ocean

Formula 1 Racing Challenge: 2024 Mexico Grand Prix

Luca Ordronneau
Discord: lucanew#3793

10-24-2024

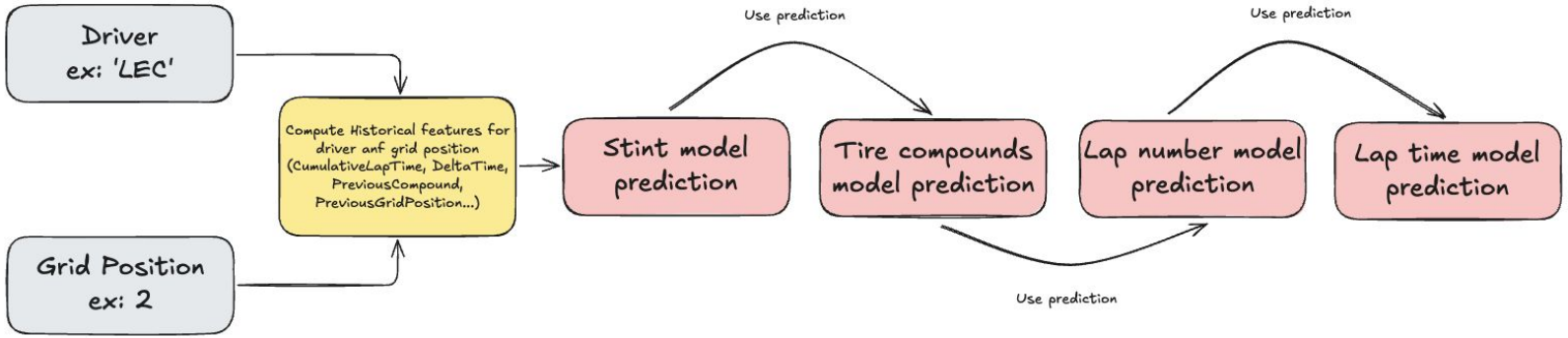
Introduction

This report provides an F1 race prediction model, specifically applied to the Mexican Grand Prix. The project involves preprocessing **ALL historical F1 data from 2018 to 2023 (using the FastF1 API)**, followed by predictive modeling for key race variables such as stint numbers, tire compounds, lap numbers, and lap times for each driver based on their grid position.

Using this project, users can input a list of **drivers** and their respective **grid positions** and receive predicted values for their stints, tire choices, and expected lap performance during the race. This pipeline integrates **multiple machine learning models** for each predictive task

In addition to the predictions, this report also provides access to visualizations for each driver at their respective grid positions. These visualizations help illustrate and analyze the predictions of the model's output.

Prediction Flow



Feature Engineering

Categorical Embedding Feature

Handle Categorical Variables Efficiently: Embeddings convert categorical data into continuous numerical vectors, enabling the model to process and learn from non-numeric features.

Capture Semantic Relationships: They represent categories in a multidimensional space where similar entities are placed closer together, allowing the model to recognize and utilize inherent relationships between categories.

Reduce Dimensionality (2-dimensions): Compared to methods like one-hot encoding, embeddings create lower-dimensional representations, which reduces computational complexity and the risk of overfitting, while retaining meaningful information.

DriverEmbeddings:

Description: Encoded representations of drivers, where each driver is mapped to a continuous vector capturing their characteristics.

Benefit: Allows the model to capture similarities and differences between drivers, improving generalization and handling categorical variables effectively.

EventNameEmbeddings:

Description: Embeddings for race events (e.g., specific Grand Prix), encapsulating track-specific characteristics in a numerical format.

Benefit: Helps the model account for the unique features of each track, such as layout and conditions, which influence performance and strategy.

Feature Engineering

Cumulative Mean Shifted Features

Capture Historical Trends: Cumulative features aggregate historical data (e.g., weighted cumulative means), allowing the model to learn from past performance and trends over time.

Smooth Out Variability: By averaging over multiple observations, they reduce the impact of anomalies and noise in the data, leading to more stable predictions.

Enhance Contextual Understanding: Incorporating cumulative information provides a broader context for the model, improving its ability to make informed predictions based on long-term patterns.

