Final Project Technical Report

Bryce Grover, Jason Kim

Introduction

In competitive swimming, the ability to predict future performance can be a very valuable asset for athletes, coaches, and analysts. Accurate forecasting can guide training regimens, optimize performance schedules, and enhance strategic decision-making. This report details a project aimed at developing predictive models for swimming performance using historical Olympic data and personal career data from two athletes (the authors of this report), Bryce and Jason.

The project focuses on three tasks: forecasting future Olympic results, predicting individual athlete performance, and identifying peak performance periods within a year.

Task 1: Forecasting Olympic Winning Times and Teams

Utilizing historical data from Olympic 100-meter freestyle races from 1914 to 2020, the objective is to develop a model that predicts the winning time and the winning team for the next five Olympic Games. The 100-meter freestyle event was chosen due to its highly competitive nature. The event is considered 'deep,' meaning many athletes compete in the event, and it is usually won by very small differences in time.

Task 2: Predicting Individual Athlete Performance

Using the personal swimming career data of Bryce and Jason, the aim is to create a model that forecasts their next 100-yard* freestyle times. There were two reasons for choosing the 100-yard freestyle for prediction. Firstly, like the 100-meter freestyle, the 100-yard freestyle is a highly competitive event, making the predictive model valuable. Secondly, data availability influenced the choice. The 100-yard freestyle is the most commonly swam race for both Jason and Bryce, so it provides sufficient data to train the model effectively.

*Collegiate swimming and Olympic swimming use two different pool formats. Collegiate swimming uses a 25-yard pool, while Olympic swimming uses a 50-meter pool. Bryce and Jason

are collegiate swimmers and thus compete in a 25-yard pool, hence the change in metrics from meters in Task 1 to yards in Task 2.

Task 3: Identifying Peak Performance Periods

Using the same personal career data for Bryce and Jason, a model is designed to predict the month of the year in which Bryce and Jason are likely to achieve their best performance in the 100-yard freestyle. Knowing when an athlete performs best in a given season is valuable knowledge that can optimize training cycles and ultimately lead to greater competitive success.

Analysis

Olympic Data Overview

The dataset, `Olympic_Swimming_Results_1912to2020.csv`, contains historical results of Olympic 100-meter freestyle races from 1914 to 2020. The primary features in the dataset include:

- **Year:** The year of the Olympic event.
- **Gender:** The gender category of the race (Men/Women).
- **Team:** The country represented by the athlete.
- **Stroke:** The swimming stroke used in the race (Freestyle).
- **Distance (in meters):** The distance of the race.
- **Results:** The race times recorded.

Below are the summary statistics.

```
  Location              Year      Distance (in meters)    Stroke
Length:4359        Min.   :1912   Length:4359          Length:4359
Class :character   1st Qu.:1968   Class :character     Class :character
Mode  :character   Median :1988   Mode  :character     Mode  :character
                   Mean   :1983
                   3rd Qu.:2004
                   Max.   :2020
    Relay?          Gender              Team             Athlete
Min.   :0.0000   Length:4359        Length:4359        Length:4359
1st Qu.:0.0000   Class :character   Class :character   Class :character
```

```
Median :0.0000   Mode  :character   Mode  :character   Mode  :character
Mean   :0.1698
3rd Qu.:0.0000
Max.   :1.0000
  Results               Rank
Length:4359        Min.   :0.000
Class :character   1st Qu.:2.000
Mode  :character   Median :4.000
                   Mean   :3.165
                   3rd Qu.:4.000
                   Max.   :5.000
```

Olympic Data Cleaning

A function was implemented to convert the race times from HH:MM:SS.ss format to a SS.ss format for consistent scale of the data. This conversion was essential to ensure all times were comparable and could be used effectively in the analysis. Any entries with non-numeric values in the `Results` column were removed, and the data set was filtered to include only 100-meter freestyle events to focus on a single event type.

Next, additional columns such as `Results_in_seconds` were created to standardize race times, facilitating easier comparison and analysis. Categorical variables (Gender, Team, Stroke, Distance) were OneHotEncoded.

