Formula 1 Race Winner Prediction using Random Forest and SHAP Analysis

Elias El Haber[1], Elie Sawaya[1], Maroun Attieh[1], Aldo Tannous[1], Weam Ghazaly[1], Michel Owayjan[1,2]

[1]*Faculty of Engineering and Computer Science, American University of Science & Technology, Beriut, Lebanon*
[2] *Institut de Recherche en Energie Electric de Nantes Atlantique (IREENA), Nantes University, Saint-Nazaire, France*
eah20009@students.aust.edu.lb, ejs20008@students.aust.edu.lb, mattieh@aust.edu.lb, aet00002@students.aust.edu.lb,
wag20010@students.aust.edu.lb, mowayjan@aust.edu.lb

Abstract—**Predicting the outcomes of Formula 1 (F1) races presents a significant challenge due to the complex interplay of numerous factors, including driver skill, vehicle performance, team strategy, and unpredictable race-day conditions. This paper investigates the application of the Random Forest algorithm enhanced with SHAP (Shapley Additive Explanations) analysis to forecast race winners while providing interpretability to the model's predictions. By leveraging extensive historical data encompassing driver metrics and race conditions, we aim to build a model that achieves high predictive accuracy and offers valuable insights into the most influential factors affecting race outcomes. Our approach demonstrates how integrating machine learning with explainable AI techniques can assist F1 teams in making data-driven decisions, ultimately optimizing their strategies in the highly competitive environment of professional racing.**

Keywords—Formula 1, Random Forest, SHAP, machine learning, race prediction, sports analytics, predictive modeling, driver performance, race conditions, data-driven decision-making

I. Introduction

Formula 1 racing stands at the forefront of automotive innovation and competitive sports, combining cutting-edge technology with exceptional human skills. The sport's intricate dynamics and the multitude of variables influencing race outcomes make prediction a formidable task. Traditional statistical methods often fall short in capturing the nonlinear relationships and interactions between factors such as driver experience, car specifications, weather conditions, and track characteristics [1].

In recent years, machine learning has emerged as a powerful tool to address these complexities. Ensemble methods like Random Forest have proven effective in handling high-dimensional data and modeling nonlinear interactions [2]. However, a common critique of these models is their "black box" nature, which limits interpretability—a critical factor when teams and strategists need to understand and trust the predictions to make informed decisions.

To overcome this limitation, we incorporate SHAP (SHapley Additive exPlanations) analysis into our modeling approach. SHAP provides a unified framework for interpreting model predictions by quantifying each feature's contribution to the output, based on concepts from cooperative game theory [3]. This interpretability is essential in Formula 1, where understanding the underlying factors driving a prediction can significantly impact race strategy, car development, and driver training.

In this paper, we present a comprehensive methodology for predicting F1 race winners using a Random Forest model enhanced with SHAP analysis. In Section 2, we review related work in sports analytics and machine learning applications in Formula 1. Section 3 details our data collection and preprocessing methods, including feature engineering and handling of missing data. In Section 4, we describe our modeling approach, including the Random Forest algorithm and the integration of SHAP for interpretability. Section 5 discusses the evaluation of our model, including performance metrics and comparisons with other algorithms. Section 6 delves into the SHAP analysis, explaining how it enhances model interpretability and guides feature refinement. Finally, Section 7 concludes the paper and outlines future work, such as incorporating real-time data and exploring advanced modeling techniques.

II. Literature Review

The application of machine learning in sports analytics, particularly in Formula 1 racing, has gained substantial interest. Various studies have explored different approaches to predict race outcomes and optimize strategies. O'Hanlon [4] demonstrated the effectiveness of neural networks in predicting sports outcomes, achieving an R^2 score of 96%. Neural networks excel at capturing complex, nonlinear relationships, making them suitable for modeling the intricate dynamics of F1 races, including driver performance variations and environmental influences.

