F1oresight: Predicting Final Position and Race Winners in Formula 1 Using Random Forest and Historical Performance Data

Sandra Hu
University of Toronto Schools
husa@utschools.ca

Jessica Gu
University of Toronto Schools
guje@utschools.ca

Diptarko Ghanti
University of Toronto Schools
ghadi@utschools.ca

***Abstract*—Formula One (F1) is a highly data-intensive motorsport where race outcomes are influenced by driver skill, car performance, team strategy, and track-specific factors. Predicting race results is challenging due to the inherent unpredictability of mechanical failures, weather conditions, and in-race decisions. F1oresight, a machine-learning model, models the relationship between qualifying performance, recent driver consistency, and race outcomes. Using historical race and qualifying data collected via the FastF1 API, features such as average position over the last three races and grid advantage were engineered to capture recent driver performance and track effects. Random Forest models were trained to predict finishing positions. By isolating stable performance indicators from unpredictable race-day variables, the models provide insights into the drivers' raw pace and expected outcomes. Results demonstrate that machine learning can effectively estimate race rankings and winning probabilities.**

I. Introduction

Formula One is the highest level of international single-seater motorsport, where teams compete across global circuits in a season-long championship. Race outcomes depend on driver skill, car performance, team strategy, and track conditions.

Machine learning (ML) techniques have demonstrated strong performance in structured prediction tasks across sports domains, including match outcome prediction and player performance forecasting. Unlike traditional statistical approaches, ML models can capture nonlinear relationships between variables such as team performance, driver consistency, and starting grid position. In motorsport, qualifying position has been shown to be a strong indicator of race performance, yet race outcomes remain uncertain due to variability in driver skill, team strategy, reliability, and track-specific characteristics [2].

Building on this relationship, F1oresight, applies machine learning techniques to model the connection between qualifying performance, recent driver consistency, and race outcomes. By training predictive models on historical race data, the project aims to estimate finishing positions and the probability of victory for the top drivers.

A. Motivation

In recent years, Formula 1 (F1) has evolved into one of the most data-intensive sports in the world. Each race weekend generates extensive structured data. As the availability of historical race datasets has expanded, predictive modeling has become an increasingly important tool in sports analytics for understanding performance trends and forecasting competitive outcomes [4].

However, many existing approaches attempt to explicitly model highly unpredictable race-day events, which can introduce additional uncertainty into predictions. As a result, isolating the influence of more stable performance indicators, such as qualifying pace and recent driver consistency, remains an area that warrants further investigation [3].

This project investigates whether machine learning techniques can model the relationship between qualifying performance, recent driver consistency, and race outcomes in Formula 1. Using historical race and qualifying data, Random Forest models are trained to predict both finishing positions and the probability of a driver winning a race.

B. Related Works

In modern Formula One strategy analysis, Monte Carlo simulation methods are commonly used to model uncertainty in race conditions. By randomizing variables such as tire degradation, safety car probability, and pit lane loss time, teams can generate thousands of simulated race scenarios to evaluate the robustness of different strategic decisions. However, implementing a full stochastic race simulation requires detailed telemetry inputs and extensive computational modeling.

Machine learning methods such as Random Forest and other ensemble models have been applied to motorsport datasets to predict race outcomes and driver performance, often outperforming traditional regression-based approaches in predictive accuracy [1]. However, race outcomes remain partially unpredictable due to factors such as race incidents, mechanical failures, and strategic variability.

C. Problem Definition

Trying to predict Formula 1 race results can be a struggle because of the sheer amount of unpredictable variables that can affect a driver's final position in a race. From mechanical failures, dramatic weather shifts, to seemingly spontaneous strategy calls, the fastest driver on the track often finds themselves off the podium from factors outside of their control. In an attempt to account for as many factors as possible, many traditional Formula 1 predictors use Elo rating systems that reward an individual driver's ability to overcome

unpredictable circumstances or Monte Carlo simulations that calculate the variables into probabilities for an algorithm. This project attempts to bypass calculating the unpredictable by investigating the high correlation between a driver's qualifying performance and the final points distribution of specific tracks. By modeling the non-linear relationships between a driver's qualifying performance, their historical consistency, and race outcomes, this project predicts both the expected ranks and the win probabilities of the top contenders. In doing so, raw pace is isolated from the influence of major factors like starting position to provide a cleaner way of evaluating the drivers.

