Northwestern University

Machine Learning-Based Prediction of Formula One Race Outcomes

A Data-Driven Study of Top-10 Finish Prediction Using Historical Race Data

Prepared by:

Sara Alsiyat - Qifan Yang - Boqi Niu

MSDS 422: Practical Machine Learning

Dr. Irene Tsapara

Northwestern University

January 2026

Abstract

Formula One race outcomes are influenced by multiple factors, including driver performance, constructor strength, circuit characteristics, and race conditions. Finishing in the top 10 of a race is particularly important because it determines whether a driver earns championship points and reflects competitive performance. Historical race data provides an opportunity to study these factors and understand patterns in race outcomes. This project examines whether machine learning methods can be used to predict whether a driver finishes in the top 10 of a Formula One race using historical data.

The study uses a publicly available Formula One World Championship dataset from Kaggle, covering races from 1950 to recent seasons. The dataset is provided in multiple relational tables and includes race results, qualifying data, driver and constructor information, race status, and circuit details. The unit of analysis is at the driver–race level, where each record represents a driver's result in a specific race. The dataset contains both numerical attributes, such as finishing position, grid position, points, laps completed, and season indicators, and categorical attributes, including driver, constructor, circuit, country, and race status.

The target variable for this project is a binary indicator representing whether a driver finishes within the top 10 positions in a race. The purpose of the study is to evaluate whether historical race information can be used to support structured prediction of race outcomes. The project is intended to provide a data-driven foundation for analyzing Formula One performance and exploring the potential of machine learning techniques in competitive sports analytics.

Problem Statement

Formula One race results depend on many factors, including driver ability, constructor performance, track characteristics, and race conditions. Although race outcomes are often seen as unpredictable, teams and analysts use historical data to better understand what drives performance. Finishing in the top 10 is especially important because it determines whether a driver earns championship points and reflects overall race success.

This project focuses on predicting whether a driver will finish in the top 10 of a Formula One race using historical race data. Machine learning methods are used to combine information about drivers, constructors, circuits, and recent performance. The goal is to understand whether race outcomes can be explained using measurable factors rather than chance, and whether predictive models can support performance analysis and strategic decision making in Formula One.

Research objective

The main objective of this project is to build and evaluate machine learning models that predict top-10 finishes in Formula One races using historical data. The study examines how factors such as qualifying position, constructor team, driver performance history, track characteristics, and recent race results affect finishing position. The project also explores whether some circuits show more variation in race outcomes, which may suggest more unpredictable races compared to tracks where performance is more consistent. In addition, the research considers related outcomes such as points finishes and Did Not Finish (DNF) events to better understand race reliability and performance risk. Finally, different machine learning models will be compared to assess their accuracy, interpretability, and usefulness for decision-support purposes.

Annotated Bibliography

Race to the Podium: Separating and Conjoining the Car and Driver in F1 Racing – Alsiyat

This paper studies what really drives race outcomes in Formula One by separating the impact of the driver from the impact of the car and constructor. The authors use historical Formula One race results and financial data from multiple seasons, focusing on races between 2012 and 2019. The dataset includes information on drivers, constructors, race finishing positions, qualifying positions, retirements (DNF), team budgets, and driver salaries. The main goal of the study is to understand how much each factor contributes to a driver's finishing position and whether success in Formula One is driven more by individual skill, team resources, or the interaction between both. The paper applies statistical models to race-level data and focuses on measurable outcomes such as finishing position and podium results.

The results show that both driver skill and constructor performance play an important role in determining race outcomes, but their effects are not equal. While driver ability matters, the interaction between the driver and the team explains a large portion of race performance. High-performing drivers tend to outperform their teammates even in similar cars, while strong constructors provide advantages through better technology, strategy, and reliability. The study also shows that race outcomes are not random and that historical performance can be used to explain and anticipate future results.

This paper is highly relevant to our project because it directly supports our goal of predicting whether a driver will finish in the top 10 of a race. From a decision-support perspective, the clear separation between driver and constructor effects justifies including both driver-level and team-level features in our machine learning models. In my professional experience working with national-level data and KPIs, separating contributing factors is critical for building reliable and explainable models. The paper's findings also support using interaction features, such as driver–constructor combinations, which aligns well with our planned feature engineering using the Kaggle Formula One dataset.

NeuralAC: Learning Cooperation and Competition Effects for Match Outcome Prediction - Alsiyat

This paper introduces NeuralAC, a machine learning framework designed to predict match outcomes by modeling both cooperation and competition effects between participants. The authors evaluate the model using historical match outcome data from competitive team-based environments, where each observation represents a match involving multiple participants working together against opponents. The dataset includes team compositions, participant identities, and match results, allowing the model to learn how interactions influence outcomes rather than relying only on individual performance metrics. NeuralAC uses neural networks to capture these interaction patterns and compare them against traditional prediction models.

The key contribution of this paper is showing that modeling interaction effects leads to more accurate predictions than treating participants independently. The results demonstrate that cooperation within a team and competition between teams both strongly influence outcomes. The experiments show that NeuralAC performs better in complex competitive settings where outcomes depend on coordination, shared resources, and strategic interactions. This highlights the limitation of models that rely only on individual features.

This paper is important for our project because Formula One race outcomes also depend on interaction effects, especially between the driver and the constructor. While Formula One appears to be an individual sport, performance depends heavily on teamwork, car reliability, race strategy, and coordination. From a decision-support perspective, this paper supports going beyond driver-only features and incorporating interaction features in our models. In applied analytics work, capturing how factors work together often produces more useful insights than analyzing them in isolation. NeuralAC provides strong methodological support for experimenting with models that can capture complex relationships when predicting top-10 finishes using historical Formula One data.

A Data-Driven Analysis of Formula 1 Car Races Outcome - Yang

This article gives us a statistical analysis to help understand the most influential factors that can impact the outcome of Formula 1 races. The outcomes are measured as total championship points scored by a driver over the period of a season. The main motivation is that F1 teams collect vast amounts of data, but yet few studies or analysis systematically analyzes which variables really affect race results. In the study, the researchers compiled race data from 2015 to 2019 by web-scraping and processed the data into a dataset with 21 features such as pit frequency, tyre usage percentages, laps led, penalties, race positions, starting positions, accidents, and driver completion rates. First, correlation analysis was conducted and it turned out that there are strong relationships among position-related variables and between tyre usage and penalties. This suggests that there are complex dependencies among factors. Due to the high feature interdependence, the authors used Principal Component Analysis (PCA) to reduce the dimensionality. The first four principal components can capture around 70% of the total variance. This indicates that the major patterns can be represented in a lower-dimensional space without significant loss of information.

Next, a linear regression model is used to predict total points scored by drivers (based on the reduced feature set). Key findings include that finishing more races correlates strongly with higher points, while tyre usage of certain compounds such as soft, medium, hard also significantly affects performance. Additionally, a better average starting position or pole position is associated with more points. This illustrates the importance of qualifying. The model achieved an R-square of 99%, which suggests a strong explanatory power on this dataset.

The authors conclude systematic statistical analyses can discover meaningful insights into F1 race performance.

Advanced Machine Learning Approaches for Formula 1 Race Performance Prediction: A Comprehensive Analysis of Championship Point Forecasting - Yang

This study presents a machine learning framework designed to predict Formula 1 race performance and championship points. It used a dataset which spans 74 years of F1 history from 1950 to 2024. This includes 589081 individual lap times from 1125 total races. The authors' goal was to overcome limitations in prior works by using complex feature engineering, algorithm comparisons, and rigorous validation methods which are specially geared towards complex motorsport analysis.

Their approach begins with building meaningful features from qualifying race data, lap time data, circuit characteristics, and temporal dynamics across different eras of racing from 1950 to 2024. These engineered features capture strategic and performance factors such as grid position effects, lap-to-lap variations, and historical performance patterns across different circuits. Various machine learning models are evaluated, including traditional regression, ensemble methods, and gradient boosting techniques. It also comes with hyperparameter tuning and cross-validation to ensure robust generalization.

Among the tested algorithms for example gradient boosting such as XGBoost consistently outperforms others, yielding exceptional predictive accuracy with an R-square score of 0.999, RMSE of 0.197, and MAE of 0.125 on forecasting championship points. Feature importance analysis indicates that race position is the dominant predictor, which contributes around 75.8 % to model performance, with seasonal variations accounting for another 23.8 %. Cross-validation results demonstrate strong generalization (mean R-square roughly equals to 0.993 plus or minus 0.013).

The paper argues that its findings not only advance predictive modeling performance in motorsport but also hold some practical implications for F1 teams, broadcasters, and strategy analysts. It also offers insights into how historical performance and context can influence race outcomes.

The Use of Machine Learning in Predicting Formula 1 Race Outcomes - Niu

This paper examines the application of machine learning techniques to predicting Formula 1 race outcomes, with a particular focus on driver finishing positions and constructor championship points. Using historical Formula 1 data from the 2010 to 2023 seasons, the study develops a structured predictive modeling pipeline based on TabNet, a deep learning architecture specifically designed for tabular data. TabNet is chosen for its strong predictive performance as well as its built-in interpretability through feature attention mechanisms.

Two separate models are constructed: a driver-level model that predicts individual finishing positions and a constructor-level model that forecasts total team points per race. The models rely exclusively on pre-race features such as grid position, number of laps, constructor affiliation, and derived indicators like overtaking potential. Data preprocessing includes cleaning incomplete records, encoding categorical variables, feature normalization, and chronological train–test splitting to prevent data leakage. Hyperparameter tuning is performed using Optuna to optimize model performance.

Empirical results demonstrate that the driver model achieves strong predictive accuracy (R^2 = 0.75, RMSE = 2.87), outperforming the constructor model, which exhibits slightly higher variability due to team-level complexity (R^2 = 0.71, RMSE = 3.92). Visualization and residual analyses indicate low systematic bias and good alignment between predicted and actual outcomes. The study concludes that interpretable deep learning models like TabNet can effectively predict Formula 1 race outcomes using structured pre-race data. Limitations include the absence of real-time race variables such as weather, safety cars, and telemetry, which are identified as promising directions for future research.

Predicting Formula 1 Race Outcomes: A Machine Learning Approach - Niu

This paper investigates the use of machine learning techniques to predict Formula 1 lap times and race outcomes using historical race data from 2014 to 2023. The primary objective is to assess how accurately lap times for individual drivers can be forecasted and how these predictions translate into overall race results. The study emphasizes the importance of modeling temporal dependencies in racing data, given that lap performance is influenced by prior laps, race conditions, and driver history.

The author constructs a large-scale dataset comprising over 214,000 laps across 203 races and 55 drivers, incorporating core performance metrics, race conditions (such as pit stops and safety cars), and driver- and team-specific indicators. Several baseline models—including linear regression, decision trees, and random forests—are first evaluated to establish reference performance levels. While random forests outperform simpler models, their prediction errors remain large relative to the typical time gaps between drivers, limiting their usefulness for accurate race outcome prediction.

To better capture sequential patterns, the study focuses on Long Short-Term Memory (LSTM) networks. Multiple LSTM architectures are developed and refined through increased training epochs, dropout regularization, custom feature scaling, and a novel composite loss function that combines lap time error, positional accuracy, and historical driver performance. Experimental results demonstrate that minimizing lap time error alone does not necessarily lead to better race outcome predictions, highlighting the importance of incorporating relative and contextual performance measures.

Evaluation on selected races from the 2023 and 2024 seasons—particularly the Abu Dhabi and Bahrain Grands Prix—shows that later LSTM models substantially outperform baseline approaches in predicting podium finishers, top-five placements, and race winners, especially in races with fewer unexpected incidents. The paper concludes that LSTM-based models, when properly designed and evaluated, offer significant promise for predictive analytics in Formula 1. Future work is proposed to explore transformer-based models and to extend prediction scope to include pit stop strategies and safety car events, aiming for a more comprehensive race forecasting system.

NBA Winner Prediction: A Hybrid Framework Incorporating Internal and External Factors – Alsiyat

This paper studies how machine learning can be used to predict the outcome of NBA games by combining both internal and external factors. The authors use historical NBA regular season data from the 2012–13 season to the 2021–22 season, collected from professional sports databases. The dataset includes game-level information such as team performance statistics, win–loss outcomes, home and away status, Elo ratings, and contextual factors like home court advantage and player tiredness due to back-to-back games. The main objective of the study is to improve prediction accuracy by incorporating contextual and situational features in addition to traditional performance metrics. The authors apply supervised machine learning models and evaluate their performance using cross-validation.

The results show that combining internal team performance indicators with external contextual factors leads to better prediction accuracy compared to models that rely only on basic statistics. Features such as differences in Elo ratings, recent team performance, home court advantage, and tiredness were found to be important predictors of game outcomes. Among the tested models, ensemble-based approaches such as random forest achieved the strongest performance. The study demonstrates that sports outcomes are influenced by both historical performance and situational conditions, and that machine learning models can capture these relationships effectively.

This paper is relevant to our project because it provides a clear example of how combining multiple types of features improves outcome prediction in competitive sports. From a decision-support perspective, the paper supports our approach of integrating driver-level, constructor-level, and contextual race information when predicting top-10 finishes in Formula One. Similar to how NBA outcomes depend on both team strength and external conditions, Formula One race results depend on driver skill, car performance, circuit characteristics, and race context. In applied analytics work, incorporating both internal and external factors leads to more reliable and explainable models. The methodology used in this paper helps justify our feature selection strategy and supports the use of supervised machine learning models for structured performance prediction using historical Formula One data.

Bayesian analysis of Formula One race results: disentangling driver skill and constructor advantage - Yang

A very interesting and valuable question is posed through the article regarding analytics in Formula 1 racing by determining how driver skill and constructor advantage affect performance. This concept is explored through Bayesian analysis within the work of Erik-Jan van Kesteren and Tom Bergkamp. The beauty of sports analytics, in a very simple and instinctive way, is the ability to rank competition participants by their individual ability; however, in high-speed racing and high-stakes competitions (i.e., Formula 1) there are many influences and other types of factors that come into play and affect the outcome of the competition very differently than in many other forms of sport. The probing and inquisitive thinking of the millions of fans and devotees of this sport, regarding questions like “What is the effect of the driver’s skill versus the constructor?” remain unanswered.

In order to accomplish this, the authors have created a new Bayesian multilevel (hierarchical) model for rank-ordered logit regression that directly uses data from race finishing positions during the hybrid era of Formula One (2014–2021). The current model improves upon previous efforts by not losing information from point-based systems; it also uses the performance differences of teammates and driver movements between teams to differentiate between driver and constructor effects.

The model computes an entire finish order (rather than aggregate points) allowing for greater details regarding competitive dynamics.

Through this Bayesian framework, the authors are able to measure the skill of each driver and the advantage of each constructor (in log-odds terms) in relation to their competitors (analogous to Elo ratings in chess) with credible intervals to measure uncertainty. The findings show that team performance (constructor) has a large impact on the outcomes of races; approximately 88% of the variance in results is associated with the car, and less than half of the remaining variance can be attributed to driver skill and/or other factors. Also, in terms of driver performance during this time period, the top two drivers were Lewis Hamilton and Max Verstappen, while the top three constructors were Mercedes, Ferrari and Red Bull.

This is a good study that explains exactly how the proposed method can be used to create counterfactual scenarios (such as what if a driver had been in a different type of car) that can be used to rank each individual driver and constructor independently while providing a statistical basis for all of this. The authors conclude that this methodology advances quantitative analytical rigor for measuring the performance of Formula One drivers but that it may extend beyond Formula One to other sporting and competitive settings where the interaction between multiple variables in performance measurement is relevant.

Learning to Identify Top Elo Ratings: A Dueling Bandits Approach - Niu

This paper addresses the problem of efficiently identifying top-performing players under the Elo rating system by reducing the number of matches required to accurately estimate player strength. Traditional Elo-based evaluation relies on repeated random or predefined match-ups, which can be inefficient, especially when the goal is to quickly identify the strongest players among a large pool of competitors. To overcome this limitation, the authors formulate the match scheduling problem as a dueling bandits problem and propose an adaptive, data-driven framework for selecting informative match-ups.