WeightedCumulativeMeanLapTimeSeconds:

Description: This feature represents the driver's historical average lap times, calculated as a weighted cumulative mean over past races.

Benefit: It helps the model understand the driver's performance trend over time, smoothing out anomalies and emphasizing consistent patterns.

WeightedCumulativeMeanSpeedI1:

Description: The cumulative average speed of a driver in the first sector (SpeedI1) of the track across previous races.

Benefit: Captures long-term performance in specific track sectors, allowing the model to predict future speeds based on historical data.

Feature Engineering

Shifted Features

Prevent Data Leakage: Shifting features ensure that only past information is used for predictions, maintaining the chronological integrity of the data and avoiding the use of future information.

Model Temporal Dependencies: They capture recent changes or events that can influence the target variable, helping the model understand how the immediate past affects the future.

Improve Predictive Accuracy: By including lagged variables, shifted features allow the model to account for autocorrelation and temporal patterns, enhancing its ability to forecast outcomes.

ShiftedCompoundEmbedding:

Description: The embedding vector of the tire compound used in the previous stint, shifted to align with the current stint.

Benefit: Enables the model to consider the influence of the last tire choice on the current decision without causing data leakage, as it only uses past information.

ShiftLapNumber:

Description: The lap number from the previous stint or lap, shifted to the current context.

Benefit: Helps the model understand temporal dependencies by incorporating how previous laps affect future performance, enhancing predictive accuracy.

Number of stints

Model & Performance

Model Card

Type: Regression XGBoost Model **- Target:** *Number of Stints*

Feature Importance

CircuitMaxLapNumber (0.1374): The most influential feature, representing the race's maximum lap number, indicating that longer races require more stints.

EventNameEmbeddings and **Year** (0.0593, 0.0452, 0.0387): Event characteristics and the season year significantly impact stint strategies due to track features and regulatory changes.

WeightedCumulativeMeanPitTimeSeconds and **Speed Metrics** (0.0531, 0.0526, 0.0495): Historical pit times and driver speeds suggest that faster teams and drivers influence the number of stints.

Compound Embeddings (0.0394, 0.0358): Tire compound choices play a crucial role, emphasizing the importance of tire strategy in determining stint numbers.

GridPositionEmbeddings (0.0332, 0.0324): Starting position affects stint planning due to race dynamics and overtaking opportunities.

NumberOfRaces (0.0309): Driver experience influences stint strategies, implying that seasoned drivers may adopt different approaches.

Hyperparameter Optimization

GridSearchCV with 5-fold cross-validation was used to tune hyperparameters over a grid of 108 combinations, totaling 540 fits.

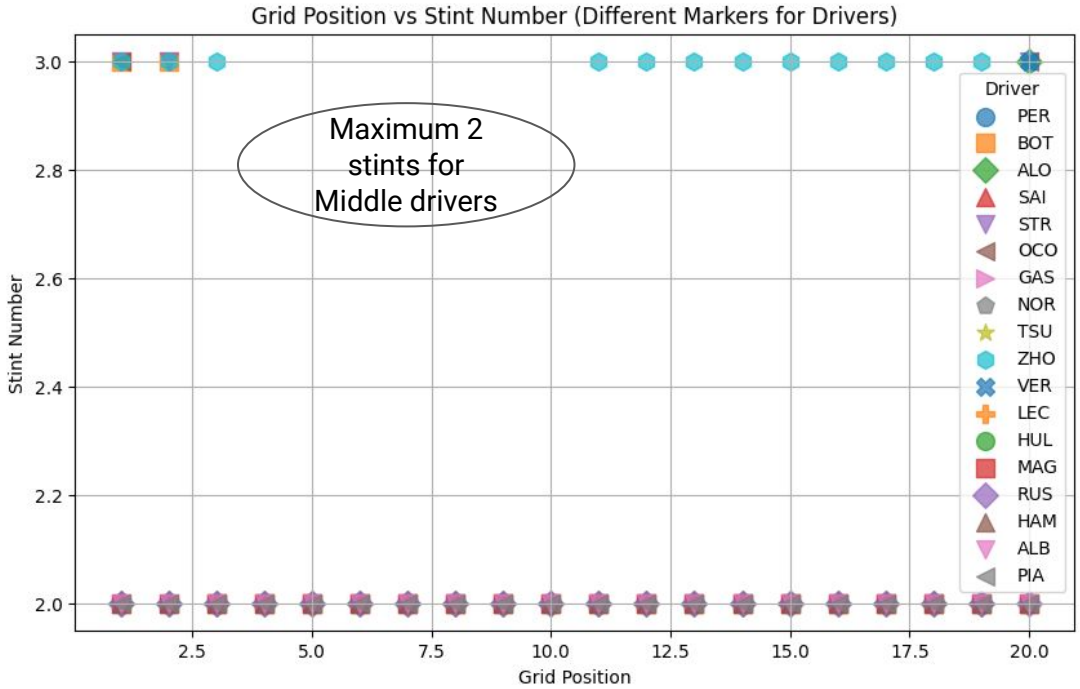
Best Hyperparameters Found:

