Using AI to enhance performance in Formula 1

Nayan Kumar

Using Artificial Intelligence to Enhance Performance in Formula 1

Abstract

Formula 1 is a sport in which every millisecond matters and engineers and drivers must make decisions in the blink of an eye. Every car generates gigabytes of data per race from telemetry, tires, track conditions, and other sources. It puts pressure on teams to work fast and provide solutions before the next race. With the complexities and unstructured nature of this information and the high-pressure environment in which the teams operate, the teams have an issue to address, which is how to mine all these insights and use them to inform strategy during the race to make the team more efficient. For years, this was all done through manual interpretation of data by engineers and analysts, but now that AI is starting to change how teams prepare and race. The work discussed in this paper will demonstrate the use of supervised learning (A machine learning technique) and more specifically gradient boosting regression model like XGBoost to help predict key race metrics, which can help to improve a f1 team's performance. The model will be trained on structured datasets which include driver inputs, environmental factors, historical race logs, and be able to predict race outcomes and react to real-time events more effectively. XGBoost will be used to capture the complex nonlinear relationships in the data and provide more accurate predictions under different conditions. By leveraging AI-powered insights, Formula 1 teams can reduce uncertainty, optimize strategies, and gain a competitive advantage. The datasets required for this model, evaluation metrics like MAE and RMSE, and the potential benefits and challenges of deploying the model in this high-pressure motorsport setting will also be discussed.

Introduction

(A) Problem Definition

Formula 1 is one of the most advanced and competitive sports today. Formula 1 cars can reach speeds up to 200 miles per hour, and decisions need to be made at that speed as well. However, these decisions need to be accurate and effective. Every car is equipped with

hundreds of sensors that send telemetry data about vehicle performance, driver behavior, and environmental conditions. Formula 1 teams have been using experienced engineers and data analysts to interpret the data and develop an effective race strategy. However, the volume and complexity of data has been increasing steadily and slowly going beyond human ability, real-time analysis and rapid decision making are becoming increasingly difficult. Artificial intelligence is an opportunity to disrupt race strategy and performance optimization.

(B) Previous Work (Literature Review)

The use of machine learning and AI in Formula 1 has garnered significant research attention. Noe and Patel (2024) leveraged telemetry data and ensemble learning methods, specifically Extra Trees and Random Forest, to develop lap time prediction models. Their work demonstrated that these methods can capture complex nonlinear relationships within race data, delivering highly accurate predictions that closely match actual race performance.

Todd et al. (2025) focused on tire energy consumption, employing XGBoost regression models augmented by explainability techniques. Their study highlighted key variables influencing tire wear, such as brake pressure and temperature gradients, providing valuable interpretative insights into the model's decision process. This transparency enhances confidence in AI recommendations, a crucial aspect when integrating machine learning into competitive race strategy.

Reinforcement learning approaches, as explored by Thomas et al. (2025), simulate race environments to optimize pit stop scheduling strategies. Their AI learns optimal policies through trial-and-error interactions with the racing context, showing improved race outcomes. This paradigm extends beyond prediction, emphasizing prescriptive analytics in motorsports.

(C) Proposed Solution: Using AI in Formula 1

The main problem this research intends to solve is taking large, complicated datasets, for example, driver input, tire wear, and weather, as input to Artificial Intelligence (AI) to make more optimal decisions faster. Formula 1 teams need to decide at any point during a race how long to go before a pit stop, how to adjust to a car with different tire wear levels and how to adjust to other external changes like track temperature.

Manual or rule-based analysis can be slower to identify nuanced relationships or adjust to changes. This paper details how supervised learning in the form of gradient boosting regression models (e.g. XGBoost) can be used to make predictions of a continuous race characteristic (lap time, tire degradation, fuel usage, etc.). These predictions can be used to inform decisions made during a race and over a longer timescale on car development.

(D) Supervised vs. Unsupervised Learning

One primary difference in AI modelling is if a model is a form of supervised or unsupervised learning. The former requires the presence of labelled outputs to train on, such as the example of clustering race laps with similar performance. In unsupervised training, this information does not exist, but it is not applicable to this study as it would require the use of supervised learning with historical inputs (e.g. throttle position, track temperature) and known outputs (e.g. lap time). Supervised models are more applicable to Formula 1 because the sport produces large quantities of labelled, structured data after each session. This is an ideal training environment to create models with clearly defined input–output relationships and can be expected to generalize well to other predictive tasks (lap time, tyre wear etc.).

(E) Type of Datasets Needed

In order to create a model with accurate predictions a wide range of data needs to be acquired. This can include: telemetry information (speed, RPM, throttle and brake input, gear changes), track temperature, humidity, rain data, tire information (compound, temperature, wear), pit stop, sector times and data from events (racing incidents etc.)

Data Category	Example Features	Source
Telemetry	Throttle %, Brake Pressure, Speed, Gear	Car Sensors
Tire Data	Compound Type, Temp (°C), Wear %, Pressure	Tire Sensors, Logs
Environmental	Track Temp, Air Temp, Rain %, Wind Speed	Track Sensors/Weather
Race Events	Lap Time, Pit Stop Time, Sector Times	Official Race Logs

Fig 1 : Shows what type of data is needed.

This diverse and structured data is what can be used to train the supervised models to make the predictions during the races in real time. Gathering all of these datasets and synchronizing them accurately is one of the first major steps in AI based decision making.

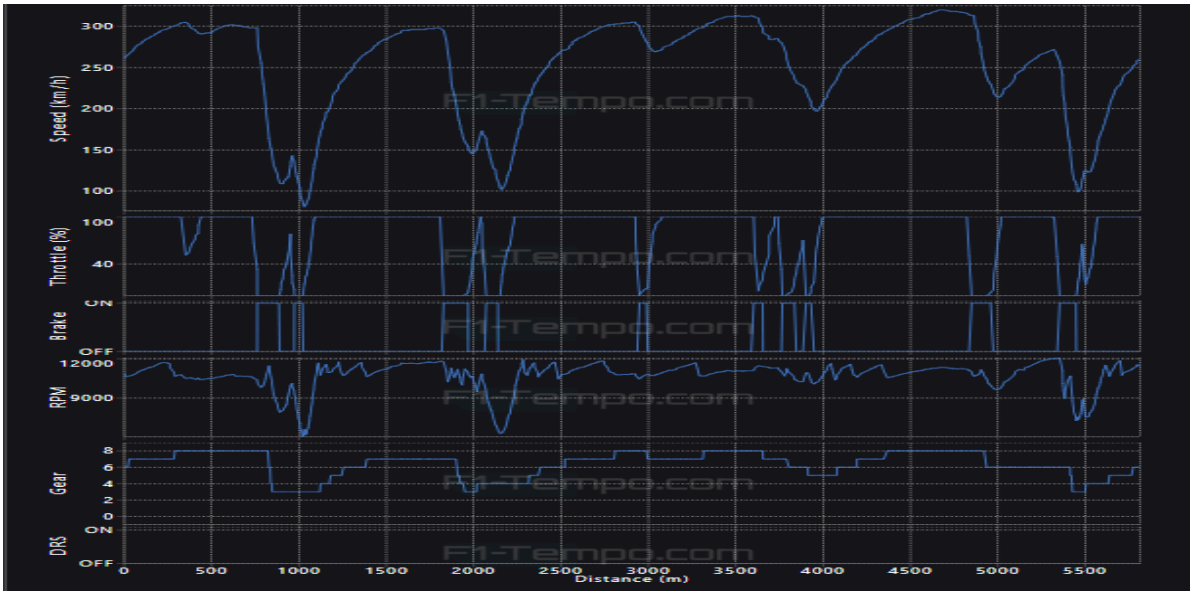

Fig 2 : Max Verstappen's car telemetry data 49th lap in the British Grand Prix Source : F1 tempo

(F) Modeling Approach & Paper Structure

For this paper, using a gradient boosting regression machine learning algorithm is the best and most effective way to approach the problem. The algorithm I want to use is very powerful in that it can easily model complex and nonlinear relationships, and it works well with the real-world noise in the data. The rest of this paper is organized as follows. I will then discuss how the data will be collected and prepared for the machine learning algorithm. Next, I will cover the modeling approach. I will then talk about how to evaluate the model with various metrics, including MAE, RMSE, and R2. Finally, I will then conclude with some discussion about the results, challenges, and potential directions for using AI in a real-time, high-stakes Formula 1 race.