II. Methodology

To provide room to expand later and for ease of maintenance and updates, the process of building the prediction system follows this simple structure:

1) Data collection
2) Preprocessing data
3) Feature engineering
4) Model training

A. Data collection

Telemetry data and historical Formula 1 race results were collected from the FastF1 API, a Python library. For each racing event, the race session and the qualifying session results were retrieved. For each of the drivers recorded to have participated in that event, their team, their grid position, final position, qualifying position, and status were all recorded along with the year, the circuit, and the circuit's ID as well. These data points were only taken from race weekends that followed conventional race formats with available, completed results to standardize the amount and quality of the data that was pulled.

B. Preprocessing data

Machine models are unable to understand the raw data, so the collected data was preprocessed to turn all the information into numerical inputs. The names of the drivers and the team names were encoded using Label Encoding, while circuit identifiers were changed to numeric indices using factorization. This way, the data could be mathematically operated on and specific track characteristics that influence the outcomes of races could be taken into account.

C. Feature engineering

In order to help with pattern recognition and to account for recent performances, the average position over the last three races (AvgPositionLastThree = rollingmean(FinalPosition, 3)) and the grid advantage (GridAdvantage = QualifyingPosition - Grid Position) were both calculated. This way, the recent performance of a driver would be taken into account along with grid penalties changes that could have occurred between qualifying and race day. Forward fill and backward fill methods were used when values were missing from DNS or DNQ, ensuring data continuity. StandardScaler was used to normalize all numerical features and to improve the stability of the model training.

D. Model training

Using the Random Forest algorithm, two machine learning models were trained, the Finishing Position Prediction Model using regression and the Race Winner Prediction Model using classification. Random Forest algorithms are skilled at resisting overfitting and have a strong ability to capture non-linear relationships, which is well suited for this project. Since predicting the final positions in a race can be solved using regression, a Random Forest Regressor was implemented with 100 decision trees for balance, a maximum depth of 10 levels to prevent overfitting, and a random state of 42 as a fixed seed. Because winning a race can be boiled down into a binary classification problem of winning for one or not winning for zero, a Random Forest Classifier was trained to calculate the probability of a driver taking pole position. The maximum depth was reduced to a conservative 5 levels. Since each driver's probability of winning is calculated in a percentage, they can be ranked by their chance of taking pole position. Since the dataset is relatively small, 80 percent was used for training while 20 percent was used for testing. The output values are rounded for interpretation.

III. Results

Three methods were used to evaluate the quality of the model's prediction. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used for the regression model that calculated the finishing position prediction, and accuracy was used for the binary classification model that calculated the probability that drivers would win a race. The equation for MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

where i is the index of the driver, n is the total number of race samples, y_i is the final position for the driver, and $\hat{y}_i$ is the predicted final position for the driver. It reveals the magnitude of error. The equation for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

By squaring the error before finding the average, large outliers become more visible. This gives an indication of how consistent the model performs.

The RMSE was calculated to be 3.18. This value indicates that though the model has demonstrated general consistency, it is not immune to deviations. These deviations can be explained by the unpredictable factors in motorsports like mechanical failure or shocking strategy calls.

The evaluation of the regressor showed that there was a significant gap between the training error at around 1.3 to 1.4 MAE and the validation error at around 3.2 MAE.

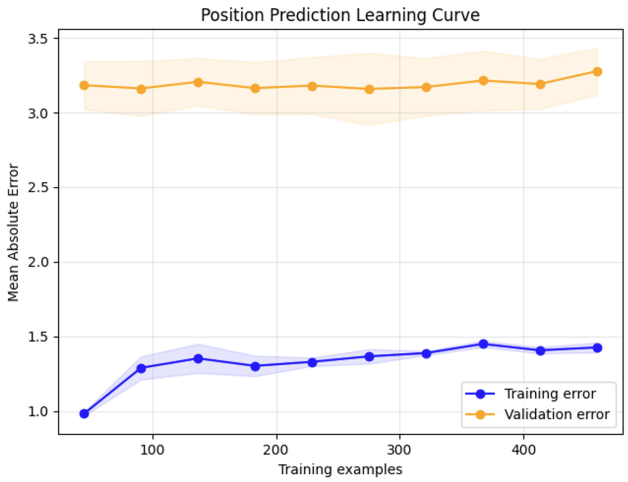

Fig. 1. Plot of the regressor position prediction learning curve. It was evaluated using MAE and RMSE.

