Simulating 2025 F1 Lap Times With a Random Forest Approach Based on Historical Race Data

STA 160: Practice in Statistical Data Science (Spring 2025)

Department of Statistics | University Of California, Davis

Professor Fushing Hsieh

Howard Bui (920136355) | howbui@ucdavis.edu

7 June 2025

1 Introduction

1.1 What is Formula 1?

Formula 1 (F1), often referred to as the pinnacle of motorsport, is the highest class of racing for open-wheel racing cars. F1 is known for its high-speed races, cutting-edge technology, and its drivers who come from different countries. Each season includes a series of races called Grand Prix, where drivers compete for points in both driver and constructors (teams) championships, and are held on circuits (race tracks) in various countries around the world. These circuits take place on two main types of circuits: closed-track circuits, which are built specifically for racing, and street circuits, which are made by temporarily closing public roads in a host city (Formula 1, 2024).

1.2 Problem Statement and Motivation

This study seeks to evaluate the predictive capability of statistical learning models with respect to F1 lap times, focusing specifically on circuits featured in the 2025 season. Motivated by the interaction between driver and constructor capability, and track characteristics, a regression-based predictive framework was developed by using historical data from the Kaggle *Formula 1 Lap Time Prediction* competition, which spans races from 1996 to 2023. The key question addressed in this study is: **Can a driver's lap time be accurately predicted based on relevant features on circuit, driver, team, and race conditions?**

1.3 Factors That Influence Lap Performance

1.3.1 Technical and Strategic Contributors

Lap performance in Formula 1 is influenced by a complex set of technical and strategic variables that determines how quickly a driver completes a lap, which includes the characteristics of the track itself, as well as the driver and constructor's ability to manage tire wear, fuel usage, and on-track decisions (Mugge & Zandieh, 2024). Specific tracks such as Autodromo Nazionale di Monza yield faster lap time due to long straight segments, whereas technical circuits such as Circuit de Monaco tends to produce slower lap times due to its tight corners, narrow turns, and limited overtaking zones (F1StatBlog, 2016). In addition to circuit layout, lap performance is also influenced by the driver's consistency and skill, particularly in managing tire degradation, fuel consumption, and in-race decision-making. Car and constructor performance are also key contributors, with competitive teams typically producing faster and more reliable laps due to engine power, aerodynamic efficiency, and technical setup. Further variability arises from fuel load and pit stop timing. Early laps are slower due to heavier fuel load, and later laps quicken as the car becomes lighter. On the other hand, pit stops introduce necessary time losses, making their timing a strategic factor that affects lap outcomes.

1.3.2 Why Lap Time Prediction Matters in Context

Given the numerous factors that contribute to lap time variability, the ability to model and predict lap performance offers a deeper perspective to understanding how races evolve. A lap time reflects not only driver input but also the outcome of car setup, track configuration, race strategy, and surrounding conditions. As such, accurate lap time prediction has meaningful applications across the F1 community (Hudson, 2024). For drivers and constructors, predictive lap time insights may assist in optimizing pit strategies, evaluating qualifying runs, and comparing driver consistency under different race scenarios. For analysts and commentators, such models provide a foundation for constructing hypothetical race projections, which support both pre-race briefings and live commentary. Fans of the sport also benefit in that lap time prediction supports data-driven engagement through applications such as fantasy leagues, betting systems, and social media fan accounts that focus on simulations. Modeling lap time predictions from raw race data provides a powerful resource for understanding *why* drivers are fast, not just *who*.

2 Data Overview and Exploratory Data Analysis (EDA)

2.1 Data Overview

The dataset utilized is derived from the Kaggle competition *Formula 1 Lap Time Prediction,* which provides per-lap historical data from F1 races between 1996 and 2023. Each observation represents a completed lap and includes categorical and numerical features such as driver identity, constructor, circuit, lap number, pit stop information, and the recorded lap time in milliseconds. To ensure consistency with the project objective of forecasting lap times for 2025 races, the dataset was restricted to circuits listed on the 2025 season calendar that were also represented in the original data. This filtering process resulted in a refined dataset 'train_2025' comprising 22,393 laps across 17 circuits. Additionally, feature engineering was employed to construct 'is_pit_lap' and 'race_year' variables in order to enhance interpretability. These were derived from 'pitStop' and 'date', respectively.

2.2 Exploratory Data Analysis

2.2.1 Lap Time Distribution

The distribution of lap times for the filtered 2025 circuit was right-skewed, with most laps falling between 70,000 and 120,000 milliseconds (70-120 seconds). The distribution revealed multiple peaks, which are likely due to different track layouts, race conditions and the presence or absence of pit stops.

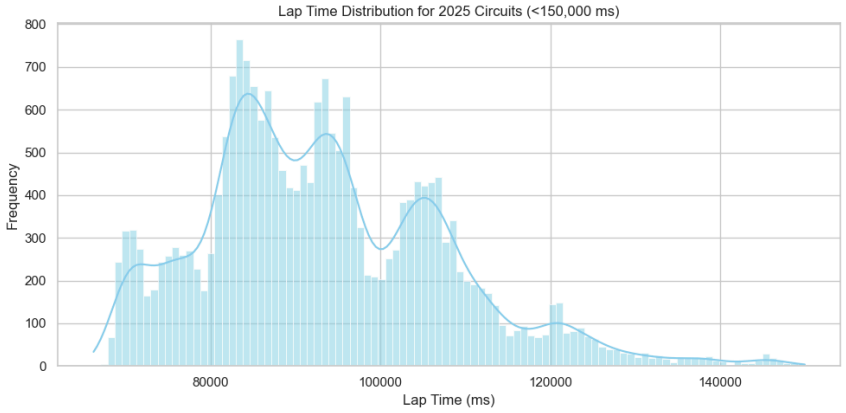

Figure 1: Histogram of Lap Time Distribution

2.2.2 Average Lap Time by Circuit

Lap times vary significantly across circuits, depending on track length and unique layouts. For example, the Red Bull Ring exhibited the shortest average lap time, while Baku City Circuit and Spa-Francorchamps showed much longer lap times. The bar chart below shows that substantial heterogeneity was observed across circuits.

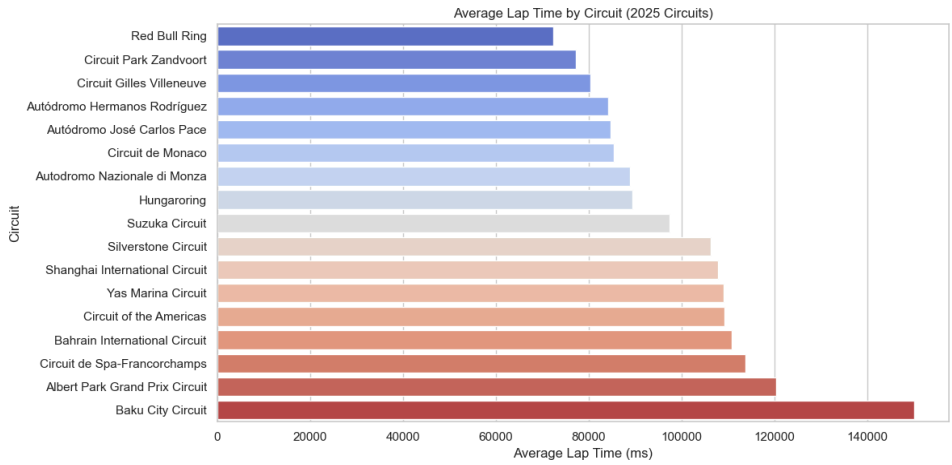

Figure 2: Bar Chart of Average Lap Time by Circuit

2.2.3 Average Lap Time By Constructor

Constructors historically linked to higher-tier performance such as Brawn GP were associated with lower lap times, while lower-tier constructors such as HRT exhibited the opposite pattern. While the dataset contains constructors that are no longer active as of the 2025 season, their inclusion was maintained to support model robustness and to capture broader performance across teams.

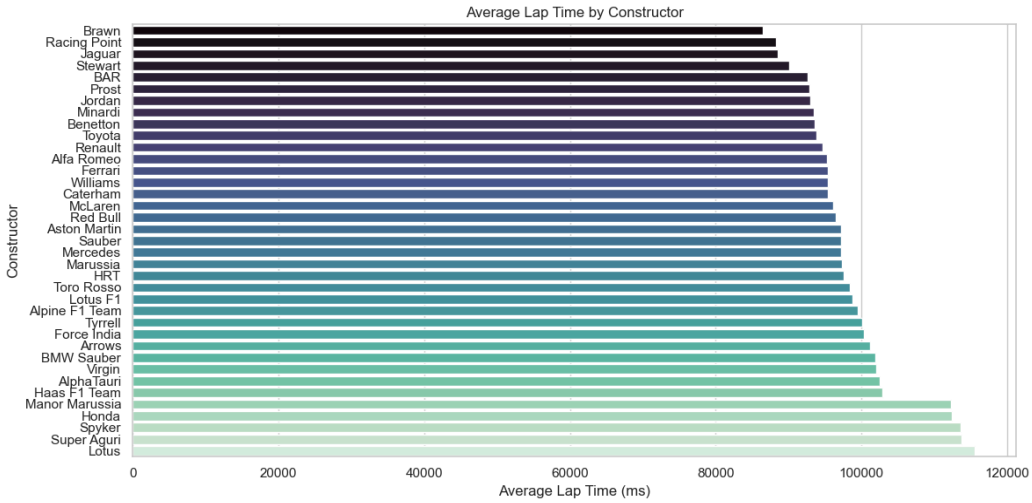

Figure 3: Bar Chart of Average Lap Time by Constructor

2.2.4 Pit Stop Influence

Laps designated as pit-in ('is_pit_lap' = 1) displayed significantly higher times, which validates the predictive relevance of pit-related indicators.

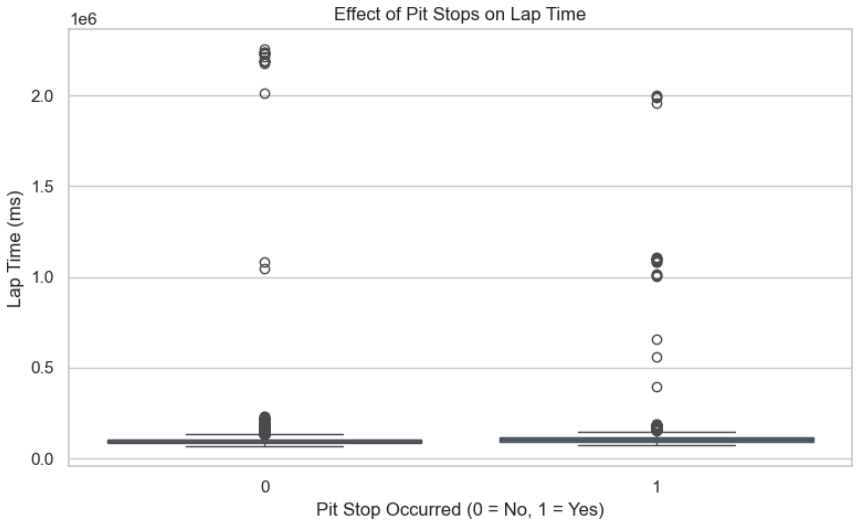

Figure 4: Box Plot of Pit Stop Effect on Lap Time

3 Methodology

3.1 Data Cleaning and Preprocessing

After the dataset had been narrowed to circuits on the 2025 calendar, a series of preprocessing steps were performed to ensure consistency in model training. Lap times were converted from milliseconds to seconds ('lapTime_ms' to 'lap_time_sec') for easier interpretation. Among the columns selected for modeling, all relevant fields were complete with the exception of the non-essential 'time' variable, which contained missing entries. Redundant attributes such as unique identifiers ('driverId', 'constructorId'), full names, and additional metadata were removed, yielding a modeling dataset composed of 20 variables.

