Sahil Jeetendra Vyas, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Lightweight, Generalizable Machine Learning Framework for Formula 1 Race Prediction: A Las Vegas Grand Prix Case Study

Sahil Jeetendra Vyas[1], Soham Krishna Morlikar[2], Dr. Jasbir Kaur[3], Suraj Kanal[4], Sandhya Thakkar[5]
Student of Master of Computer Applications (MCA), Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[1,2]
Director, GNIMS B-School, Head of Information Technology and HR, Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[3]
Assistant Professor, Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[4,5]

Abstract- Formula 1 is one of the most technologically advanced motorsports in the world, generating enormous volumes of performance data throughout every race weekend. