University at Buffalo Logo | 03 - PNG Logo Vector Downloads (SVG, EPS)

Finish Line Forecasters: F1 Data-Driven Foresights

A Comprehensive Model Incorporating Track, Location, and Historic Data

Anusha Udayakumar - 50542184

A Report Submitted in partial fulfillment of the requirements for the Engineering Science (Data Science) MS

DEPARTMENT OF

Institute for Artificial Intelligence and Data Science School of Engineering and Applied Sciences

University at Buffalo

# Abstract

The proposed paper is based on data analysis and machine learning on Formula 1 dataset is studied to gain valuable new insights into sports research based on driver's performance, constructor standings and many other factors which affect the final predictions. The paper aims to draw valuable insights by using various machine learning algorithms to gain future readings of the driver's performances. Furthermore, analysis based on various algorithms is used to compare and pick the best fit which provides more accuracy and performances well for unknown instances. In addition, this leads to Incorporation of historical data and cross validation, scaling, hyperparameter tuning for data selection. This paper endeavors to offer valuable insights for forthcoming research in this domain, while also outlining potential enhancements and broader applications of the project for future exploration.

# Introduction

Formula 1 Racing stands as a captivating blend of innovative technology, driver skill, and strategic prowess. The quest for predicting the unpredictable outcomes of a season is a challenge that demands a nuanced understanding of historical trends, track intricacies, and the dynamic interplay of variables. In this research endeavor, we aim to construct a robust and comprehensive predictive model, leveraging historical data, driver/team performance metrics, track characteristics, and additional factors to forecast the outcomes and standings of the 2024 Formula 1 season.

This paper explores a method using machine learning to predict winners in Formula 1 races, tapping into supervised algorithms. Recognizing its success in top-tier sports like Formula 1, the study reviews existing literature on machine learning in sports predictions. It emphasizes where the data comes from, how it is collected, and how models are evaluated. The research tackles emerging challenges in Formula 1 by adding new features to historical data and refining model accuracy through cross-validation. The goal is not just to improve winner predictions but also to set the stage for future research in sports analytics, particularly in the ever-changing world of Formula 1.

The realm of Formula One (F1), predicting race outcomes has long been a captivating challenge for both enthusiasts and experts. This intricate motorsport is characterized by high-speed thrills, strategic maneuvering, and intense competition, making it difficult to accurately forecast the victor of each race. However, with the advent of advanced statistical analysis and a growing wealth of historical data, the possibility of developing reliable predictive models is now within reach.

This research endeavor delves into the realm of F1 race prediction using statistical analysis, aiming to establish a robust and effective methodology for forecasting race outcomes. By harnessing the power of data analysis, we strive to illuminate the intricacies of this dynamic sport and contribute to the advancement of F1 prediction techniques.

Furthermore, the complexity of the sport, influenced by external factors such as weather conditions, track characteristics, and unforeseen race events, adds layers to the analytical challenges. The dual motivation for this project stems from both a scientific desire to explore methodological approaches for predicting sports rankings and an economic interest in leveraging ML for sports betting insights.

By addressing these considerations, we aim to develop a predictive model that not only provides accurate forecasts of race outcomes but also enhances our understanding of the complex interplay of factors influencing Formula 1 seasons.

# Objective

The primary objective is to develop a predictive model that transcends traditional approaches, incorporating not only historical performance but also dynamic variables that influence the modern Formula 1 landscape. By synthesizing data from diverse sources, including the Ergast API via the FastF1 Python Package, our goal is to enhance the accuracy and reliability of predictions, providing a deeper understanding of the intricate dynamics governing Formula 1 seasons.

# Literature Review

A thorough examination of existing literature reveals a landscape where predictive modeling in Formula 1 has predominantly focused on historical performance metrics, driver/team dynamics, and race-specific variables. Notable works include those exploring machine learning applications in racing predictions and analyses of historical data to predict race outcomes. While these studies contribute valuable insights, our research seeks to extend beyond these boundaries, incorporating a more diverse set of variables and considerations.