3.2 Feature Transformation

Three categorical variables, 'driver', 'constructor', and 'circuit' were encoded as integers to avoid sparsity issues. In addition to the numeric features already present in the dataset such as 'lapNumber', 'lapPosition', 'avgDriverFinish', and 'avgConstructorFinish', the final model incorporated several indicators related to pit strategy ('pitStop', 'pitCount', 'pitTime_ms') as well as the two engineered variables 'is_pit_lap' and 'race_year'; these features together captures relevant aspects of race dynamics and driver-constructor performance. The resulting feature matrix consisted of 12 variables, and the response variable was defined as 'lap_time_sec'.

3.3 Model Selection and Evaluation Strategy

Three models were evaluated to determine the most effective model for predicting lap times:

- Linear Regression serves as a baseline model, assuming a linear relationship between input features and lap time.
- Random Forest is an ensemble of decision trees, and is included for its ability to model nonlinear interactions and reduce overfitting through bootstrapping and feature randomness.
- Gradient Boosting was included as a powerful boosting method that builds trees sequentially to minimize errors.

Each model was evaluated using 5-fold cross-validation on the training dataset; this method divides the data into five equal parts, using four parts for training and one for validation in each round, ensuring that all samples are used for both training and testing across folds (Quan, 2024).

The evaluation metric used was Root Mean Squared Error (RMSE), defined as:

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

RMSE was chosen for its interpretability in seconds and its sensitivity to large prediction errors, which are critical in the context of lap time prediction.

Model	Mean RMSE (seconds)	Standard Deviation (seconds)
Random Forest	39.90	2.99
Gradient Boosting	58.36	8.43
Linear Regression	70.37	12.17

Table 1: Summary of model performance for three regression algorithms using 5-fold cross-validation

The Random Forest model achieved the lowest average RMSE of approximately 39.9 seconds, outperforming both Gradient Boosting and Linear Regression, while also demonstrating the lowest standard deviation across folds, indicating consistent performance. Thus, Random Forest was selected due to its superior mean accuracy and consistency (Quan, 2024).

4 Results

4.1 Final Model Training and Feature Interpretation

The final model was trained using Random Forest with 100 estimators and a fixed random seed. After training, feature importance scores were extracted to assess which input variables contributed most significantly to the model's predictions.

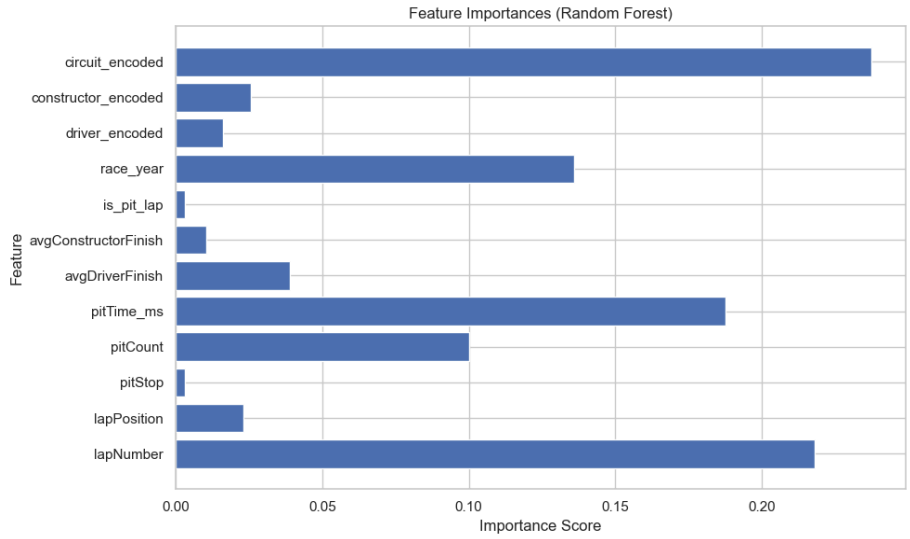

Figure 5: Bar chart of feature importances from the Random Forest regressor

Feature importance analysis revealed that 'circuit_encoded', 'lapNumber', and 'pitTime_ms' held the greatest predictive weight, while lower importance was observed for 'driver_encoded', 'lapPosition', and 'is_pit_lap'.