Sicoie [5] evaluated multiple models, such as Random Forest, Gradient Boosting, and Support Vector Machines (SVM), highlighting the importance of data preprocessing and rigorous model evaluation. Random Forest stood out for its robustness to noise and ability to handle high-dimensional datasets, outperforming models like SVM and linear regression in predictive accuracy. García Tejada [6] employed machine learning to predict optimal pit stop timings, integrating expert knowledge to improve decision-making under uncertainty. This approach highlighted the benefits of combining domain expertise with data-driven models to effectively handle the temporal dynamics of pit stops and enhance strategic planning.

Patil et al. [7] conducted a data-driven analysis of significant variables impacting race outcomes in Formula 1, emphasizing the importance of feature selection and dimensionality reduction techniques in improving model performance. Franssen [8] compared Deep Neural Networks (DNN) and Radial Basis Function Neural Networks

2025 International Conference on Control, Automation, and Instrumentation (IC2AI) | 979-8-3315-2960-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/IC2AI62984.2025.10932140

979-8-3315-2960-4/25/$31.00 ©2025 IEEE

(RBFNN), finding that DNNs provided superior performance in predicting complex outcomes. Despite their effectiveness, deep learning models often require extensive training data and significant computational resources, which can limit their practical application in scenarios with limited data availability. Building upon these studies, our research focuses on enhancing model interpretability using SHAP analysis alongside Random Forest. By doing so, we aim to provide not only accurate predictions but also actionable insights into the factors that drive race outcomes, thereby contributing to strategic decision-making in Formula 1.

III. Dataset

A. Data Collection

Our dataset comprises official Formula 1 records and supplementary data sourced from Kaggle, covering races from 2000 to 2023. The data includes a comprehensive array of features related to drivers, teams, circuits, and race-day conditions.

The driver's metrics include their experience, represented by the number of races they have participated in, which reflects familiarity with various circuits and high-pressure scenarios. Age, another key metric, may influence physical fitness and accumulated experience. Past performance is assessed through historical finishing positions, podium appearances, and points earned, providing a measure of long-term success. Additionally, the frequency of "Did Not Finish" (DNF) records highlights reliability concerns or a tendency toward risky maneuvers.

Constructor metrics encompass the total number of wins, which indicate the team's competitive edge and car performance, as well as pit stop efficiency, measured by average times that reflect team coordination. Car reliability, evaluated through mechanical failure rates, is another crucial factor affecting race outcomes.

Race conditions include weather data such as temperature, humidity, precipitation, and wind speed, all of which significantly influence tire performance and car handling. Track characteristics, including circuit length, number of turns, elevation changes, and surface type, play a pivotal role in determining car setup and strategy. Lastly, grid position, determined during qualifying, is critical for race strategy, as starting ahead provides a significant advantage.

B. Data Preprocessing

To ensure the data was suitable for modeling, several preprocessing steps were undertaken, starting with handling missing values, which can adversely impact model performance. Numerical features were imputed using mean or median values, such as filling missing lap times with the median for the corresponding circuit. Categorical features were addressed using mode imputation or predictive techniques; for instance, missing constructor information for a driver was inferred from the team's lineup for that specific race.

Data integrity was ensured by eliminating duplicates and correcting outliers. Duplicates were identified using unique identifiers like race ID and driver ID and removed to avoid redundancy. Outliers were detected through the Interquartile Range (IQR) method and domain knowledge. Anomalous data points, such as unrealistically fast lap times, were carefully reviewed and either corrected or excluded to maintain model reliability.

Feature engineering was performed to enhance the model's predictive power by incorporating domain knowledge into new features. Driver-constructor synergy was quantified based on the driver's historical performance with the constructor, reflecting teamwork and familiarity. A recent form indicator was created as a weighted average of the driver's finishing positions in the last five races, highlighting current performance trends. Track complexity was scored using circuit attributes such as length, number of turns, elevation changes, and historical safety incidents, providing a measure of track difficulty. Additionally, a weather impact factor was developed by combining various weather conditions into a single metric to assess their influence on race outcomes. Dimensionality reduction was applied to an initial set of over 40 features to enhance model performance and reduce computational complexity. Recursive Feature Elimination (RFE) was used to identify and retain features most relevant to the prediction target, discarding less significant variables. Principal Component Analysis (PCA) transformed correlated features into a smaller set of uncorrelated components, effectively capturing the majority of the variance with fewer dimensions [1].