# Research Questions

These are a few questions we wish to solve in scope of this paper.

**Utilizing Historical Data for 2024 Predictions:** How can historical data from previous F1 seasons, spanning 1951-2023, be effectively utilized to forecast the outcomes of the 2024 season? What insights can be drawn from long-term trends and patterns?

**Beyond Performance Metrics:** Apart from historical data and driver/team performance, what additional variables significantly impact the outcomes of the season? Are there novel factors that, when integrated, contribute to a more holistic and accurate predictive model?

**Realistic Predictive Accuracy:** Given the inherent unpredictability of Formula 1 races, how accurate can predictions realistically be? What are the limitations and challenges in achieving precision, and how can these be mitigated?

# Data Source and Quality

Our research draws upon the Ergast API via the FastF1 Python Package, encompassing a wealth of data from 1951 to 2023. The relatively clean and comprehensive nature of this dataset enables us to conduct a meticulous analysis of race information, competition results, participant details, and performance metrics, including lap times, pit stops, and confidential data on engines and tires.

One of the key components of the dataset is its extensive race information. This includes detailed records of individual races, each season’s schedule, and comprehensive data about the circuits where these events take place. This aspect of the dataset not only provides fundamental information about the locations and timing of the races but also offers a glimpse into the evolution of the sport over the decades. The data on circuits, for instance, can be used to study changes in track design and safety standards, reflecting the technological and regulatory evolution of Formula 1 racing.

The dataset also delves into the competitive aspects of Formula 1, providing data on constructor standings and results, as well as individual race results. This includes information about the teams (constructors) that participate in the sport, their performance over the seasons, and their standings in the championship. Additionally, the dataset covers the participants in great detail, including both the drivers and the constructors. This section is particularly valuable for analyzing the careers of drivers, the history and performance of teams, and the dynamics of driver-team relationships over the years.

Lastly, the dataset offers an in-depth look at various performance metrics that are critical to understanding the sport. This includes detailed lap times and information about pit stops, which are essential for analyzing race strategies and driver performance. These metrics provide insights into the nuances of race tactics, the efficiency of teams in the pit lane, and the overall performance of drivers during a race. Such data is invaluable for fans, commentators, and analysts seeking to understand the finer details of race strategy and driver skills in the highly competitive world of Formula 1.

# Feature Engineering

To ensure accurate and reliable analysis of lap times, driver’s performance, and other factors a comprehensive data preparation process was undertaken. This involved several key steps:

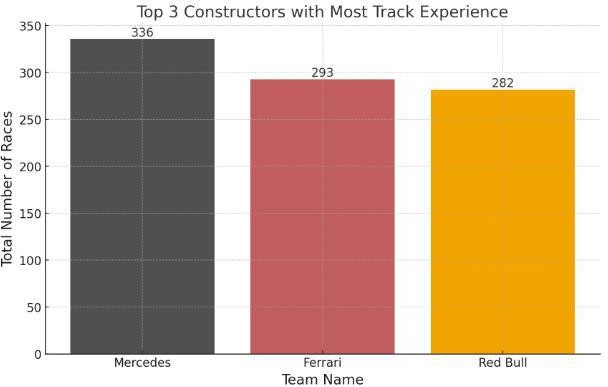
* **Missing Value Removal:** All instances containing missing values, denoted by '/N', were removed from the dataset. For example: In result dataset, columns like time and position features included ‘\N’ values for players who did not complete the race. These rows were removed as they did not add any value to the analysis. Moreover, missing data can introduce bias and skew results, therefore its exclusion was crucial for accurate analysis.
* **Non-Finish Removal:** Instances where drivers did not finish the race (dnf) or retired early were also excluded. Lap times recorded for incomplete races would have provided inaccurate and misleading information for our analysis.
* **Timeframe Refinement:** To ensure consistent and independent data distributions, the analysis focused solely on data from 2016 to 2023. This timeframe was chosen to account for potential changes in regulations, technological advancements, and driver performance over time.
* **Outlier Removal:** Outliers in lap time records were identified and removed. Outliers can significantly distort the overall picture and mask underlying trends, so their removal was essential for accurate analysis.
* **Merging Datasets:** To incorporate valuable driver-related information, the metadata dataset was merged with the results dataset. This provided a richer context for analyzing lap times, player names and allowed for correlation with other driver metrics.
* **Redundancy Elimination**: Redundant pit stop data that was already present in the results dataset was removed to avoid data duplication and ensure efficient analysis.