This difference is an indication of overfitting in the model. The training error being lower than the validation by approximately 3 places implies the model is struggling to understand new races and that there is seemingly a limit to the model's ability. The validation error running at around a consistent 3.2 races could be from the unpredictable variables that come with racing.

The classifier's evaluation revealed a validation accuracy of around 94 percent to 96 percent.

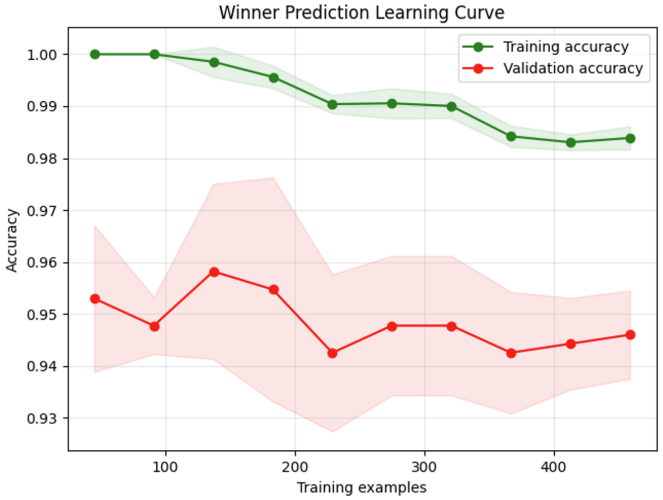

Fig. 2. Plot of the classifier winner prediction learning curve. It was evaluated using accuracy.

The training accuracy ranged from around 98 percent to 100 percent. The downward trend suggests that the overfitting issue decreases over time.

IV. Conclusion

Overall, this project shows potential for evaluating the quality of a driver, but also reveals limitations to using machine learning to predict results in motorsports. The current model heavily relies on data from qualifying and starting position. Adding more factors like weather or tire strategy would allow for the model to take into account more unpredictable factors and improve the accuracy of the predictions. This would help with decreasing the MAE. There is also a chance of data leakage occurring with this current model if it uses future data to predict past performance.

This project is reliant on the idea that the correlation between a driver's qualifying result, their most recent performances, their team, and the circuit that the race takes place on is strong enough to accurately train a machine learning model to predict the final positions of the drivers and the likelihood of their victory. By modelling the non-linear relationships between these factors and implementing a Random Forest Regressor and Classifier, the model was successfully able to predict the victor of a race to a validation accuracy of around 95 percent and a training accuracy of around 99 percent. However, when predicting the final positions of drivers, there was a consistent MAE of around 3.2 positions and an RMSE of 3.18. This indicates that unpredictable factors in Formula 1 play a significant role in deciding the final position and point distribution. As motorsports become increasingly popular and as technology grows more and more refined, it is likely that previously intuition and knowledge based strategies will shift to become more data centric. As this project has demonstrated, machine learning has a strong potential to aid teams in evaluating their drivers, performances, and predicting their results.

References

[1] E. El Haber et al., "Formula 1 race winner prediction using random forest and SHAP analysis," *Proceedings of the 2025 International Conference on Control, Automation, and Instrumentation (IC2AI)*, pp. 1270–1274, 2025.

[2] T. Muehlbauer, "Relationship between Starting and Finishing Position in Formula One Car Races," *International Journal of Performance Analysis in Sport*, vol. 10, no. 2, pp. 98–102, 2010.

[3] V. S. S, R. Coelho, and L. Chandrashekar, "Optimum racing: A F1 strategy predictor using reinforcement learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 4, pp. 5800–5805, 2025.

[4] A. Urdhwareshe, "The Use of Machine Learning in Predicting Formula 1 Race Outcomes," *Preprints*, 2025.