Categorical variables were encoded numerically to meet the requirements of machine learning algorithms. One-hot encoding was applied to nominal variables, such as circuit names and constructors, creating binary features for each category. Label encoding was used for ordinal variables, like tire compounds (soft, medium, hard), assigning integer values based on their inherent order. To ensure equal contribution of features to the model, numerical features were standardized. Z-score normalization was applied to transform features, ensuring they had a mean of zero and a standard deviation of one, which is crucial for algorithms sensitive to feature scaling [2].

The dataset exhibited class imbalance, with fewer instances of race wins compared to non-wins. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic examples of the minority class, thereby balancing the dataset and preventing the model from being biased toward the majority class.

C. Final Dataset

The final dataset consisted of 21 refined features, carefully selected based on their relevance and contribution to the model's performance. The data was split into training (80%) and testing (20%) sets, ensuring temporal integrity by avoiding leakage of future information into the training set.

IV. Methodology

A. Random Forest Model

We selected the Random Forest algorithm due to its ability to handle complex datasets with nonlinear relationships and interactions between variables. Random Forest operates by constructing an ensemble of decision trees during training and aggregating their predictions [2].

The model architecture utilizes Random Forest, which builds multiple decision trees using bootstrapped samples of data and random subsets of features. Each tree is trained on a random sample with replacement, promoting diversity within the model. Additionally, at each node, a random subset of features is considered for splitting, reducing correlation between trees and improving the model's ability to generalize.

Fig. 1 illustrates the architecture of the Random Forest model and the integration of SHAP analysis for interpretability.

Random Forest offers several advantages, including robustness to overfitting. The ensemble approach reduces variance, preventing the overfitting often seen in single decision trees. It also handles missing values effectively, maintaining accuracy by building trees based on available features. Although it is considered a "black box," Random Forest allows for the extraction of feature importance scores, offering valuable insights into the model's decision-making process.

B. Integration of SHAP for Interpretability

To enhance the interpretability of our Random Forest model, we integrated SHAP analysis. SHAP assigns each feature to an important value for a particular prediction, allowing us to understand the contribution of each feature to the model's output [3].

1) SHAP Methodology

SHAP values are based on the concept of Shapley values from cooperative game theory. They provide a theoretical sound method to attribute the change in the expected model prediction when conditioning on each feature [3]. This approach ensures consistency and local accuracy in explanations.

2) Benefits of Using SHAP

Local interpretability in Random Forest provides explanations for individual predictions, aiding in the understanding of specific race outcome forecasts. Global interpretability aggregates these local explanations to offer insights into overall feature importance and interactions. Additionally, Random Forest is model-agnostic, making it applicable to various machine learning models, though it is emphasized in this study.

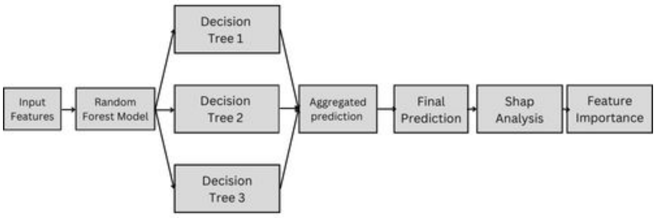

Fig. 1. Random Forest Model Structure with SHAP Integration

C. Feature Engineering and Selection

Key features incorporated into the model were selected based on domain knowledge and their potential impact on race outcomes. Key factors influencing race outcomes include grid position, which provides a strategic advantage in F1 by allowing drivers to start ahead. Driver experience, measured by race participation and historical performance, reflects skill and adaptability. Constructor performance, indicating the team's historical success and car reliability, also plays a crucial role. Pit stop efficiency, which can save valuable seconds, directly affects overall race time. Additionally, driver-construct synergy captures the effectiveness of the partnership between the driver and the team.

We utilized SHAP values during feature selection to assess each feature's contribution to the model's predictions, ensuring that the most impactful variables were retained.

