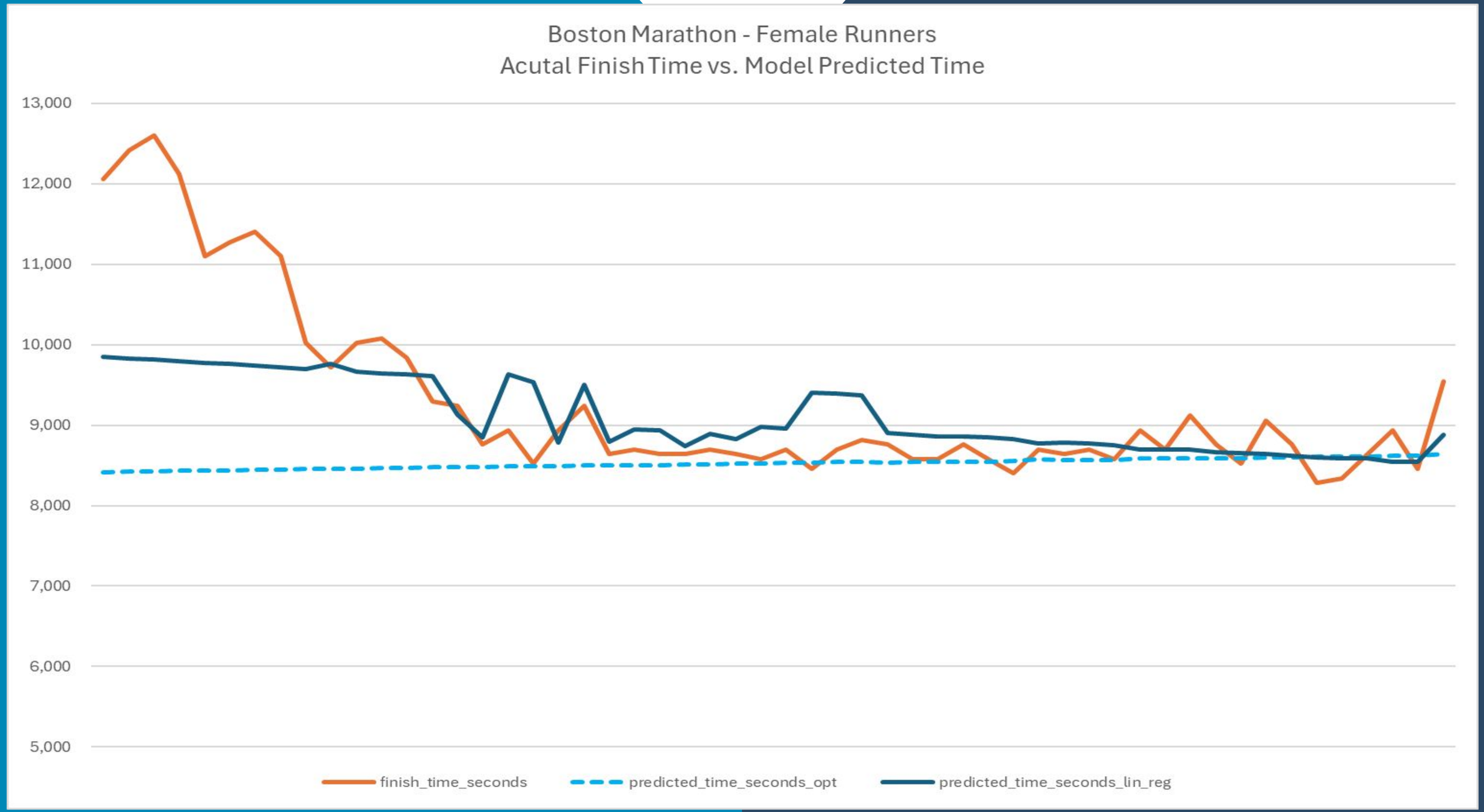
Marathon

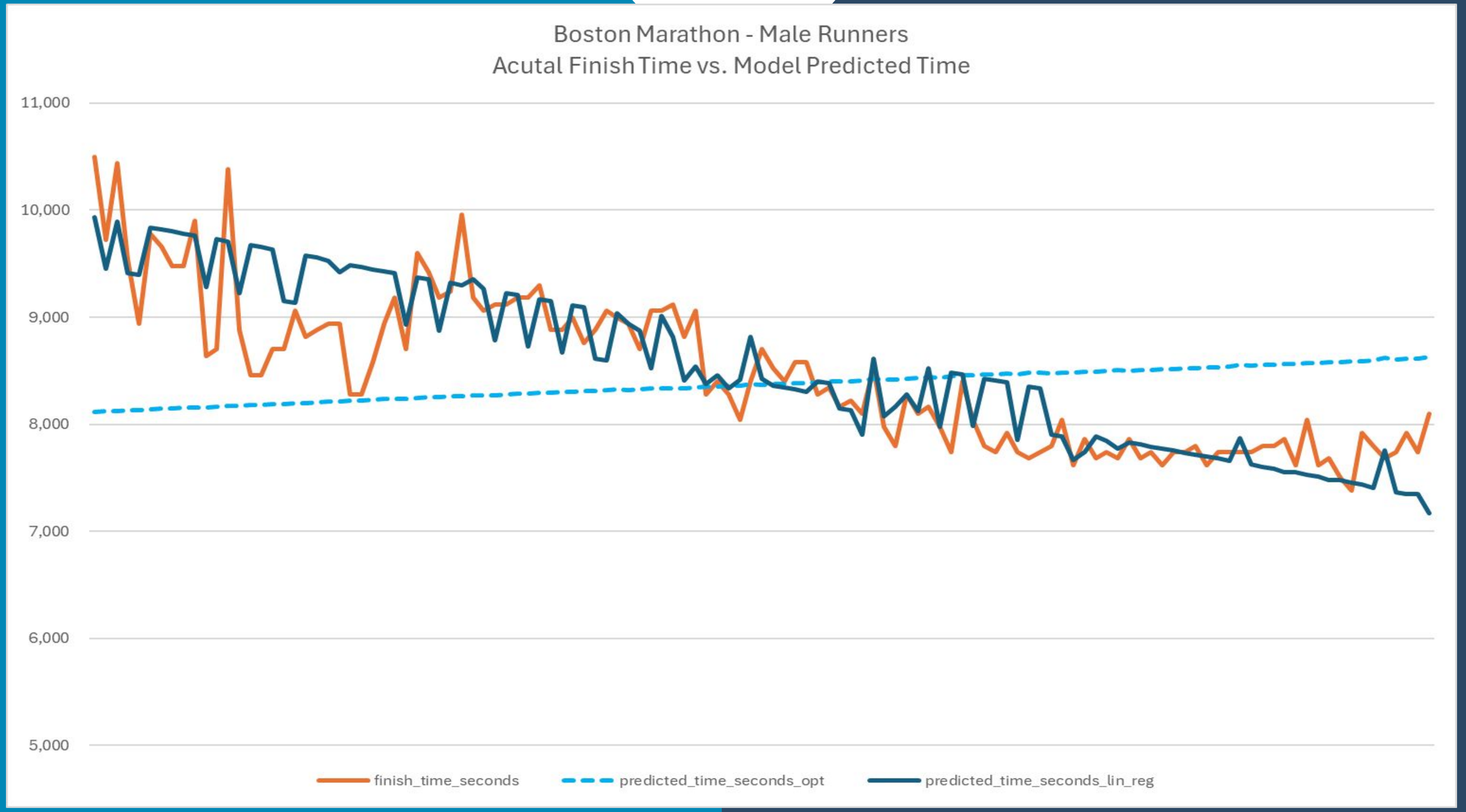
By featuring the different marathons, we can see variability across winning time in the different marathons which have different terrain and weather considerations as well as jet lag for international marathons.