4.2 Simulation of Predicted Lap Times

To explore the model's predictive capabilities in a real-world scenario, a simulated 52-lap race was constructed for Charles Leclerc at Silverstone Circuit for the 2025 British Grand Prix. Charles Leclerc was selected for the simulation due to his consistent presence at Ferrari since 2019, which provides the model with real historical examples of the driver-constructor pairing. The 'race_year' was fixed at 2023 as a proxy for unseen 2025 data. Based on recent historical performance, Leclerc's average finish position was set at 5.0, and Ferrari's constructor finish was set at 3.5. A two-stop strategy was implemented, with pit stop losses estimated at 25 seconds per event, consistent with typical pit time loss at Silverstone.

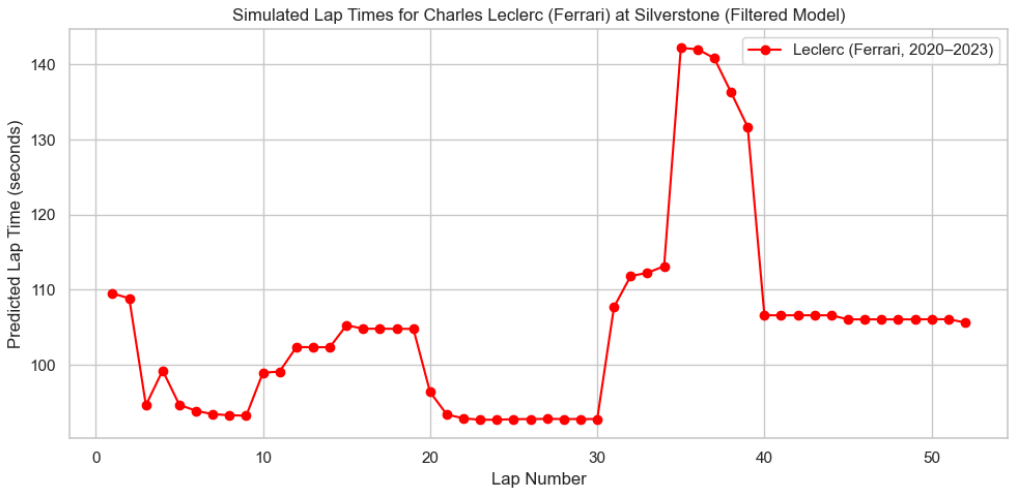

Figure 6: Simulated lap times for Charles Leclerc driving for Ferrari at Silverstone, generated using a Random Forest model trained on race data from 2020-2023.

Across the full race distance, lap times remained relatively stable during uninterrupted stints, ranging from approximately 93 to 107 seconds. Notable deviations occurred at laps 15 and 35, where sharp increases in predicted lap times reflect scheduled pit stops. Outside these intervals, lap durations followed a consistent pattern, with a modest downward trend observed in the final stint, which is likely attributable to reduced fuel load. These sharp increases during pit laps support the importance and interpretability of pit-related variables such as 'pitTime_ms' and 'is_pit_lap', which appear to be effectively incorporated into the model's structure. For context, Charles Leclerc's fastest lap during the 2024 British Grand Prix at Silverstone was recorded at 1:29.748, with typical laps falling behind 1:31 and 1:35 under dry but variable conditions (Formula 1, 2024). The model's predicted outputs, which align with this range, suggest that the simulation reproduces realistic race dynamics.

Note on Simulation Dataset Construction: The synthetic test dataset used for the simulation was independently constructed rather than derived from the original 'test.csv' file provided from the Kaggle competition. This decision was intentional to support the main goal of the project, which was to evaluate the model's predictive capability in a realistic race scenario for a known driver-constructor pairing (Charles Leclerc-Ferrari) on a specific circuit (Silverstone). The original 'test.csv' lacked the target variable (lapTime_ms) and was intended for competition leaderboard submission, not evaluation. The synthetic dataset's design was based entirely on historical trends in the training data, thus this approach enabled simulation-based insights that directly supports the project's core question (Quan, 2024).