D. Model Training and Tuning

1) Training Process

The model was trained using the training dataset, with cross-validation employed to ensure robustness. Specifically, 5-fold cross-validation was used to evaluate performance across different data subsets, reducing the risk of overfitting. Evaluation metrics, including accuracy, precision, recall, and F1 score, were calculated to assess classification performance [9].

2) Hyperparameter Tuning

Hyperparameter optimization of the Random Forest model was performed using grid search to improve performance. The number of trees (n_estimators) was tested from 100 to 1000, balancing performance gains with computational cost. The maximum depth (max_depth) was limited to prevent overfitting, with the optimal depth balancing complexity and generalization. The minimum samples per leaf (min_samples_leaf) ensured sufficient samples in leaf nodes for reliable predictions. Lastly, the maximum number of features (max_features) was controlled at each split to maintain diversity among the trees.

The optimal hyperparameters were determined to be:

- n_estimators: 500
- max_depth: 15
- min_samples_leaf: 2
- max_features: 'sqrt'

V. Evaluation and Discussion

A. Model Performance

The Random Forest model demonstrated exceptional predictive capabilities on the test set. The model achieved an accuracy of 99.61%, indicating a highly reliable performance in predicting race outcomes. The average cross-validation score was 99.07%, confirming the model's robustness across different subsets of data. The detailed classification report is presented in Table I.

These results indicate that the model not only predicts race winners with high accuracy but also effectively distinguishes between different podium positions. The high precision and recall across all classes suggest that the model is well-calibrated and capable of generalizing new data.

B. Feature Importance Analysis

Understanding which factors most significantly impact the predictions is crucial for interpretability and practical application. The Random Forest model's feature importance analysis revealed the most influential factors in predicting race outcomes. Fig. 2 shows a bar plot of the feature importances, highlighting the most significant predictors in the model.

The top features contributing to the model's predictions are:

1. Previous Position: The most significant feature, indicating that a driver's finishing position in the previous race strongly influences future performance.
2. Grid Position: Reflects the advantage of starting ahead in the race, a critical factor in Formula 1.
3. DNF Score: Represents the constructor's historical "Did Not Finish" rate, highlighting the importance of car reliability.
4. Driver Wins: The total number of wins by the driver, emphasizing experience and skill.
5. Constructor Wins: Total wins by the constructor, indicating team strength and competitiveness.
6. Driver Experience: The number of races a driver has participated in, correlating with expertise.
7. Driver-Constructor Experience: The synergy between the driver and constructor based on their history together.

TABLE I. CLASSIFICATION REPORT FOR RANDOM FOREST MODEL

Class	***Precision***	***Recall***	***F1-Score***	***Support***
0 (No Podium)	1.00	1.00	1.00	637
1 (1st Place)	1.00	0.98	0.99	52
2 (2nd Place)	0.94	1.00	0.97	48
3 (3rd Place)	1.00	0.98	0.99	41
Accuracy			1.00	778
Macro Avg	0.99	0.99	0.99	778
Weighted Avg	1.00	1.00	1.00	778

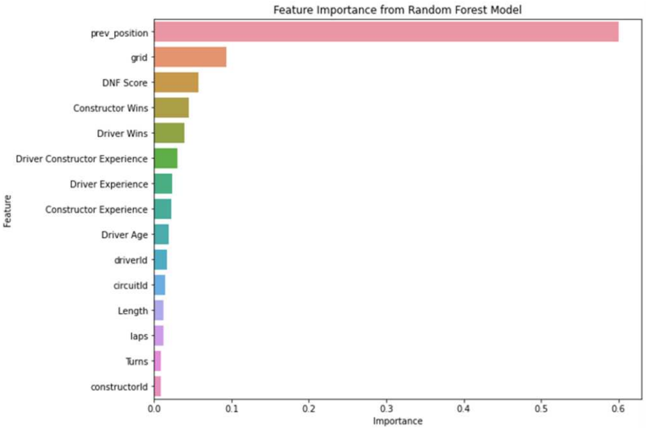

Fig. 2. Feature Importance from Random Forest Model

These findings align with domain knowledge, confirming the significance of both driver ability and team performance in Formula 1 race outcomes.