Olympic Data Preprocessing

The features selected for modeling included year, gender, team, stroke, and distance. Categorical variables were encoded using OneHotEncoder. The dataset was then split into 90/10 train-test-split.

The training and testing sets were transformed using the preprocessor to apply scaling and encoding. The transformed data was reshaped to match the input requirements of LSTM models. This preprocessing step ensured that the data was clean and ready for accurate forecasting.

Olympic Data Visualizations

Historical winning times for 100-meter freestyle events were plotted to show trends over time. There is a clear trend of consistent improvement in time.

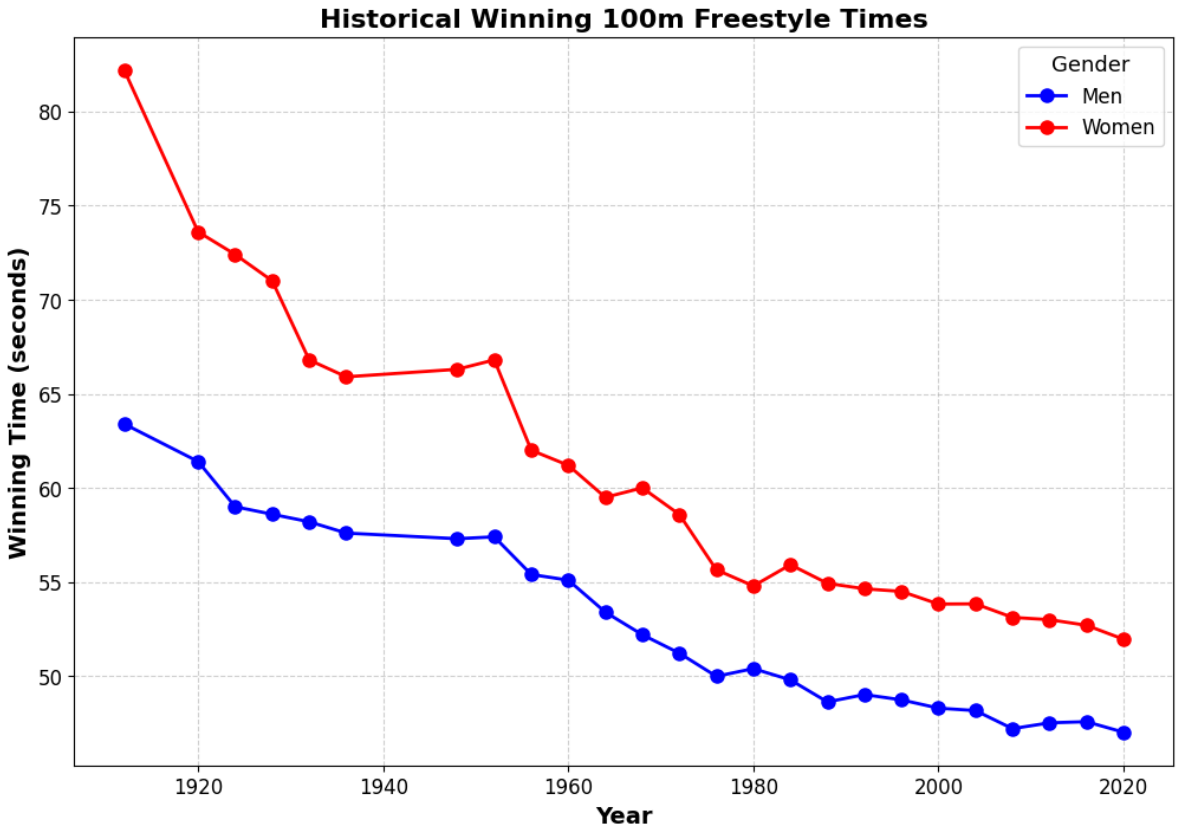

Figure 1: Visualization of historical trends

The performance of selected countries in the 100-meter freestyle over the years was also visualized to compare historical results and highlight trends. There is again, a clear trend of consistent improvement in times.

