Forecasting Outcomes in Formula 1 Racing

Erdös Institute Data Science Bootcamp, Summer 2024
https://github.com/Xaalthe71/Formula1_Forecasting_Erdos2024

Ali Arslanhan, Ryan Bausback, Edward Voskanian

Overview

This project identifies factors from Formula 1 practice sessions that influence the qualifying round and explores predicting qualifying times and grid positions with practice session data.

Dataset: https://pitwall.app/

In each of the three practice sessions, we use the best lap time, the gap to the fastest lap time (0 if you are the quickest), and the number of laps completed. Initially, we split the data into training and testing sets, reserving 15% for final validation.

Modeling

We applied linear regression, random forest NN, and KNN models to predict final grid positions and fastest lap times in the first qualifying session (Q1). For features we used the best lap time, the gap to the fastest lap, and the number of laps completed in each practice session.

- Linear Regression and KNN (Q1 Time): Five-fold cross-validation on a baseline model providing the average time in Q1, followed by linear regression models and a KNN model incorporating all possible combinations of features. The KNN model over FP1 Time, FP1 Gap, FP2 Time, FP3, Time, and FP3 laps, performed the best during final validation, achieving a mean squared error (MSE) of 16.80.

- Linear Regression (Grid Position): Linear regression model to predict drivers' final grid positions, This providing an initial benchmark for grid position predictions. This model achieved an MSE of 21.46 with $R^2 = 0.427$

- Random Forest (Grid Position): The Random Forest model was employed to predict the final grid positions using the same practice session features. This ensemble learning method improved prediction accuracy by averaging the results of multiple decision trees, leading to more robust and reliable predictions. The random forest achieved an MSE of 13.45 with $R^2 = 0.64$.

- Fully-Connected Neural Network (Grid Position): Aimed to capture complex patterns in the data that other models might miss, this network consisted of three hidden layers with ReLU activation functions and a softmax output layer. The dimensions were [9, 64, 64, 64, 20]. Two accuracy metrics were used:
    a. exact grid position
    b. within two of exact grid position

  Training accuracy: 20/44% (exact/within two)
  Testing Accuracy: 12/33% (exact/within two)

Conclusion

Predicting Formula 1 starting grids is a challenging task due to the dynamic nature of the sport and the numerous factors influencing performance. Our models, particularly the Random Forest and Fully-Connected Neural Network, showed promising results. Our present models serve as good baselines for the more sophisticated models. Future work will involve incorporating additional features such as weather conditions and experimenting with more advanced modeling techniques to enhance prediction accuracy.