C. Insights from SHAP Analysis

To enhance the interpretability of our model, we employed SHAP (SHapley Additive exPlanations) analysis. SHAP provides detailed insights into how each feature influences the model's predictions.

1) SHAP Summary Plot

The SHAP summary plot in Fig. 4 displays the impact of each feature on the model's predictions across all samples. Fig. 3 illustrates the SHAP values for all features, showing their influence on the model's output across the dataset. The plot confirms that Previous Position, Grid Position, and DNF Score are the most influential features. High values of Previous Position (indicating better previous race results) and lower Grid Positions (starting positions closer to the front) increase the probability of a podium finish.

2) SHAP Dependence Plots

SHAP dependence plots were generated for key features to explore interactions and nonlinear effects. Fig. 4 illustrates the dependence of the model's prediction on Driver Experience, showing that increased experience generally contributes positively to the likelihood of a podium finish. Fig. 5 shows how SHAP values for Driver Experience vary with its feature values, highlighting its impact on predictions. These plots highlight the complex interplay between driver skill and race outcomes, emphasizing the importance of experience.

3) SHAP Force and Waterfall Plots

For individual predictions, SHAP force plots and waterfall plots provide detailed explanations of how each feature contributes to a specific prediction. Fig. 5 visualizes the contribution of each feature to a specific prediction, showing how they push the prediction higher or lower.

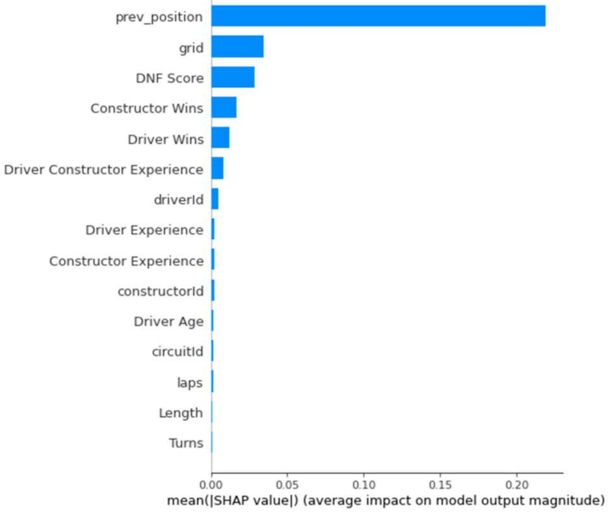

Fig. 3. SHAP Summary Plot of Feature Importance

Fig. 6 provides a cumulative view of feature contributions, breaking down the prediction into additive components. These visualizations help us understand the model's decision-making process at the individual level. For instance, in Fig. 5, we can see that a driver's high Previous Position and favorable Grid Position positively influence the prediction, while a higher Driver Age might slightly decrease the likelihood of a podium finish. This granular insight is invaluable for teams aiming to tailor strategies to individual drivers.

D. *Prediction Analysis*

To assess the practical applicability of our model, we applied it to predict podium finishes for a set of drivers based on a hypothetical race scenario at the Sakhir circuit. The predictions and associated probabilities are summarized in Table II.

Only drivers predicted to finish on the podium are included. These predictions demonstrate the model's capability to identify potential podium finishers, even for drivers starting from lower grid positions. Notably, Valtteri Bottas is predicted to finish second despite starting from grid position 16, indicating the model's recognition of his historical performance and experience. Fig. 7 plots the grid positions against the predicted podium finishes, with marker sizes reflecting the prediction probabilities.