These meticulous data preparation steps ensured that the analysis was based on a clean, accurate, and representative dataset. This allowed for reliable insights into lap time trends, driver performance, and other key factors impacting race outcomes.

# Exploratory Data Analysis

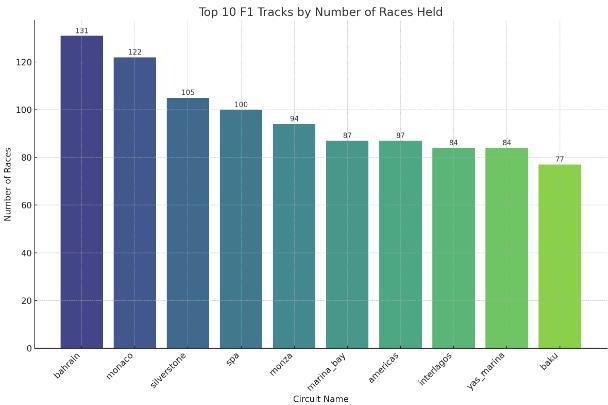
Exploratory Data Analysis (EDA) is an essential step in the data science process that involves examining datasets to summarize their main characteristics. The provided visualizations, including bar charts, a pie chart, and a

correlation matrix, are typical tools used in EDA to uncover patterns, identify anomalies, test hypotheses, or check assumptions with the help of summary statistics and graphical representations.

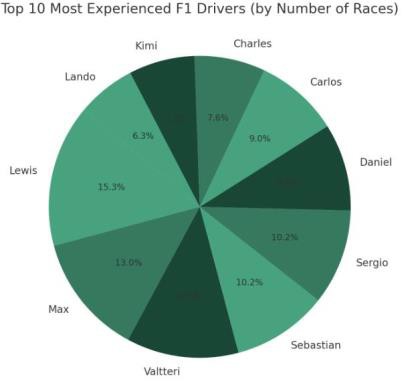


The bar chart shows the *Top 3 Racing Teams with Extensive Track Expertise* highlighting the dominance of

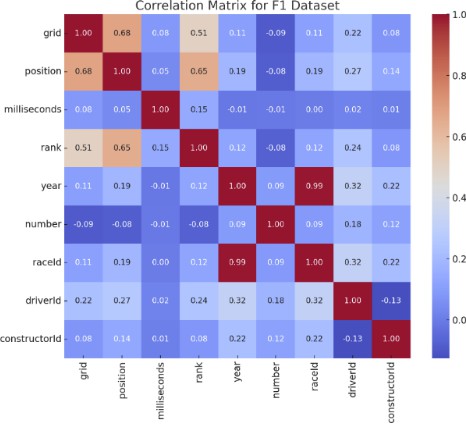
teams such as Mercedes, Ferrari, and Red Bull in motorsports. This underscores the notion that teams with a longer history and deeper experience on racing tracks are more likely to achieve success, due to their long- standing involvement in the sport.



The second bar chart lists the *Top 10 F1 Tracks by Number of Races Held* which can be used to understand which circuits have the most historical significance. The second bar chart details the *TOP 10 Formula 1 Circuits Ranked by Total Races Hosted,* shedding light on which tracks hold the greatest historical importance in the sport.



The pie chart presents the *Top 10 Most Experienced F1 Drivers* highlighting the distribution of race experience among them, with Lewis, Sebastian, and Max being notably prominent.



