Track To Triumph: Prediction Success with Formula 1 Data

Josh Arvind
The George Washington University
Advisor: Professor Majhi
Data Science Program

1 Objective

This project describes a machine learning method that uses engineered performance parameters and historical race data to forecast Formula 1 race results. Three prediction tasks were examined: evaluating the likelihood of a Did Not Finish (DNF), predicting final race positions, and projecting Top 3 finishes. Several models, including Random Forest, XGBoost, LightGBM, and Logistic Regression, were tested following feature creation and data cleaning. The results demonstrate that consistent driver performance and constructor dominance make race outcomes, particularly Top 3 finishes, quite predictable. Strong accuracy was attained by position regression models, but DNF prediction was more difficult due to the unpredictability of mechanical breakdowns and on track accidents. Relative performance markers, such points above the race average and previous podiums, were found to be the best predictors using feature relevance analysis. Overall, the study demonstrates that machine learning can accurately model Formula 1 performance trends and support real-time prediction. The project includes trained models, full documentation, and an interactive dashboard for practical use.

2 Introduction

Formula 1 (F1), which brings together elite drivers, state of the art automobile engineering, worldwide competition, and detailed racing strategies, is usually regarded as the pinnacle of international motorsport. Since the premiere World Championship season in 1950, the sport has developed into a multibillion dollar industry that is televised annually to millions of fans in more than 20 nations. Every race, known as a Grand Prix, is held on a different circuit; some permanent tracks, some temporary street layouts, and tests various facets of driver talent and vehicle performance. Formula 1 is primarily a technological and strategic conflict. Teams design and construct their own vehicles, continuously refining materials, braking systems, power units, and aerodynamics. Race outcomes can be affected by even a small performance improvement, measured in thousandths of a second. As a result, Formula 1 has developed into a testing

ground for innovation, with cutting edge simulation models, energy recovery systems, carbon fiber construction, and hybrid engines all directly arising from F1 research

Success in racing requires much more than just quick driving. To make split second strategy decisions, teams must consider tire wear, fuel consumption, weather, pit stop timing, and track position. While negotiating tight turns, fast straights, and close racing against rivals, drivers must strike a balance between control, speed, and risk. Another important factor is reliability; unforeseen circumstances, collisions, and mechanical malfunctions can significantly alter results.

The volume of data produced is a characteristic that sets modern Formula One apart. Detailed timing data, GPS monitoring, tire temperatures, engine performance metrics, aerodynamic measurements, and race control updates are all generated with each lap. As a result, Formula 1 has evolved into a data-rich analytical environment in addition to a sports event. To analyze this data and obtain a competitive edge, teams hire engineers, machine learning experts, and data scientists.These features complex variables, quantifiable patterns, and high performance make Formula 1 a perfect subject for predictive modeling. Trends in driver skill, constructor dominance, circuit-specific performance, and dependability can be seen in past race results. Researchers and spectators may examine these trends, pinpoint KPIs, and project future race results thanks to machine learning.

In conclusion, Formula 1 racing is a special combination of data science, engineering, sport, and real time, decision making. Its extensive statistical record and long history serve as a basis for sophisticated analytics that may turn unprocessed data into insightful knowledge about race strategy, performance, and forecasting. Prior research has demonstrated that historical race data can be used to identify consistent performance patterns among drivers and constructors. Carvalho [1] showed that long-term driver success can be quantitatively evaluated using historical race outcomes, while Patil et al. [2] applied machine learning techniques to predict Formula 1 race results using structured performance data. Adjustments for team strength and competitive context have also been shown to significantly improve driver performance evaluation [3]. More recently, Rane [4] demonstrated that separating driver and constructor effects through linear modeling yields more accurate outcome predictions.

3 Impact

This research demonstrates that sports analytics can bridge data science and fan engagement. By making predictions accessible through interactive dashboards, complex insights become understandable to broader audiences beyond engineers and strategists. Key message: Machine learning in Formula 1 is not only about achieving high predictive accuracy, it is about uncovering the performance patterns that drive outcomes and communicating these insights effectively.