- colsample_bytree: 0.8
- learning_rate: 0.01
- max_depth: 5
- n_estimators: 300
- subsample: 0.8

The MSE of **0.6087** indicates that the model is reasonably accurate, reducing error significantly compared to a naive baseline.

Number of stints

Predictions



Tire Compound

Model & Performance

Model Card

Type: Classification XGBoost Model **- Target:** *Tire Compound*

Feature Importance

ShiftedCompoundEmbedding_1 (0.1243) and **ShiftedCompoundEmbedding_0 (0.1080)**: These are the most influential features, representing the embeddings of the previous tire compound used. This indicates that the last compound choice heavily impacts the next one.

Stint (0.0572) and **StintNumber (0.0488)**: The current stint number and total number of stints in the race significantly affect tire compound selection, reflecting strategic adjustments during the race.

EventNameEmbeddings_0 (0.0539) and **EventNameEmbeddings_1 (0.0410)**: Encoded representations of the event highlight that track-specific characteristics influence tire choices due to varying conditions and requirements.

CircuitMaxLapNumber (0.0475) and **Year (0.0474)**: The maximum lap number of the event and the year suggest that race length and season-specific regulations play a role in determining tire compounds.

GridPositionEmbedding_1 (0.0424) and **GridPositionEmbedding_0 (0.0359)**: Starting grid positions, encoded as embeddings, imply that initial race positions affect tire strategy decisions.

WeightedCumulativeMeanCompoundEmbeddings_0 (0.0288) and **WeightedCumulativeMeanCompoundEmbeddings_1 (0.0253)**: Historical average embeddings of tire compounds used emphasize the importance of past compound choices on current decisions.

Hyperparameter Optimization

GridSearchCV with 5-fold cross-validation was used to tune hyperparameters over a grid of 32 combinations, totaling 160 fits.

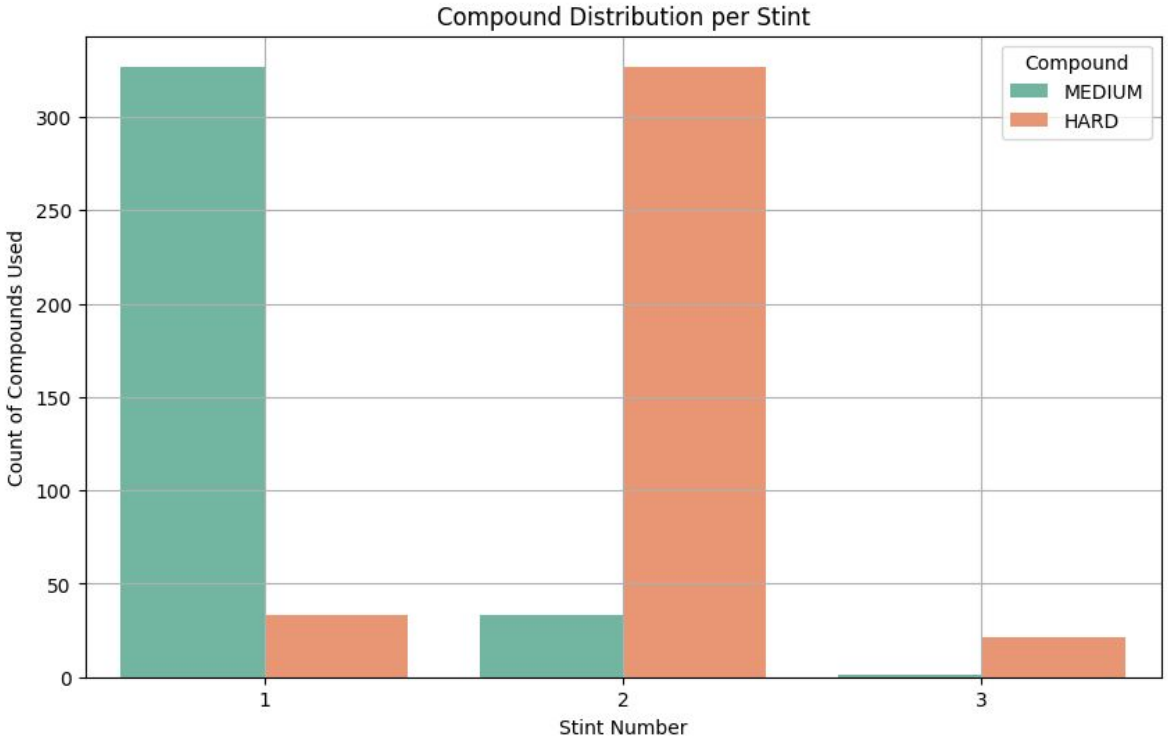
Best Hyperparameters Found:

- colsample_bytree: 0.8
- learning_rate: 0.1
- max_depth: 6
- n_estimators: 200
- subsample: 1

The model achieved an accuracy of **71.63%** on the test set, indicating it correctly predicts the tire compound in approximately 72% of cases.
The model performs well on major classes like **HARD**, **MEDIUM**, and **SOFT** tires, with precision and recall scores ranging from **69%** to **79%**.

Number of stints

Predictions



Laps per Stint
Model & Performance

Model Card

Type: Regression XGBoost Model **- Target:** *Laps per Stint*

Feature Importance

CompoundEmbedding_0 (0.1626) and **StintNumber** (0.1526): The most influential features, indicating that the type of tire compound used and the total number of stints significantly affect the number of laps in a stint.

CircuitMaxLapNumber (0.0680): Represents the maximum lap number of the event, suggesting that longer races influence stint lengths.

ShiftLapNumberToGo (0.0646) and **ShiftLapNumber** (0.0641): Previous lap numbers and laps remaining are crucial, indicating that both past performance and remaining race distance impact stint planning.

EventNameEmbeddings_0 (0.0524) and **EventNameEmbeddings_1** (0.0334): Encoded event characteristics highlight that track-specific features affect the number of laps per stint due to varying conditions.

Stint (0.0323) and **ShiftLapNumberByStint** (0.0264): The current stint number and laps in previous stints emphasize the importance of race progression and strategic adjustments over time.

Hyperparameter Optimization

GridSearchCV with 5-fold cross-validation was used to tune hyperparameters over a grid of 108 combinations, totaling 540 fits.

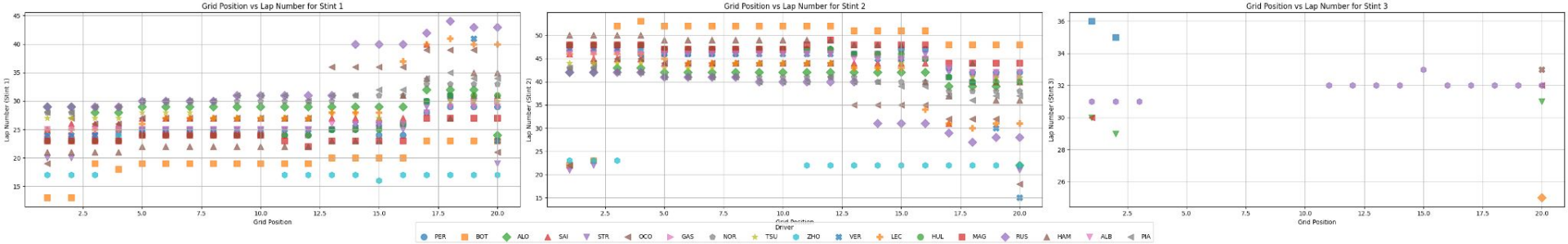
Best Hyperparameters Found:

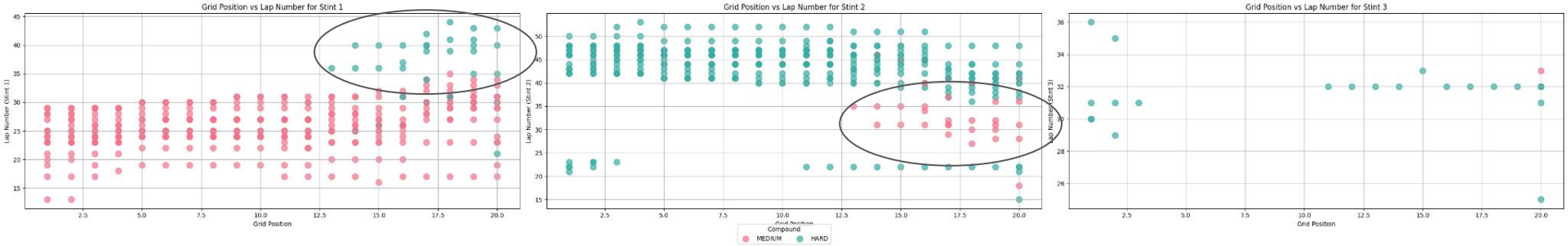
- colsample_bytree: 1
- learning_rate: 0.05
- max_depth: 6
- n_estimators: 200
- subsample: 0.8

The **Test Mean Squared Error (MSE)** of **60.2893** quantifies the average squared difference between the predicted and actual number of laps per stint on the test dataset.
This means that, on average, the model's predictions deviate from the actual number of laps per stint by approximately **7.76 laps** (RMSE)

Laps per Stint

Predictions





Back

Average Lap Time per Stint

Model & Performance

Model Card

Type: Regression XGBoost Model **- Target:** *Lap Time per Stint*

Feature Importance

WeightedCumulativeMeanLapTimeSeconds (0.3121): The most influential feature, representing the driver's historical average lap times. This indicates that a driver's past performance heavily influences their current lap times.

CircuitMaxLapNumber (0.2033): The maximum lap number of the event, suggesting that the length of the race significantly impacts lap times due to factors like fuel load and tire degradation.

EventNameEmbeddings (0.1032 and 0.0330): Encoded representations of the event highlight that track-specific characteristics affect lap times because different circuits have varying layouts and conditions.

CompoundEmbeddings (0.0647 and 0.0588): Representations of the tire compounds used, emphasizing the importance of tire choice on lap performance, as different compounds offer varying levels of grip and durability.

LapNumber (0.0328) and **LapNumberToGo** (0.0193): The current lap and remaining laps in the race indicate that lap times fluctuate throughout the race due to changing fuel loads, tire wear, and track conditions.

Year (0.0235): Suggests that season-specific factors like regulations or technological advancements impact lap times across different years.

Hyperparameter Optimization

GridSearchCV with 5-fold cross-validation was used to tune hyperparameters over a grid of 108 combinations, totaling 540 fits.