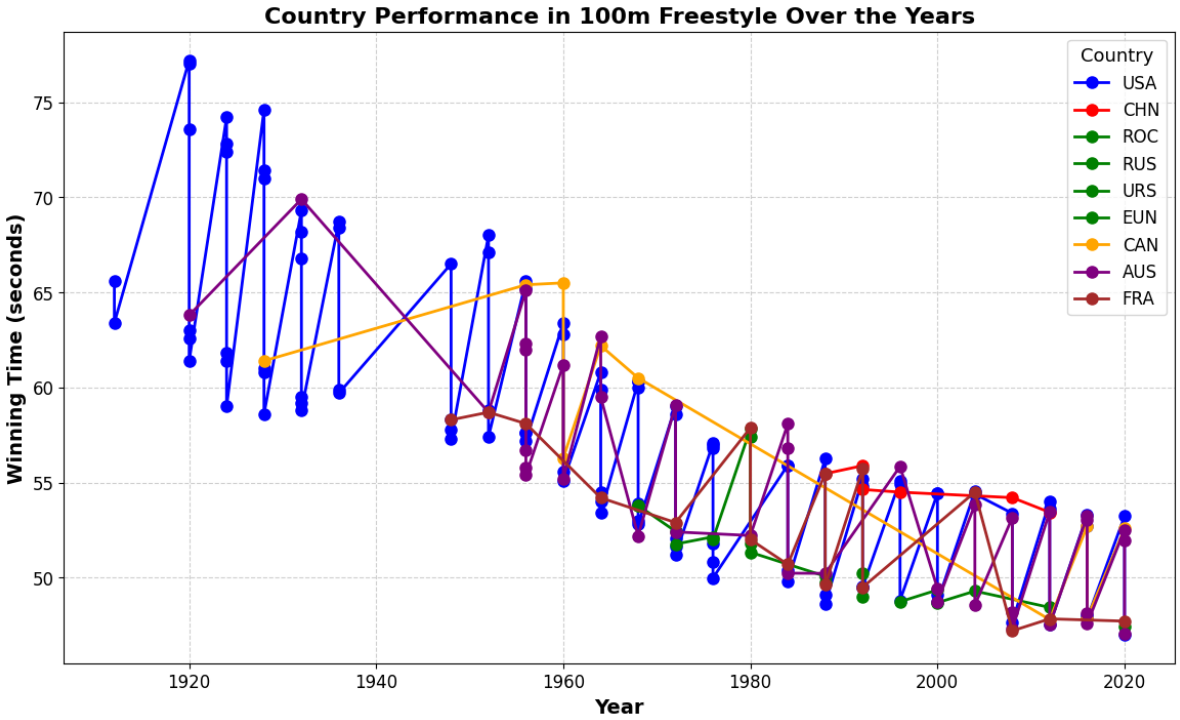

Figure 2: Individual countries' historical performance

Personal Data Overview

The dataset, `MergedData.csv`, contains the personal swimming career data for two athletes, Bryce and Jason. The objective is to create a model that forecasts their next 100-yard freestyle times. The dataset includes features such as:

- **Date:** The date of the performance.
- **Name:** The name of the athlete.
- **Event:** The event name, specifically "100 FR SCY Male" for 100-yard freestyle in short course yards.
- **Time:** The recorded time for the event.
- **Time_in_seconds:** The converted time in seconds for consistency.

Personal Data Cleaning

A function was implemented to convert the race times from HH:MM:SS.ss format to a SS.ss format for consistent scale of the data, just like in Task 1. The `Date` column was also converted to a datetime format to facilitate time series forecasting. The data was filtered to include only the relevant events for Bryce and Jason, ensuring that the analysis was focused on the 100-yard freestyle events, which are the most commonly swum races by both athletes, providing sufficient data for model training. Any entries with non-numeric values in the `Results` column were removed.