Complementing these visualizations, the correlation matrix adds another dimension to EDA by providing a quantitative view of the relationships between variables. In the F1 dataset's correlation matrix, we can discern significant relationships, such as the strong positive correlation between 'grid' and 'position' and a strong negative correlation between 'grid' and 'rank'.

Although 'year' and 'race ID' exhibit a high correlation coefficient of 0.9, it is important to note that correlation does not imply causation. In this context, the strong correlation between these variables merely indicates a consistent pattern or association in their values over time, rather than suggesting that changes in one variable directly cause changes in the other. 'Year' represents a chronological sequence, while 'race ID' likely denotes categorical or grouping information. Their high correlation might stem from a temporal or categorical arrangement rather than a meaningful causal relationship. Correlation is a valuable tool in understanding relationships between variables, it is essential to exercise caution in inferring causation from correlation alone.

By combining these visual and quantitative tools, EDA becomes a powerful approach to understand complex datasets. It helps stakeholders in the F1 community — from team strategists and sports analysts to fans — to gain a multifaceted understanding of the sport. These insights can be instrumental in shaping strategies, making predictions, and enhancing the overall appreciation of Formula 1 racing. The correlation matrix, especially, serves as a guide for further statistical analysis and modeling, ensuring that the relationships between variables are appropriately considered and utilized.

# Model Development

The primary methodology involves constructing a principal pipeline consisting of a Python object, ColumnTransformer, which compresses numerical and categorical operations (feature scaling and encoding). The second component is the model, initialized with default parameters. Hyperparameter tuning is carried out through cross-validation using RandomizedSearch to identify the best-performing parameters in terms of accuracy scores.

# Multi Linear Regression

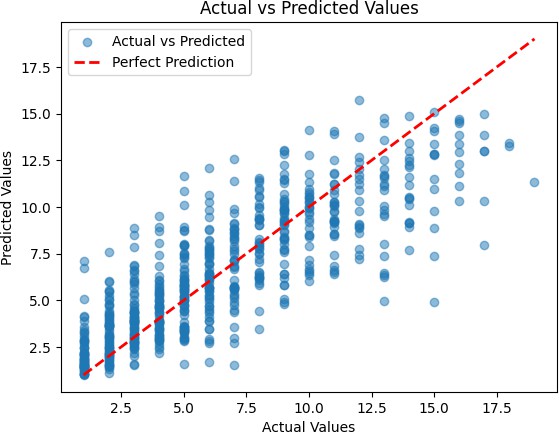
Multiple Linear Regression is chosen for predicting F1 winners due to its capability to consider multiple factors simultaneously, like driver performance, team statistics, track characteristics offering insights into their collective impact on race outcomes in a comprehensible manner.

In MLR, the model is represented as Y = β0 + β1X1 + β2X2 + ... + βn\*Xn + ε, where Y is the dependent variable, β0 is the intercept, β1 to βn are the coefficients for independent variables X1 to Xn, and ε is the error term. MLR estimates these coefficients by minimizing the difference between predicted and actual values, following the least squares method.

A few points to note here is:

1. MLR assumes a linear relationship between variables like driver performance, team statistics, track characteristics and the predicted outcome.
2. MLR expects consistent variability in the errors (residuals) across the range of predicted values. For instance, the variance in the prediction error for a driver's position should ideally remain constant across different levels of variables like team performance or track characteristics.
3. MLR assumes that the discrepancies between actual and predicted race outcomes should align with a bell- shaped curve.

**Multi Linear Regression Equation**: **y = -3.27 + (0.004526031637268779 \* X1) + (0.0006961757625142864 \* X2) + (0.0035575708096485343 \* X3) + (0.34940529558169764 \* X4) + (-1.3284679164647824e-07 \* X5) + (0.34307431891059503 \* X6) + (- 0.002618159250708658 \* X7)**