Results / Perspective

(A) Data Acquisition Strategy

The quality of predictions from any ML model are largely dependent on the underlying data and the richness of features within. To that end, I plan to acquire data from a variety of sources and through different streams, some of which include: car telemetry, environmental conditions, tire characteristics and information about the circuit and race outcomes of previous races.

The telemetry dataset will consist of driver inputs and the car's instantaneous reaction to these inputs, i.e. Throttle position, brake pedal, gear shift position, steering angle, speed, engine temperature, etc., all gathered using the car's onboard sensors.

In addition to the telemetry information, I will gather information about track and ambient environmental conditions including but not limited to: track temperature, wind speed/direction, rain, humidity, etc. These can be gathered via onboard as well as trackside sensors. Weather has a large impact on overall grip, tire degradation and pit strategies. Tire characteristics such as compound, track and core temperatures, and tire degradation levels will be gathered using IR sensors and telemetry from the team's tire logs. Race events and annotations such as lap times, pit stops, sector times, safety car periods, DRS utilization, etc. will be used as labels for the supervised learning task.

All datasets will be cleaned, time-stamped, and normalized to remove noise and outliers. Data from anomalous laps (e.g. crashes, red flags, etc.) will be removed from the training set. The resulting joined dataset will be formatted (e.g. Pandas DataFrames, SQL, etc.) and split into train, validation and test sets.

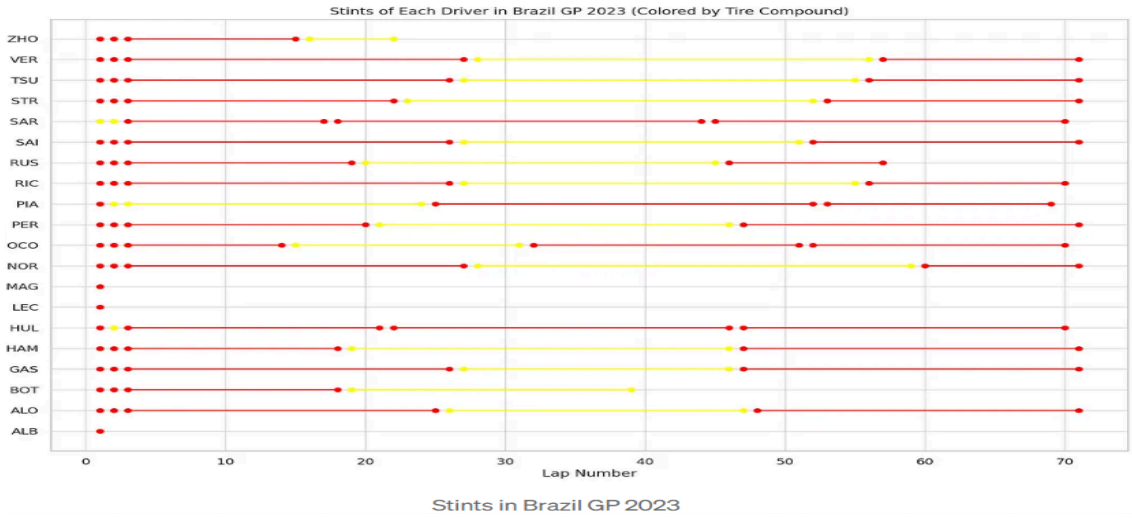

Fig 3 : Stints of Each Driver in Brazil GP 2023 (The type of tire they used during the race). Source : Tyre strategies in F1 using python

(B) Machine Learning Solution Details

In this project, I would like to explore the possibility of forecasting continuous race outcomes (lap time, tire wear, when to pit etc.) using supervised learning approaches. I propose using gradient boosting regression.

Gradient boosting models like XGBoost are state of the art for tabular data. They outperform in terms of performance, handle missing values gracefully and allow one to track feature importance. They work by combining the predictions of many decision trees which are added sequentially to correct the errors of the previous trees.

Gradient boosting can learn nonlinear relationships between input and output better than linear regression. This is crucial for something like a Formula 1 car that has a highly nonlinear relationship between what it's doing and the lap time or tyre degradation.

The inputs to this model would be telemetry data (throttle %, gear, brake % etc.), weather conditions (temperature, humidity, etc.) and tyre data (percentage tyre left, projected degradation etc.)

The output could be lap time, percentage tyre left or distance to the car in front.

The evaluation metrics that would be considered are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R^2 score. These will give a good idea of how close the predictions are on average, how the model deals with large errors and how well the model is doing overall in capturing the trends in the data.

(C) Supervised vs. Unsupervised Learning

The approach of this project is supervised learning, which means the algorithm is trained with examples of input and the expected output. Formula 1 is a perfect application for this because it has telemetry and environmental sensors that measure many quantities, but the sport is known for being very structured, where the result of these measurements is always well defined, like the length of a lap time or the amount of tire wear or time in a pit stop. These are called labels.

Supervised learning allows the algorithm to know how telemetry and environmental inputs correlate with these results and can then predict the output accurately.

Unsupervised learning would be the opposite. It is used when the data is unlabelled, so the algorithm is looking for hidden patterns and can cluster data (for example, identifying different laps where the driver was aggressive or had a different tire temperature), but it does not predict an actual value.

Unsupervised approaches could be useful for data exploration, but not for predicting a numerical metric like a race result in real time. This is why supervised learning is the right fit for this project.

(D) Classification vs. Regression

Machine learning models can be categorized by what type of data the model will predict. Is the model solving a classification problem or a regression problem?

Classification solves problems that can be described as predicting discrete categories or labels. Classification can be applied to Formula 1 to solve problems that can be described as a yes or no question such as “Will the driver pit on the next lap (yes or no)?” or “Will the driver’s tire last another 10 laps (true or false)?”.

Regression solves problems that are best described by a numeric value. Regression can be applied to Formula 1 to answer questions like “What will the driver’s lap time be?” or “How much percentage of tire wear will be experienced over the next 3 laps?”

Since the purpose of this study is to predict continuous race variables such as lap time (seconds), tire wear (%), or pit stop duration (milliseconds), it was determined that regression is the more appropriate modeling choice since classifying the prediction to a discrete category would not make sense and would limit the flexibility and accuracy of a real-time Formula 1 decision-making model.

(E) Choice of Model

For this study I propose XGBoost, a gradient boosting regression algorithm, which is driven by the nature of the data and the problem’s complexity. XGBoost is known for its efficiency with

large and complex datasets, which is typical in F1 racing data. It operates by constructing a series of decision trees in a gradient boosting framework, where each tree corrects the errors of the previous tree. This approach enables the model to learn and model complex, nonlinear relationships between the car's sensor readings, environmental conditions, and race performance metrics like lap time and sector splits.

XGBoost also offers built-in mechanisms for assessing feature importance, which enables teams to understand which variables— tire temperature, fuel load, or aerodynamic settings—are most influential in determining lap time and performance.

(F) Evaluation Metrics

To evaluate our models accuracy and reliability I propose to use 3 metrics : Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 Score.

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. In other words, it is the average over the test sample of the absolute differences between prediction and actual observation. It is linear and scales with the same unit as the target variable. For instance, if the target variable is lap time or tire wear, the MAE will be in the same units.

RMSE is a quadratic scoring rule that measures the average magnitude of the error. It is the square root of the average of squared differences between prediction and actual observation. It is more sensitive to outliers and large errors than MAE, as it squares the errors before averaging them. This can be useful to penalize large errors that may have a negative impact on race strategy or performance.

R2 Score, also known as coefficient of determination, measures how well the regression model fits the data. It is a value between 0 and 1 that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R2 score means a better fit, or a stronger relationship between the input variables (e.g. tire temperature, throttle position, etc.) and the output variable (e.g. lap time).