Summary statistics were computed to provide insights into the distribution of race times for each athlete. These statistics helped to understand the variability in the performance of each athlete.

```
     Time               Name             Attachment         Meet Name
 Length:1076        Length:1076        Length:1076        Length:1076
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     Date               Event               Cut               Place
 Length:1076        Length:1076        Length:1076        Length:1076
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Personal Data Preprocessing

The preprocessing steps included converting the `Time` values to seconds and normalizing the data. Normalization was essential to scale the data within a range suitable for training neural networks. The MinMaxScaler from scikit-learn was used for this purpose. To prepare the data for training the LSTM models, sequences were created from the time series data. Each sequence consisted of 2 consecutive time steps, which served as the input features, with the subsequent time step as the target variable. This step was necessary to transform the time series data into a format that could be used by the LSTM model.

The exploratory analysis and data preprocessing steps were crucial in preparing the dataset for modeling. By converting, normalizing, and sequencing the data, and by splitting it into training and testing sets, the dataset was made ready for building and evaluating predictive models for the 100-yard freestyle times of Bryce and Jason. These steps ensured that the data was clean, well-structured, and suitable for accurate forecasting.

Personal Data Visualizations

Time series plots were generated to visually inspect the variation in their performance over time. This visual inspection was crucial in identifying trends and patterns in the athletes' performance.

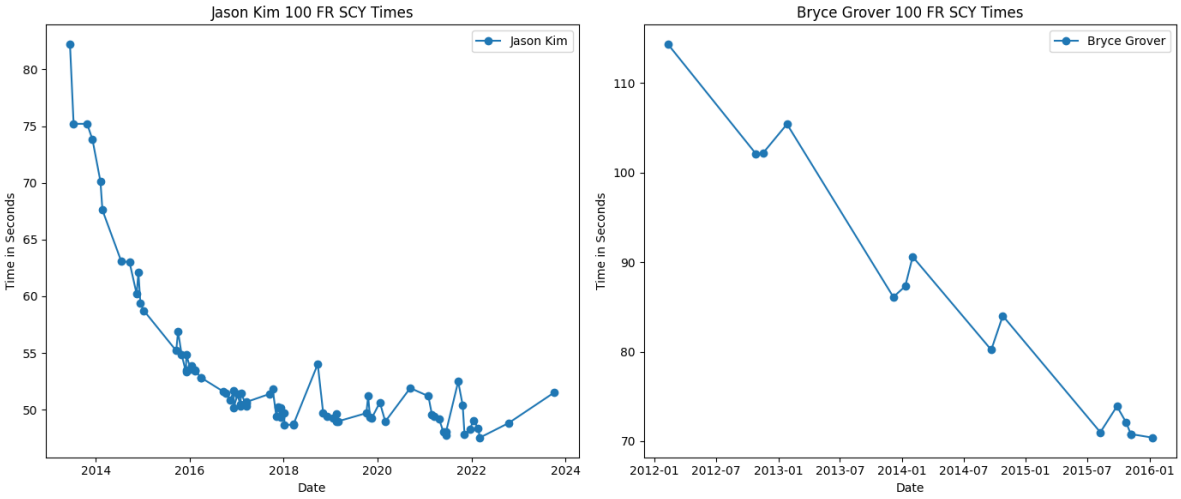

Figure 3: Bryce's and Jason's event history.

Methods

Task 1 LSTM Model

To predict the winning time and team for future Olympic 100-meter freestyle events, a Long Short-Term Memory (LSTM) model was developed. The model architecture included an initial LSTM layer with 128 nodes and L1L2 regularization to mitigate overfitting. Additionally, a dropout layer with a rate of 0.3 was added to prevent overfitting. A second LSTM layer with 64 units and L1L2 regularization was added followed by another dropout layer with a rate of 0.3. The final dense layer had 32 nodes and used ReLU activation function, and a final dense layer with a single unit for regression output.

The model was compiled using the Adam optimizer with a learning rate of 0.0001 and mean squared error as the loss function. Early stopping with a patience of 20 epochs was used to prevent overfitting and restore the best weights. The model was trained for 200 epochs with a batch size of 1, using a validation split from the test set to monitor performance.

Task 2 LSTM Model

Separate LSTM models were developed for Bryce and Jason to predict their next 100-yard freestyle times. Each athlete's data was split into training and testing sets. The model architecture for both athletes included an LSTM layer with 64 nodes and L2 regularization, followed by dropout layers with a rate of 0.5 to prevent overfitting. A dense layer with a single unit was used for regression output.

The models were compiled using the Adam optimizer and mean squared error as the loss function. Early stopping callbacks with a patience of 20 epochs were used to prevent overfitting and restore the best weights.

Task 3 LSTM Model

To predict the month in which Bryce and Jason are likely to achieve their best performance in the 100-yard freestyle, an LSTM model was developed using their personal career data. The model architecture included an LSTM layer with 64 nodes and L2 regularization, followed by a dropout layer with a rate of 0.2. A dense layer with a single unit was used for regression output.

The model was compiled using the Adam optimizer and mean squared error as the loss function. Early stopping with a patience of 30 epochs was employed to avoid overfitting and to restore the best weights.

Results

Task 1 LSTM Model Results

The LSTM model in the first task performed reasonably well.

Its predictions for the winners for the next five Olympic Games are displayed below.