The paper introduces two online algorithms, MaxIn-Elo and MaxIn-mElo, which actively choose which pairs of players should compete at each step. Instead of scheduling matches uniformly at random, the algorithms focus on players whose Elo ratings are uncertain but potentially high-performing. This is achieved by maintaining confidence bounds on Elo estimates and prioritizing match-ups that maximize expected information gain. Elo ratings are updated online using stochastic gradient descent, allowing the system to scale efficiently while using constant memory per iteration.

From a theoretical perspective, the authors show that MaxIn-Elo achieves sublinear cumulative regret, meaning that the algorithm becomes increasingly efficient over time compared to random or naive scheduling strategies. The framework is also extended to multidimensional Elo (mElo), which allows the model to capture intransitive competitive relationships, such as scenarios where no single player consistently dominates all others. This extension is particularly important in competitive environments where performance depends on interaction effects rather than a simple global ranking.

The proposed methods are evaluated through extensive experiments on both synthetic data and real-world competitive games. Results show that MaxIn-Elo and MaxIn-mElo converge significantly faster and produce more accurate rankings than baseline approaches, including random sampling and existing dueling bandit methods. In both transitive and intransitive settings, the algorithms demonstrate superior performance in identifying top players with fewer comparisons.

Overall, the paper demonstrates that combining adaptive match scheduling with online Elo updates provides a more efficient and scalable approach to ranking competitors. While the study focuses primarily on identifying the best player and assumes stationary skill levels, the framework offers valuable insights for ranking and outcome prediction tasks in complex competitive systems.

References

Vopani. “Formula 1 World Championship (1950 - 2024).” Kaggle, January 29, 2025. https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020.

Rockerbie, Duane W., and Stephen T. Easton. “Race to the Podium: Separating and Conjoining the Car and Driver in F1 Racing.” Applied Economics 54, no. 54 (July 5, 2022): 6272–85. https://doi.org/10.1080/00036846.2022.2083068.

Gu, Yin, Qi Liu, Kai Zhang, Zhenya Huang, Runze Wu, and Jianrong Tao. “Neuralac: Learning Cooperation and Competition Effects for Match Outcome Prediction.” Proceedings of the AAAI Conference on Artificial Intelligence35, no. 5 (May 18, 2021): 4072–80. https://doi.org/10.1609/aaai.v35i5.16528.

Patil, Ankur, Nishtha Jain, Rahul Agrahari, Murhaf Hossari, Fabrizio Orlandi, and Soumyabrata Dev. “A Data-Driven Analysis of Formula 1 Car Races Outcome.” Communications in Computer and Information Science, 2023, 134–46. https://doi.org/10.1007/978-3-031-26438-2_11.

Bansal, Aayam. Advanced Machine Learning Approaches for Formula 1 Race Performance Prediction: A Comprehensive Analysis of Championship Point Forecasting, 2025. https://doi.org/10.13140/RG.2.2.20910.01607.

Urdhwareshe, A. 2025 "The Use of Machine Learning in Predicting Formula 1 Race Outcomes" Preprints. https://doi.org/10.20944/preprints202504.1471.v1

Jafri, Ali. *Predicting Formula 1 Race Outcomes: A Machine Learning Approach*. Capstone project, Fall 2024.

https://aliabdullahjafri.com/static/media/Ali_Jafri_CapstoneProject1_Fall2024.c7244022875d46bec5d9.pdf.

Zheng, Xi. "NBA Winner Prediction: A Hybrid Framework Incorporating Internal and External Factors." *2022 4th International Conference on Big Data Engineering*, May 26, 2022, 71–80. https://doi.org/10.1145/3538950.3538960.

Kesteren, Erik-Jan van, and Tom Bergkamp. "Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage." Journal of Quantitative Analysis in Sports 19, no. 4 (July 25, 2023): 273–93. https://doi.org/10.1515/jqas-2022-0021.