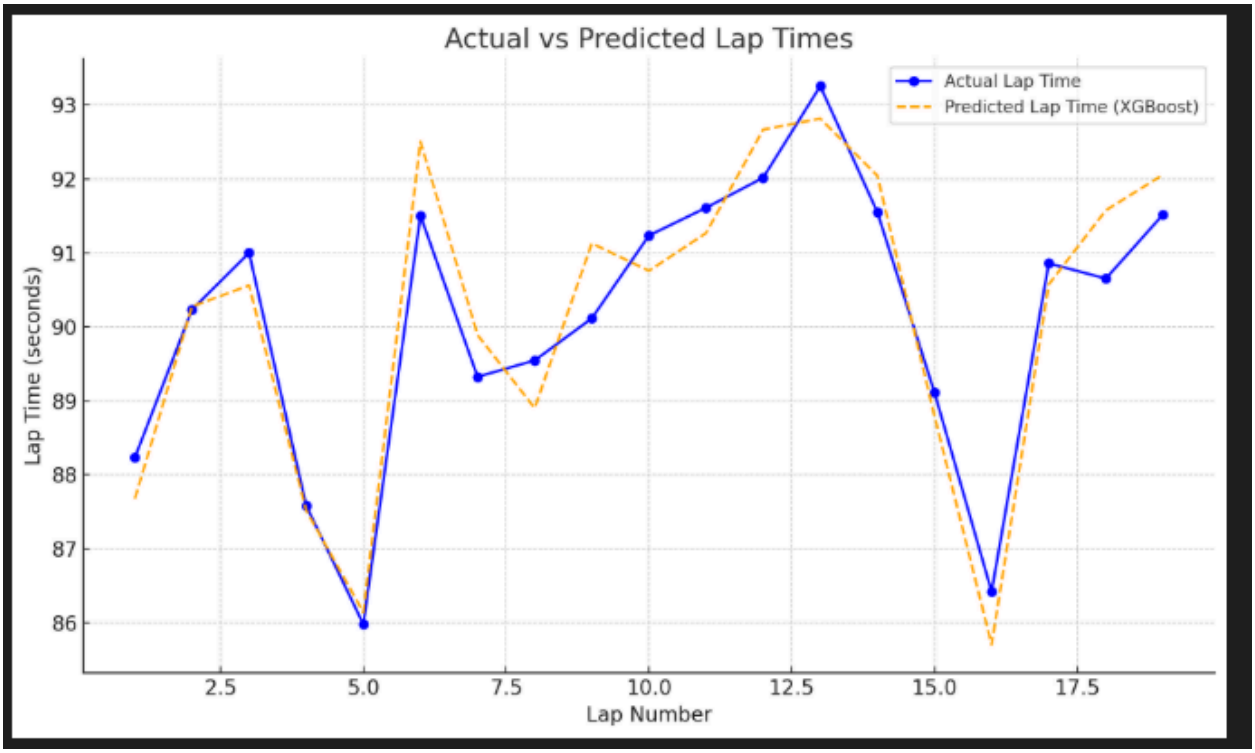

Fig 4 : Actual vs Predicted Lap time (Synthetic Data)

(G) Benefits of the Proposed Idea

Using AI in Formula 1 can help teams make faster and smarter decisions. Races move quickly, and teams often have just a few seconds to decide when to pit or change strategy. A trained model can predict things like tire wear or the best lap to pit, helping teams plan ahead instead of reacting too late.

Another benefit is that models like XGBoost are easy to understand. They show which factors (like throttle or tire temp) are affecting results the most. This helps engineers explain decisions to team leaders and drivers. Also, once the model is trained, it can work across different tracks and weather, making it a tool that teams can use all season.

(H) Literature Support for the Proposal

Other researchers have shown that supervised learning models work well in Formula 1.

For example, Noe and Patel (2024) used real lap times and telemetry to build a model that predicted performance and pit stops. Even with public data, their model matched what real teams did.

Todd et al. (2025) used XGBoost to predict tire wear and added tools that explained how driver behavior (like braking) affected the tires. Their system also lets teams test “what-if” changes like what happens if the driver uses less brake pressure.

These studies show that regression models like XGBoost are accurate, flexible, and explainable perfect for the fast, data-heavy world of F1.

Discussion

Challenges with the Proposed Solution

One of the biggest challenges using AI in Formula 1 is getting enough high-quality data to train the models. F1 teams keep their telemetry and performance data as highly confidential because it gives them a competitive edge. As a result, most of this data isn't available to the public or researchers, which makes it hard to build models that work well across different race tracks, weather conditions, and team setups.

Even when data is available, it's often messy. Telemetry can include sensor errors, missing values, or inconsistencies due to communication delays. This makes cleaning and organizing the data a time-consuming task. On top of that, the unpredictable nature of Formula 1 racing like crashes, sudden rain, or safety cars adds complexity that models may struggle to handle if such events don’t appear often enough in the training data. Because supervised models learn from past examples, they may not do well when facing rare or unusual situations they’ve never seen before.