```
2/2 [==============================] - 1s 5ms/step
Predicted winners for the next five Olympics:
    Year Gender Team     Stroke Distance (in meters)  Predicted Time
0   2024    Men  USA  Freestyle                 100m       46.065823
6   2024  Women  USA  Freestyle                 100m       50.146767
12  2028    Men  USA  Freestyle                 100m       45.814133
18  2028  Women  USA  Freestyle                 100m       49.738129
24  2032    Men  USA  Freestyle                 100m       45.577896
30  2032  Women  USA  Freestyle                 100m       49.357090
36  2036    Men  USA  Freestyle                 100m       45.355804
42  2036  Women  USA  Freestyle                 100m       49.001759
48  2040    Men  USA  Freestyle                 100m       45.146637
54  2040  Women  USA  Freestyle                 100m       48.670284
```

Figure 4: Olympic Model's Forecasts

The model's predictions for the men's next immediate race is reasonable. The current Olympic records for men is held by Caeleb Dressel (USA) with a time of 47.02 seconds. As shown, the model predicts that the next Olympic win for men will be a race completed in 46.07 seconds. This is certainly possible.

For the women's races, the next immediate prediction is just as reasonable. The current Olympic records for women is held by Emma McKeon (Australia) with a time of 51.96 seconds. The model predicts that the next Olympic win for women will be a race completed in 50.15 seconds.