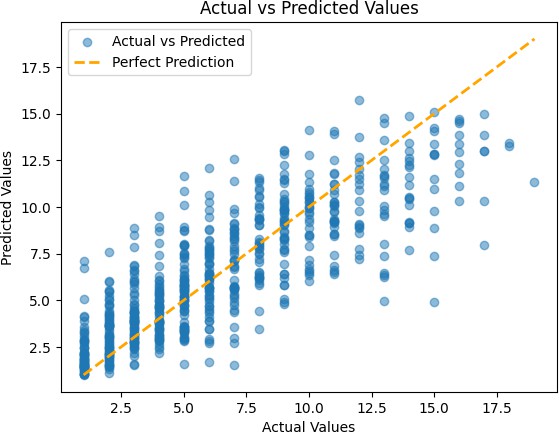


# Random Forest Regression

Random Forest Regressor, an ensemble learning solution, was chosen for its ability to handle noise and deal with categorical variables prevalent in the dataset. It is characterized by fast runtime and resistance to overfitting.

The Random Forest Regressor operates by randomly selecting subsets of data for each tree via bootstrapping, while at each node, using a random subset of features for splitting to reduce correlation and overfitting. Multiple trees are built based on these subsets, each making individual predictions. For regression task, the final prediction is generated by averaging the predictions from all trees, yielding a more robust outcome less susceptible to overfitting than a single decision tree.

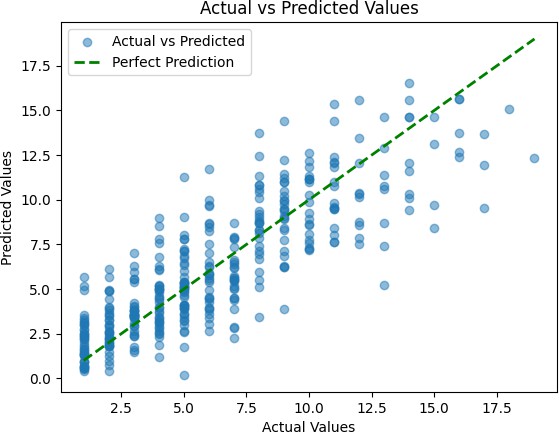
The Random Forest Regressor excels in reducing overfitting through aggregation, handles large and complex datasets adeptly, and provides insights into feature importance for enhanced interpretability.



# Gradient Boosting Regression (XGBoost)

Gradient Boosting Regressor, chosen as an extension to the ensemble solution, optimizes multiple parameters and typically outperforms random forests. The implementation involves boosting iterations, controlled through shrinkage and bagging. XGBoost minimizes a loss function by adding new models that complement the existing ones.

It stands out due to its incorporation of diverse regularization techniques like L1 (Lasso) and L2 (Ridge) regularization, which control model complexity and curb overfitting. Additionally, it employs tree pruning, discarding less impactful splits, thus optimizing tree structure for efficiency. Its inherent ability to handle missing values negates the necessity for imputation, and its optimization for speed through parallel processing sets it apart from traditional gradient boosting, making it exceptionally efficient in building trees.



# A table with numbers and letters Description automatically generatedModel Performance

**Results**

Our analysis of predictive models reveals the prowess of Random Forest Regression (RFR) as a robust contender in unraveling the complexities within Formula 1 data. Employing ensemble learning, RFR adeptly dissects the intricacies of race outcomes. In parallel, XGradient Boosting Regression (XGBR) demonstrates iterative refinement, showcasing its adaptability for precision.

Moving beyond numerical metrics, qualitative nuances emerge, shedding light on the distinctive impacts of each algorithm. While lap times and grid positions wield significant influence, Constructor ID surfaces as a nuanced factor, revealing the pivotal role of team strategies, shaped by constructors, in determining race outcomes.

## Evaluation:

The evaluation phase extends beyond conventional metrics, exploring the models' performance under diverse scenarios. Sensitivity analysis uncovers the models' responses to minor input deviations, providing insights into their robustness. Robustness testing further highlights adaptability and generalization capabilities, even in the face of outliers, underscoring the reliability of our predictive models.