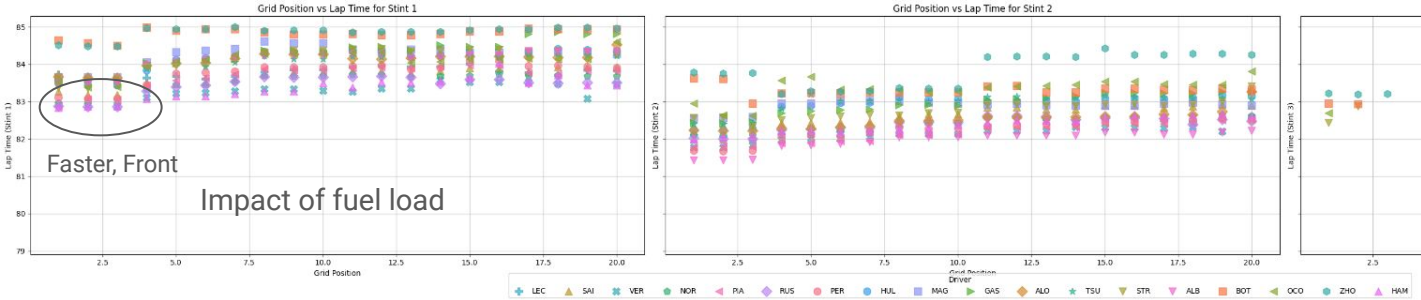
Best Hyperparameters Found:

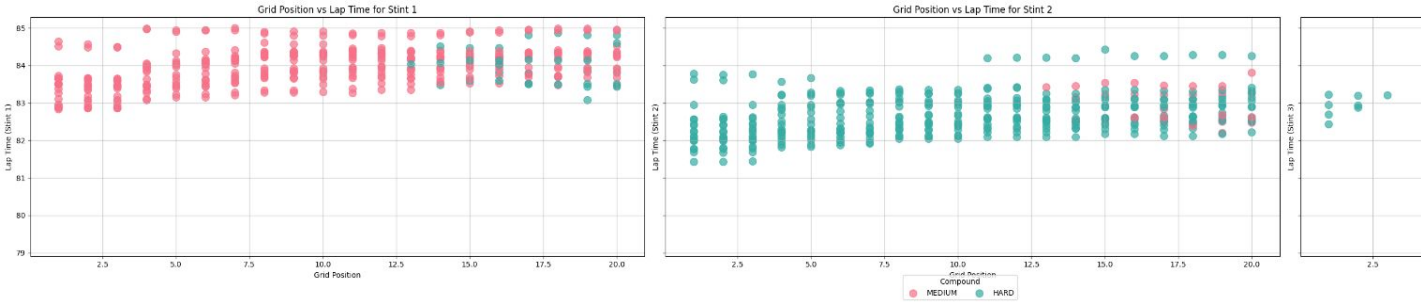
- colsample_bytree: 0.8
- learning_rate: 0.05
- max_depth: 6
- n_estimators: 300
- subsample: 0.8

Test MSE: **4.4903.** The RMSE of approximately **2.12 seconds** means that, on average, the model's predicted lap times deviate from the actual lap times by about **2.12 seconds** per lap.

Average Lap Time per Stint

Predictions





Conclusion

Aspect	What the Model Does	What the Model Does Not Do
Driver Performance and Variability	- Adapts predictions for different driver types (front runners, mid-field, back runners) by incorporating **DriverEmbeddings** and historical performance metrics like **WeightedCumulativeMeanLapTimeSeconds** and **NumberOfRaces**. - Accounts for varying starting grid positions using features like **GridPosition** and **GridPositionEmbedding**, effectively handling differences between drivers.	
Adaptability to Race Conditions	- Utilizes **EventNameEmbeddings** to capture track-specific characteristics that may influence race conditions and strategies. Take indirectly conditions from previous year with historical erformance	- Does not explicitly incorporate real-time external variables such as current weather conditions, on-track incidents, or other unpredictable events during a race. I decided to not include weather conditions (and its a risk) to focus on other features, maybe in future development
Innovative Methodology and Model Flexibility	- Employs advanced feature engineering techniques like **cumulative features** (e.g., WeightedCumulativeMeans) and **shifted features** to capture temporal dynamics without data leakage. - Uses **embedding features** for drivers, teams, events, and compounds to represent categorical variables in a meaningful numerical format, **capturing complex relationships**. - Allows input for individual drivers and is adaptable to different race scenarios.	
Interpretability and Explanation of Predictions	- Provides overall **feature importances** to highlight which variables most significantly impact predictions, aiding in understanding the model's focus areas. - Uses interpretable features such as starting grid position, historical lap times, and tire compounds, making the general logic behind predictions more accessible.	- Lacks detailed explanations on how specific feature interactions influence each prediction.