The problem with these predictions, however, lies in the nature of records. They can only get faster. If a record in not broken in an Olympic year, the last one stands and is not replaced by a slower time. Because of this, the LSTM model learned to never predict a winning time slower than the previous winning time. This does not reflect the fact that every winning time is nor a record-breaking time.

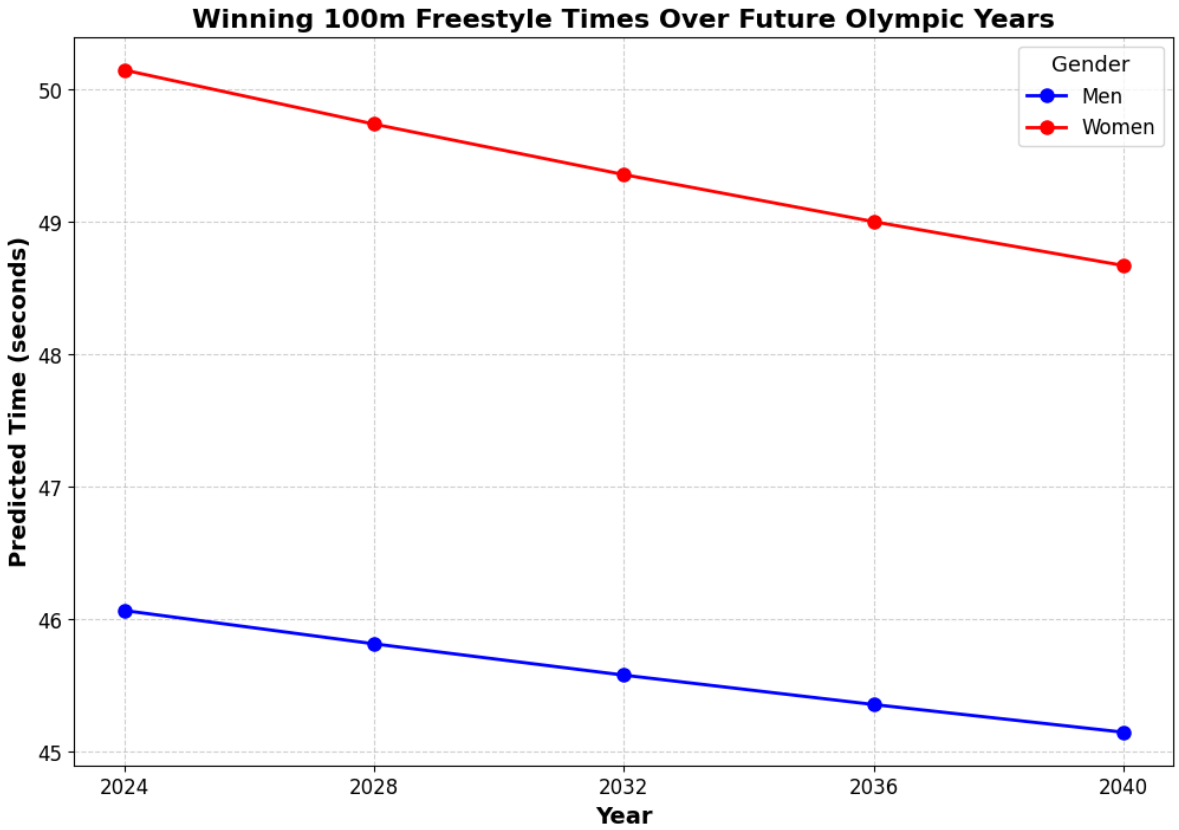

Figure 5: Olympic model's forecasts continue to always improve upon the last

As seen in the above graph, it can be seen that the times for men and women are always improving without fail. This makes the model not particularly useful for making forecasts, into the far future.

Task 2 LSTM Model Results

The LSTM model trained in Task 2 to predict the next 100 yard Freestyle time for Bryce and Jason performed well. Below are the model's race predictions for Jason and Bryce.

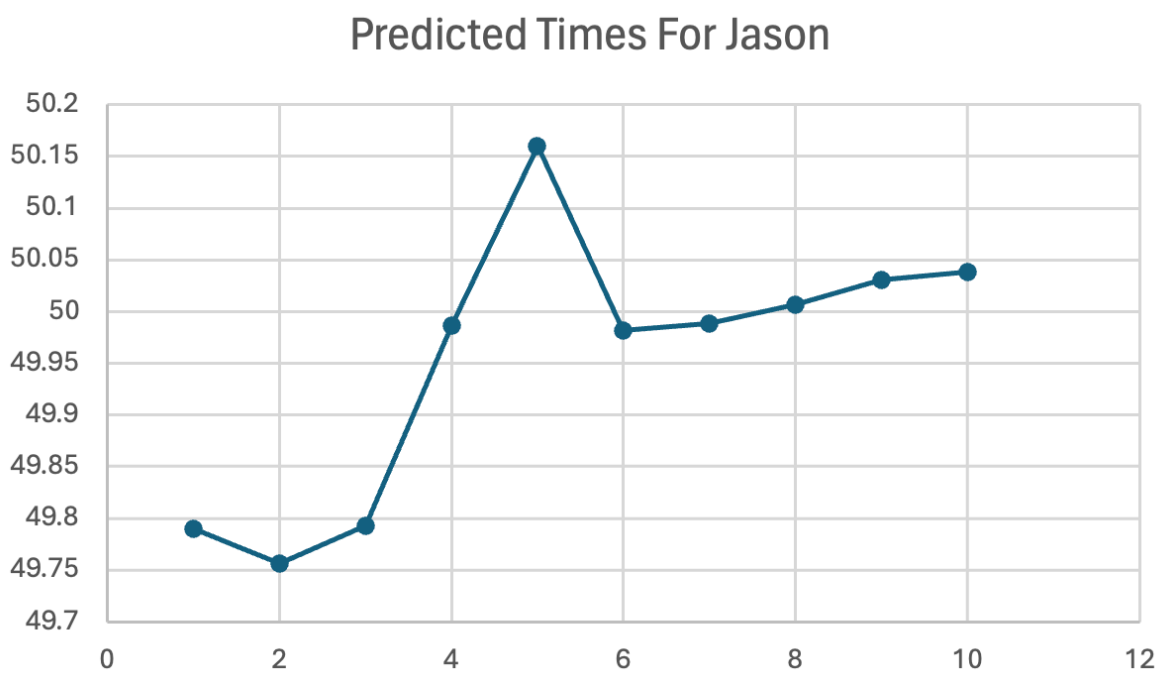

Figure 6: Next 10 time predictions for Jason

The model trained on Jason's data performed well. The average time predicted for Jason was about 50 seconds which is very close to reality.

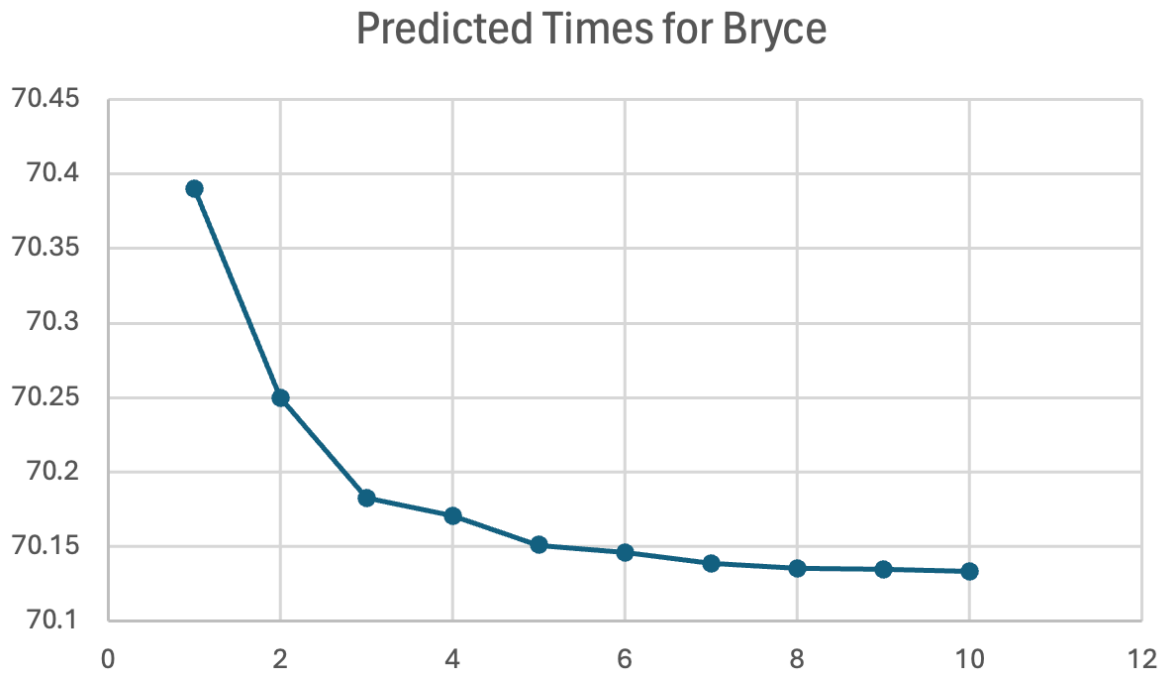

Figure 7: Next 10 time predictions for Bryce