Gender

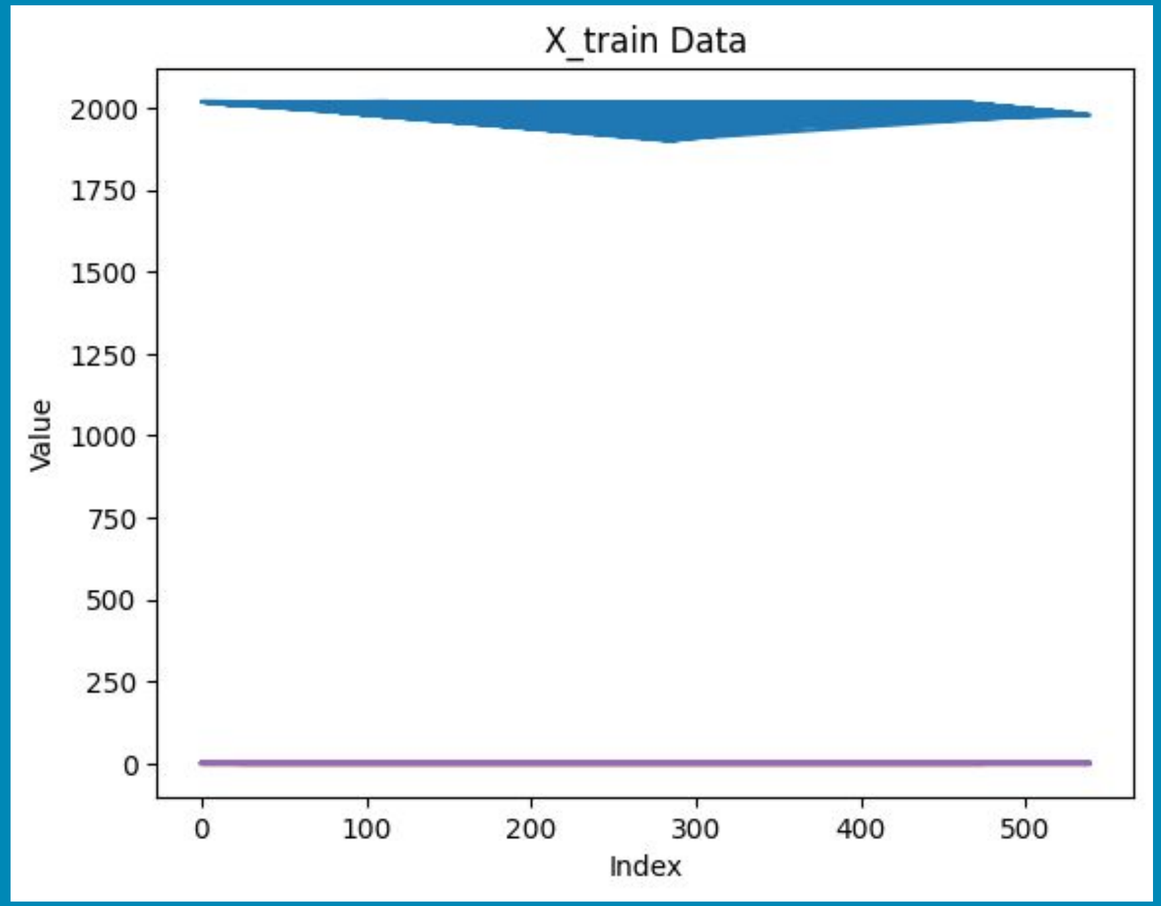
Gender is a consideration as there are different times for men and women,


Boston Marathon
Acutal Finish Time vs. Model Predicted Time
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
finish_time_seconds
predicted_time_seconds_opt
predicted_time_seconds_lin_reg