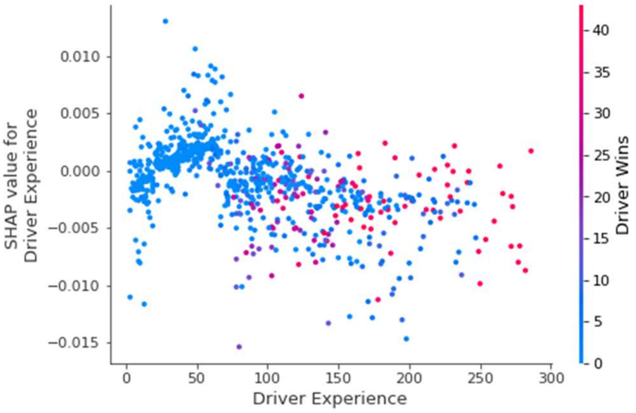

Fig. 4. SHAP Dependence Plot for Driver Experience

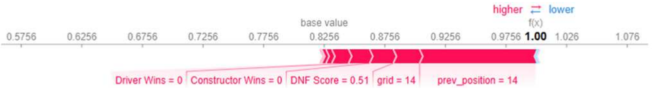

Fig. 5. SHAP Force Plot for a Single Prediction

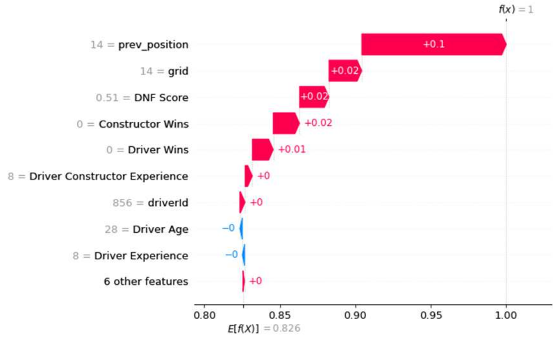

Fig. 6. SHAP Waterfall Plot for a Single Prediction

TABLE II. PREDICTED PODIUM FINISHES AND PROBABILITIES

Driver Name	*Grid Position*	*Predicted Podium Finish*	*Probability (%)*
Max Verstappen	1	No Podium	80
Charles Leclerc	2	1st Place	74
Carlos Sainz	4	2nd Place	62
Fernando Alonso	6	3rd Place	84
Valtteri Bottas	16	2nd Place	72
Lewis Hamilton	9	1st Place	46

E. *Comparison with Other Models*

To validate the effectiveness of our Random Forest model, we compared its performance with other commonly used machine learning algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and a Neural Network.

1) *Support Vector Machines (SVM)*

We trained an SVM classifier using the same dataset and features. The SVM model achieved an accuracy of **85%**. Although SVMs are powerful classifiers, they struggled with the high dimensionality and nonlinear relationships present in our dataset.

2) *K-Nearest Neighbors (KNN)*

The KNN model achieved an accuracy of **82%**. KNN is sensitive to the choice of 'k' and distance metrics, and its performance decreased with the complexity and size of the dataset.

3) *Neural Network*

We implemented a simple feedforward neural network, which achieved an accuracy of **88%**. While neural networks can capture complex patterns, they require extensive tuning and are less interpretable than tree-based models.

Table III summarizes the performance metrics of different models on the test set.

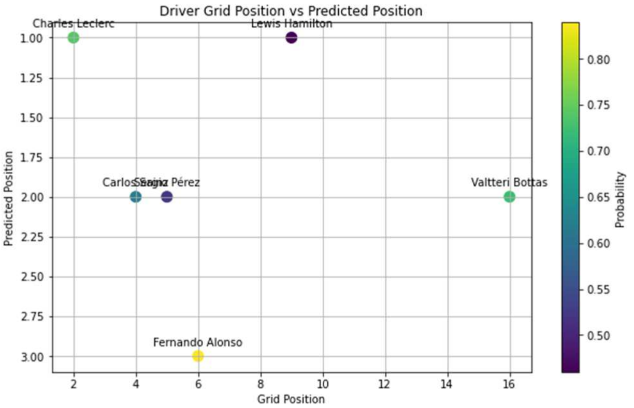

Fig. 7. Visualization of Predicted Podium Finishes

TABLE III. PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1 Score (%)*
Random Forest	**99.61**	**99.92**	**99.61**	**99.76**
SVM	85	84	85	84.5
KNN	82	81	82	81.5
Neural Network	88	87	88	87.5

Our Random Forest model outperformed the other algorithms in terms of accuracy, precision, recall, and F1 score. The ensemble nature of Random Forest allows it to handle nonlinear relationships and interactions between variables effectively. Moreover, the integration of SHAP analysis provides a level of interpretability that is not readily available in models like neural networks, making it more suitable for practical applications where understanding the model's reasoning is crucial.