The model trained on Bryce's data did not perform as well as the model trained on Jason's data. This is likley due to a shallow data set that did not have enough data for the model to learn patterns and relationships in the data. The average time predicted for Bryce was about 70 seconds which is not close to realistic time for the 100 yard Freestyle.

The models in Task 2 does not suffer from the same "nature of records" problem that the model in Task 1 does. The data used to train the models contained data for every instance of a 100 yard Freestyle race that Jason and Bryce ever swam, not just their year-on-year current personal record. The model trained on Jason's data predicted he will race faster times, then slower ones, then faster again. This shows that the model learned that not every race results in a faster time.

Task 3 LSTM Model Results

The model in Task 3 yielded good results. The model concluded that the month in which Jason is most likely to perform his best in is September. For Bryce, August.

These conclusions are good ones because across most swimmer who compete collegiately, they see high performance in the fall. This is due, usually, to training over the summer but not competing in any races. The model likely learned that every fall, on paper, Jason and Bryce

suddenly have well-performing races. This is due to the fact that over the rest of the year, the difference in time from one race to another is small because a new official time is clocked every weekend as opposed to the summer where the time is not updated from February until about August (due to the seasonal pattern of swim meets at the collegiate level). All of this is to say, the model learned a useful and accurate relationship between the time of year, and peak swimming performance. Depicted below is a visualization of the model's conclusion.

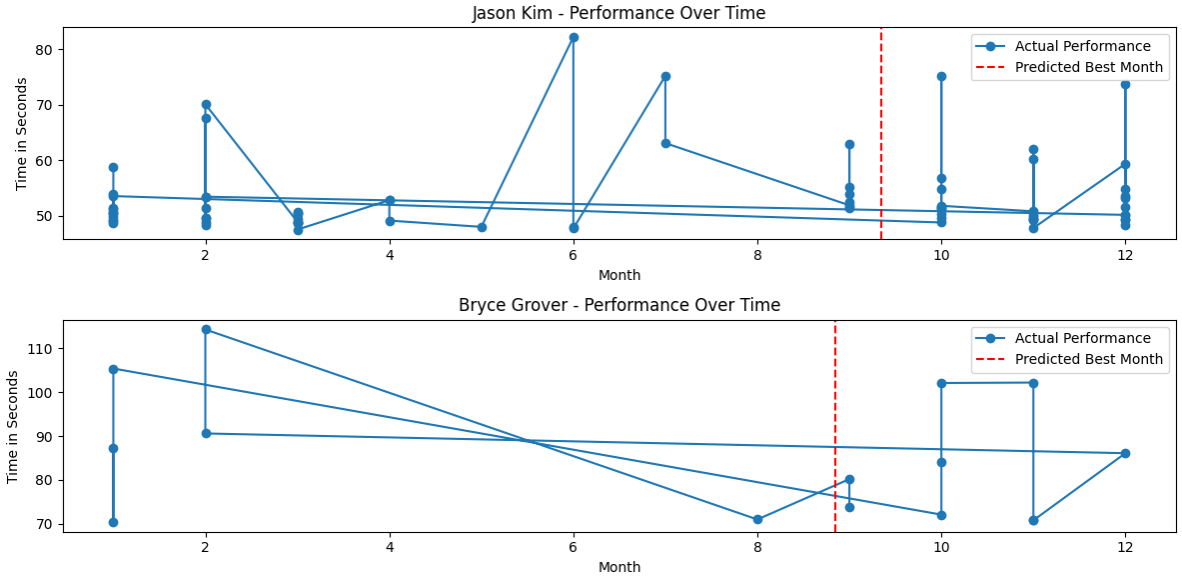

Figure 8: Prediction of the month swimmers Jason and Bryce perform best in.

Reflection

Through this assignment, we learned how valuable the quality of data is. In Task 1, the Olympic model was predicting the next winning time of the 100 meter Freestyle at the Olympics but it was constrained to think the winning time had to be an Olympic record because that is all it was trained on.

Additionally, in Task 2, the model trained on Bryce's data struggled to make realistic predictions because there was not enough historical data for the model to learn meaningful relationships in. However, both model were able to understand that the next race's time did not have to be the fastest one ever swam for either Jason or Bryce because the data reflected instances of races where their times were slower than the preceding ones.

In the future, we would train the Olympic model on a data set containing the winning Olympic times specifically – NOT the standing Olympic record year on year.

Overall, we feel this report describes robust efforts to train and evaluate models that demonstrate out knowledge of using deep learning to forecast future swimming race times.