Another complication is that F1 rules and car designs change almost every season. This means that any model built on old data may need constant retraining, which increases the computational workload and adds to the challenge of keeping the AI system current and reliable.

Gaps in the Literature

While many studies look at AI in F1, most focus on just one thing like lap time or tire wear. Few combine multiple factors, like fuel, driver input, and tire life, into one model. Also, not much research checks whether a model trained on one season or race can work on another.

Some researchers use supervised learning, others use reinforcement learning, but not many try combining them to make smarter, more complete systems. There’s also more work to be done in making AI easy to understand and explain to engineers under race pressure.

Future Work and Immediate Next Steps

Future work should focus on solving the data availability problem. One idea is to use federated learning, where teams can train models locally without sharing their raw data. This way, data stays private, but teams can still benefit from larger shared models.

Another step is to explore deep learning models like LSTMs or Transformers that can better understand time-based data. These models could be especially good at tracking how race conditions evolve over multiple laps.

Combining supervised learning for prediction with reinforcement learning for decision-making could lead to even smarter systems like the ones that not only predict what will happen but also recommend what to do next. In addition, using generative AI to simulate rare race events (like crashes or safety cars) might help improve how models handle unpredictable situations.

Researchers should also keep working on explainability. Engineers need to understand why a model made a certain prediction, especially when millions of dollars and race outcomes are on the line.

Limitations of the Proposed Solution

Even though AI has a lot of potential in Formula 1, there are still some important limitations. One of the biggest problems is getting enough good data. F1 teams usually keep their telemetry and sensor data private, which makes it hard to train models that work well in real races. Without detailed and accurate data, the AI might not perform as expected.

Also, while models like XGBoost are great at making predictions from structured data, they don't always capture how things change over time. Racing is very fast-paced and depends on many changing conditions, so more advanced models might be needed to handle that better.

Another issue is that racing is unpredictable. Things like crashes, weather changes, or new rules can affect the outcome of a race, and these situations are hard for AI to predict. Since the models are trained on past data, they might not work well if something totally new happens.

Conclusion

Artificial intelligence is set to transform decision-making in Formula 1 by shifting from intuition-based strategies to data-driven ones. AI models, especially supervised learning techniques like XGBoost, can predict key metrics such as lap times, tire wear, and fuel consumption, which can help teams make faster, smarter race-day decisions. These tools reduce reliance on human guesswork by providing real-time, data-backed insights. However, challenges remain. Access to high-quality telemetry data is limited, and many models are trained on simulations rather than real race data. For AI to become fully integrated into F1 operations, teams will need more robust datasets, improved model adaptability, and seamless integration into race strategies. Combining regression, classification, and reinforcement learning will be essential for building systems that not only predict outcomes but also recommend optimal actions. Teams that embrace AI now will gain a lasting competitive edge which can shape the future of Formula 1.

References

- Noe, A., and R. Patel. *Using Telemetry Data for Machine Learning-Driven Lap Time Prediction and Race Strategy*. 2024.
- Todd, M., et al. *Explainable AI Models for Tire Energy Use in Formula 1*. 2025.
- Thomas, E., et al. *RSRL: Race Strategy Reinforcement Learning for Pit Stop Scheduling*.

2025.

- “F1 Tempo.” *Www.f1-Tempo.com*, www.f1-tempo.com/.
- Garcia, Raul. “Tyre Strategies in Formula 1 Using Python - Python in Plain English.” *Medium*, Python in Plain English, 11 Nov. 2023, python.plainenglish.io/tyre-strategies-in-formula-1-using-python-1df4df19bf85?gi=15252c09842d. Accessed 28 July 2025.
-
- “McLaren Racing - Dell Technologies.” *Www.mclaren.com*, www.mclaren.com/racing/partners/dell-technologies/.
- “How Formula 1® Uses Generative AI to Accelerate Race-Day Issue Resolution | Amazon Web Services.” *Amazon Web Services*, 18 Feb. 2025, aws.amazon.com/blogs/machine-learning/how-formula-1-uses-generative-ai-to-accelerate-race-day-issue-resolution/.