4 Approach and Timeline

Downloading and Cleaning Data (week 1-2)

Historical Formula 1 datasets were downloaded from F1DB and recent season data were queried using the OpenF1 API. Data preprocessing included standardizing column formats, handling missing values, and ensuring consistent labeling of drivers, constructors, races, and circuits.

Exploratory Data Analysis (Weeks 3–4)

Exploratory analysis was conducted to identify historical trends in wins, podium finishes, DNFs, and fastest laps. Circuit-specific performance patterns were examined, along with the relationship between qualifying performance and final race outcomes.

Feature Engineering (Weeks 5–6)

Aggregate features were created, including average pit stop time, number of pit stops, driver performance consistency metrics, and constructor-level performance indicators. Categorical variables such as constructors and circuits were encoded, and weather or track condition features were incorporated when available.

Modeling (Weeks 7–8)

Baseline models, including Logistic Regression and Linear Regression, were established for comparison. Advanced models such as Random Forest, XGBoost, LightGBM, and ensemble approaches were trained for classification and regression tasks. Model performance was evaluated using accuracy, F1-score, precision, recall, ROC-AUC, RMSE, MAE, and R^2.

Visualization and Dashboard Development (Weeks 9–10)

An interactive dashboard was developed to explore results dynamically. Visual components included lap-time trends, pit strategy comparisons, team performance across seasons, and a visual race replay showing driver positions lap by lap.

Final Report and Presentation (Weeks 11–12)

Results, model evaluations, and visualizations were compiled into a structured final report, an interactive dashboard, and a presentation summarizing key findings.

5 Dataset

Multiple publicly accessible Formula 1 datasets are used in this work to facilitate predictive modeling. The F1DB database provided constructor records, driver statistics, and historical race results [1]. The FastF1 Python module, which allows for the organized extraction of race and qualifying performance information, was used to retrieve telemetry, lap timing, and session-level data [2].Formula 1 and FIA datasets were used to validate official race classifications, standings, and technical data [3]. When combined, these resources offer a thorough

and trustworthy basis for examining race level variability, constructor dominance, and driver performance.

1. **F1DB (2024):** A structured Formula 1 database containing historical race results, driver statistics, constructor information, and seasonal records.
2. **FastF1 (2024):** A Python package providing access to telemetry data, lap timing information, and advanced race analytics.
3. **Formula 1 Official (2024):** Official FIA and Formula 1 datasets containing race classifications, standings, and technical updates.

6 Models

Several machine learning models were selected to address different predictive tasks while balancing interpretability and performance.

Logistic Regression A baseline classification model used for binary prediction tasks such as Top 3 finishes and DNFs. It provides interpretable probabilistic outputs and serves as a performance benchmark, according to LaValley [2] .

Random Forest Classifier An ensemble learning method that aggregates multiple decision trees to capture nonlinear interactions between drivers, constructors, and circuit features, as Fatima et al. [1] states.

XGBoost A gradient boosting framework optimized for structured data and high predictive accuracy. It is well suited for modeling complex feature interactions in Formula 1 race outcomes as Fatima [1] states .

LightGBM A highly efficient gradient boosted decision tree model designed for speed and scalability, particularly effective for ranking and classification tasks, according to Ke et al. [3].

Linear Regression Used as a baseline regression model for predicting final race position as a continuous variable, enabling comparison with more advanced regression approaches, as Kumari et al.[4] mentions.

Ensemble Methods Multiple model predictions were combined to improve robustness and overall predictive performance by reducing variance and model bias as Zhou [5] mentions.

7 Analysis

1 . Top 3 Finishes Logistic regression was highly outperformed by Ensemble, XGBoost, and LightGBM models, highlighting the significance of nonlinear feature interactions,

like recent form and team dominance. Table 2 displays the nearly perfect accuracy and ROC-AUC values across top models for classification performance measures for Top 3 predictions.