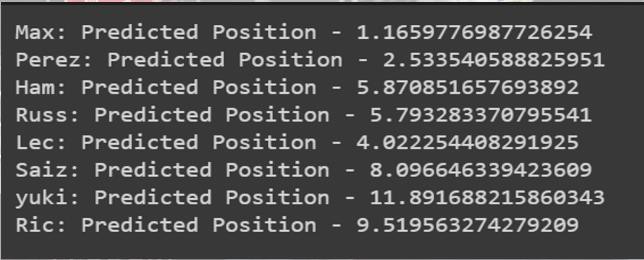
## Key Influencers:

The discussion on key influencers transcends surface-level explanations. Grid position evolves into a strategic advantage, guiding models in anticipating initial conditions that shape race dynamics. DriverID becomes a narrative of individual driving styles, drawing parallels with historical performances to enhance predictions.

Race ID's temporal context becomes a focal point, recognizing the evolution of strategies and performances over seasons. The impact of milliseconds on precision in timing gains prominence, emphasizing the critical role of split-second intervals in Formula 1 races. Circuit ID, as a beacon for track-specific characteristics, becomes a crucial factor in deciphering the unique challenges posed by different circuits.

# Predictions:

Our predictions extend beyond forecasting driver standings to anticipate potential race dynamics. Integrating historical data and recent trends, the models serve as a comprehensive tool for enthusiasts and analysts to understand unfolding narratives in Formula 1 racing. Expected correct predictions for the top 10 finishers, with varying margins of error, showcase the models' versatility in accommodating uncertainty.



Here's a glimpse of our predictions for prominent drivers on 29th feb 2024 at Bahrain:

* *Max Verstappen:* Confidently positioned at the top, a testament to his consistent excellence.
* *Sergio Perez:* Anticipated to secure the fourth position, recognizing his skills in a competitive field.
* *Lewis Hamilton:* Intriguingly predicted at ninth, showcasing the dynamic nature of Formula 1.
* *George Russell:* Positioned eighth, highlighting the model's adaptability in recognizing emerging talents.
* *Charles Leclerc:* A strong second-place prediction underscores the model's precision.
* *Carlos Sainz:* Expected to secure the tenth position, revealing the model's ability to navigate competition.
* *Yuki Tsunoda:* Predicted at sixteenth, acknowledging challenges and unpredictability in F1 races.
* *Daniel Ricciardo:* Anticipated to finish thirteenth, illustrating the nuanced nature of our model.

# Conclusion

In conclusion, our findings affirm the commendable performance of ensemble models, particularly Random Forest and XGBoost, in forecasting race outcomes. These models not only exhibit strong predictive accuracy against historical data but also demonstrate an understanding of nuanced feature patterns influencing race results. The incorporation of comprehensive features, including driver experience and track-specific factors, enhances predictive capabilities. While acknowledging the inherent unpredictability of Formula 1 races, these models offer valuable insights into potential outcomes, contributing to a more informed understanding of the dynamic interplay of factors in each racing season.

# Future Scope

**Weather Conditions:** Looking ahead, we aim to refine our predictive models by incorporating real-time weather data. This means not just anticipating optimal tire choices but also adjusting strategies on the fly based on the evolving weather conditions during races.

**Regulation Changes:** In the future, we plan to stay ahead of the game by proactively considering upcoming rule changes. This involves making our models adaptable to evolving regulations, providing insights into how these changes might impact team performance and overall race outcomes.

**Technological Advancements:** As technology marches forward, so will our analysis. We're exploring ways to integrate the latest tech innovations in car design, simulation tools, and tire compounds. By doing so, our models can offer insights into how these advancements might reshape the dynamics of Formula 1 racing.

In a nutshell, our focus is on real-time adaptability, staying ahead of regulatory shifts, and keeping up with the latest tech trends to ensure our models remain at the forefront of understanding the ever-changing world of Formula 1.

# References

* + <https://www.f1-predictor.com/>
  + <https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020>
  + <https://www.formula1.com/en/racing/2023.html>
  + <https://f1metrics.wordpress.com/>
  + https://medium.com/hoyalytics/xgboost-theory-and-application-4801a5dba4fb