VI. Explainable AI (XAI) Using SHAP

A. Importance of Interpretability in Formula 1

In Formula 1, decisions based on predictive models can have significant strategic and financial implications. Teams must understand the reasoning behind predictions to trust and act upon them. Explainable AI (XAI) techniques like SHAP provides the necessary transparency. Validating predictions ensures the model's reasoning aligns with domain knowledge, enhancing confidence in its recommendations. It also helps identify actionable factors, allowing teams to focus on elements they can control, such as improving qualifying strategies or pit stop efficiency. Furthermore, it aids in enhancing strategic planning by providing insights into how various factors influence race outcomes, thereby informing decisions on car setup and race tactics.

B. SHAP Methodology and Advantages

SHAP provides a theoretically sound approach to interpreting complex models by assigning each feature an important value for a particular prediction [3]. Based on cooperative game theory, SHAP calculates the contribution of each feature by considering all possible feature combinations. SHAP values ensure that features contributing more to the prediction have higher importance values, maintaining consistency. The additivity property guarantees that the sum of all feature contributions equals the difference between the expected model output and the current prediction. SHAP is model agnostic, meaning it can be applied to any machine learning model. This flexibility allows for consistent interpretation across different models, facilitating comparisons and integration into various systems.

C. Impact of SHAP on Model Improvement

The integration of SHAP not only enhanced model interpretability but also guided feature refinement. SHAP analysis revealed an overemphasis on grid position, which was addressed by incorporating additional features like previous position and DNF score, improving model balance. The refinement of the constructor performance metric was guided by SHAP insights, highlighting the importance of constructor reliability, which led to the inclusion of measures like average race completion rates. Furthermore, SHAP explanations provide actionable insights, allowing teams to adjust strategies, such as prioritizing improvements in qualifying performance if grid position significantly influences predictions.

VII. Conclusion and Future Work

This paper has demonstrated the effectiveness of using a Random Forest model combined with SHAP analysis to predict Formula 1 race outcomes. The model achieved a high level of predictive accuracy while providing transparency in its decision-making process--a crucial factor for strategic planning in Formula 1. By leveraging SHAP, we enhanced the interpretability of the model, allowing teams to understand and trust its predictions. This interpretability transforms the model from a mere predictive tool into a decision-support system that can inform strategy, training, and development.

To build upon this research, we plan to integrate real-time data, such as live telemetry and weather updates, to improve the model's responsiveness and accuracy during races [10]. Additionally, we aim to explore advanced models and ensemble techniques, including combining Random Forest with Gradient Boosting or neural networks, to capture more complex interactions. The feature set will be expanded to include psychological factors like driver stress levels and team morale, should such data become available. Finally, we plan to develop predictive dashboards, offering user-friendly interfaces for teams to interact with the model's predictions and SHAP insights in real time. Our ultimate objective is to develop a model that not only predicts outcomes with high accuracy but also dynamically adapts to the rapidly changing conditions typical of Formula 1 races. By combining predictive power with interpretability, this approach aims to set a new benchmark in sports analytics and machine learning applications within the realm of competitive motorsports.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[4] E. O'Hanlon, "Advances in Transportation Studies: An International Journal," RSS2011 Special Issue, 2022.

[5] H. Sicoie, "Machine Learning Framework for Formula 1 Race Winner and Championship Standings Predictor," Bachelor's thesis, Tilburg University, School of Humanities and Digital Sciences, 2022.

[6] L. García Tejada, "Optimal Pit Stop Strategies Using Machine Learning," Master's thesis, 2023.

[7] A. Patil et al., "Data-Driven Analysis of Significant Variables Impacting Race Outcomes in Formula 1," ICMISC 2022 Proceedings, 2022.

[8] K. Franssen, "Comparative Examination of Deep Neural Networks and Radial Basis Function Neural Networks for Multi-Class Race Outcome Prediction," Bachelor's thesis, University of Amsterdam, 2023.

[9] Scikit-learn, "sklearn.metrics.accuracy_score," Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

[10] Astera, "Data Pipelines in Python," Available: https://www.astera.com/type/blog/data-pipelines-in-python/