2. Final Position Can Be Estimated with Useful Accuracy Regression models produced low error rates, reflecting realistic finishing ranges even in the presence of overtakes and race incidents. Qualifying results, constructor strength, and recent performance were the strongest predictors. The comparative performance of regression models is summarized in Table 1, which reports RMSE, R^2, and MAE values for each approach.

3. DNFs Are Difficult to Predict Due to randomness in crashes, mechanical failures, and weather, DNF classification yielded lower performance. However, teams with historical reliability issues and drivers with higher long term failure rates were identified as higher risk. More granular mechanical telemetry could improve accuracy. Table 3 displays the classification results for DNF prediction. These results suggest that adding more detailed mechanical telemetry might raise prediction accuracy even more.

   These findings are consistent with prior studies showing that podium finishes are strongly influenced by constructor dominance and persistent driver performance trends (Carvalho, 2022 [1], Patil, 2022 [2]).

Regression Model Performance

Model	RMSE	R	MAE
Random Forest	**1.55**	**0.9107**	**0.88**
LightGBM	1.56	0.9089	0.96
Ensemble	1.61	0.9030	1.00

Table 1: Regression model performance for predicting final race position.

Top 3 Classification Performance

Model	Accuracy	F1-Score	ROC-AUC
Ensemble	**99.75%**	**0.9918**	0.9999
XGBoost	99.59%	0.9865	**1.0000**
LightGBM	99.55%	0.9848	0.9999

Table 2: Classification model performance for Top 3 prediction.

DNF Classification Performance

Pit Stop Strategy Visual Analysis

Model	Accuracy	F1-Score	ROC-AUC
LightGBM	**92.58%**	0.6552	**0.9636**
Ensemble	92.29%	**0.6594**	0.9611
XGBoost	92.21%	0.6415	0.9610

Table 3: Classification model performance for DNF prediction.

8 Closing Statement

Formula 1 race outcomes, especially podium finishes, are highly predictable due to consistent driver performance and constructor reliability. With over 75 years of historical data and modern machine learning tools, accurate forecasting is increasingly achievable. This framework has practical applications for analytics platforms, team strategy, and fan engagement, demonstrating that with the right data and the right features, performance becomes predictable.

9 References

9.1 Formula 1 References

1. Carvalho, A. (2022). *Using Historical Data to Identify the Best Driver in Formula 1 History*. Doctoral dissertation, National College of Ireland.

2. Patil, A., Jain, N., Agrahari, R., Hossari, M., Orlandi, F., & Dev, S. (2022). A data-driven analysis of Formula 1 race outcomes. In *Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 134–146). Springer.

3. Phillips, A. J. (2014). Uncovering Formula One driver performances by adjusting for team and competition effects. *Journal of Quantitative Analysis in Sports*, 10(2), 261–278.

4. Rane, S. (2025). Predicting Formula 1 Race Outcomes: Decomposing the Roles of Drivers and Constructors through Linear Modeling. *arXiv preprint* arXiv:2508.00200.

9.2 Technical References

1. Fatima, S., Hussain, A., Amir, S. B., Awan, M. G. Z., Ahmed, S. H., & Aslam, S. M. H. (2023). XGBoost and Random Forest algorithms: an in-depth analysis. *Pakistan Journal of Scientific Research*, 3(1), 26–31.

2. LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18), 2395–2399.

3. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

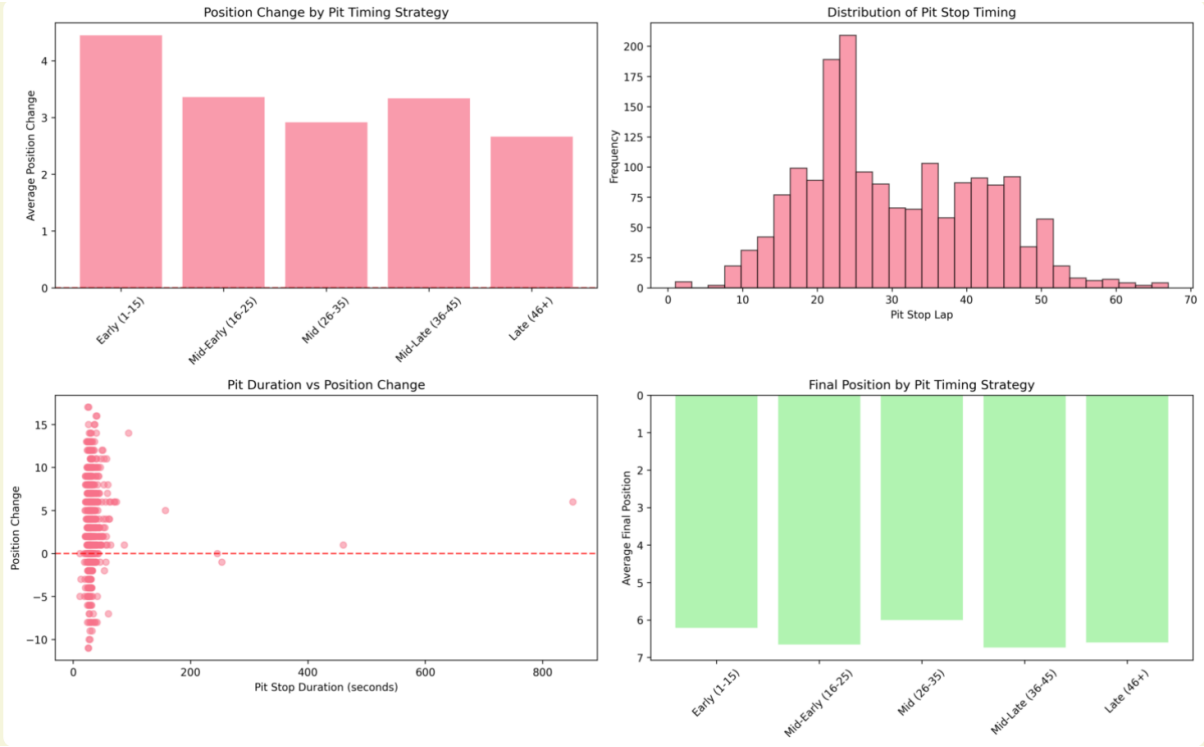

Figure 1: The first set of graphics looks at the impact of pit stop timing on race results. Drivers who stop early (between laps 1–15) typically earn the most positions, while those who stop later in the race (beyond lap 46) gain fewer positions, according to the top left graphic, Position Change by Pit Timing Strategy. This implies that early pit stops could give drivers a competitive edge, likely by allowing them to take advantage of open track space and undercut competitors.

Figure 2: The majority of pit stops occur between laps 15 and 30, peaking around lap 22, as shown in the top right chart, Distribution of Pit Stop Timing. This indicates that teams generally prefer mid race stops, aligning with standard fuel management and tire wear strategies.

Figure 3: Shorter pit stops; lasting roughly two to five seconds, are the most frequent, as depicted in the bottom left graph, Pit Duration vs. Position Change. These shorter stops typically correspond with slight or positive position adjustments, whereas longer stops lead to significant position losses, emphasizing the importance of pit crew efficiency and precision.

Figure 4: Finally, the bottom-right chart, Final Position by Pit Timing Strategy, reinforces the trend that early or mid race pit strategies tend to yield stronger overall finishes. Drivers who execute their pit stops earlier generally achieve better final positions compared to those who delay their stops toward the end of the race. Taken together, these graphs suggest that success in Formula 1 depends not only on driving skill, but also on the strategic timing and execution of pit stops. Optimal pit stop planning; balancing timing, duration, and tire strategy, can make a decisive difference in race outcomes.

4. Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, 4(1), 33–36.

5. Zhou, Z. H. (2021). Ensemble learning. In *Machine Learning* (pp. 181–210). Springer.