5 Discussion and Reflection

The study demonstrated that a data-driven model can generate lap time predictions that align with observed race behavior. The Random Forest model, trained on lap-level data from 2020-2023, was chosen for its accuracy and robustness and proved to be effective in accounting for contextual factors such as circuit characteristics and pit strategy. The study demonstrated how the model could be effectively applied to a simulated future race scenario, which would allow for the interpretation of projected race outcomes.

The initial design considered Lewis Hamilton as the subject of the simulation; this was motivated by the extensive scope of his career data, which spans multiple seasons and provides a rich foundation for modeling driver performance. However, since Hamilton only joined Ferrari starting from the 2025 season, the model had no training data pairing him with that constructor, since the most recent F1 season from the dataset was 2023. Incorporating a pairing that is absent from the training data such as the Hamilton-Ferrari pairing would have introduced an unobserved combination, thereby reducing the simulation's foundation in real-world conditions. To avoid this issue, Charles Leclerc was selected as the simulation subject, as his continuous tenure with Ferrari since 2019 ensured that the model was applied within a historically supported context, allowing for more consistent and interpretable outputs.

Limitations include the exclusion of weather, tire degradation, and driver error, which were unavailable in the Kaggle dataset. If dataset options were not limited to ones from Kaggle competitions, the model could have likely achieved greater realism and predictive power by taking advantage of a more broad and current set of race inputs.

Note on Data Integrity and Synthetic Simulation: The synthetic test dataset used to simulate Charles Leclerc's 52-lap race at Silverstone was created entirely using data derived from the original Kaggle dataset. All input features in the simulation were derived from or informed by patterns in the original Kaggle dataset. In addition to creating this simulation set, the training dataset was filtered for 2025 circuits, and additional variables were created strictly from existing data columns. No external data sources were introduced at any point, thus these modifications adhere to project guidelines requiring exclusive use of the Kaggle data, while enabling the model to be applied meaningfully in a future race scenario (Quan, 2024).

6 Conclusion

This study addressed the viability of forecasting F1 lap times using historical race data where a Random Forest model was successfully trained and validated via simulation, demonstrating predictive consistency under hypothetical 2025 scenarios. The alignment of forecasted lap durations with real-world benchmarks supports the model's applicability in F1 analytics.

7 Resources

1. Kaggle. (2024). *Formula 1 Lap Time Prediction - NWVS S00E02* [Data set]. https://www.kaggle.com/competitions/formula-1-lap-time-prediction-nwvs-s00e02
2. Formula 1. (n.d.). *Drivers, teams, cars, circuits and more; Everything you need to know about F1.* Formula1.com. https://www.formula1.com/en/latest/article/drivers-teams-cars-circuits-and-more-everything-you-need-to-know-about.7iQfL3Rivf1comzdqV5jwc
3. Hudson, M. (2024, August 22). *How F1 teams plan strategy: Secret codes and hundreds of voices.* The Times. https://www.thetimes.com/sport/formula-one/article/how-f1-teams-plan-strategy-7t93vlgdg
4. Mugge, E., & Zandieh, M. (2024). *Race day mathematic: Analyzing Formula 1 performance through data-driven models* [PDF]. Arizona State University. https://cisa.asu.edu/sites/g/files/litvpz691/files/2024-04/Mugge_E_Zandieh.pdf
5. F1StatBlog. (2016, June 28). *Longest straights in F1.* https://f1statblog.co.uk/2016/06/longest-straights-f1/
6. Formula 1. (2024, July 7). *2024 British Grand Prix fastest laps.* Formula1.com. https://www.formula1.com/en/results/2024/races/1240/great-britain/fastest-laps
7. Quan, Z. (2024). *STA 160 Discussion 7: Model evaluation and interpretation.* Department of Statistics, University of California, Davis. Unpublished course handout.
8. Quan, Z. (2024). *STA 160 Discussion 3: Supervised learning and regression methods.* Department of Statistics, University of California, Davis. Unpublished course handout.
9. Quan, Z. (2024). *STA 160 Discussion 6: Final project report writing guide.* Department of Statistics, University of California, Davis. Unpublished course handout.