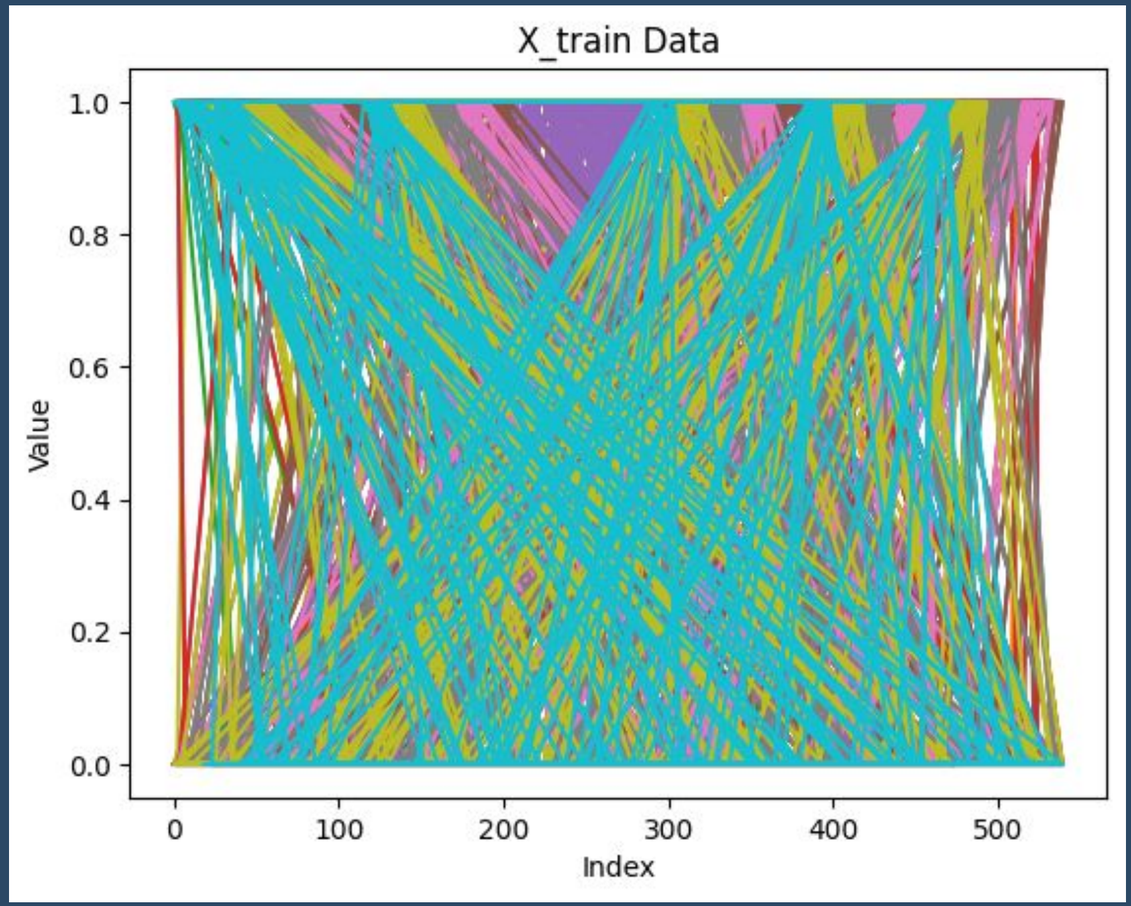

Boston Marathon - Female Runners
Acutal Finish Time vs. Model Predicted Time
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
finish_time_seconds
predicted_time_seconds_opt
predicted_time_seconds_lin_reg


Boston Marathon - Male Runners
Acutal Finish Time vs. Model Predicted Time
11,000
10,000
9,000
8,000
7,000
6,000
5,000
finish_time_seconds
predicted_time_seconds_opt
predicted_time_seconds_lin_reg

X_train Data Split



X_train Data Split Adjusted



Insights | Conclusion

- The “tuner/auto optimization” did not give the results that we looked for
- Despite the r-squared score of the linear regression; the best model was the linear activation
- Overall, the hypothesis of being able to predict the finish time of the marathon with the following vectors:
    - Year
    - Marathon
    - Country of Origin
    - Gender

QUESTIONS?

THANK YOU

FOR YOUR ATTENTION

Group 1:

Jair Solano, Nicole Perez,

Kevin Zhang, Scott Kutlick

October 2024

Appendix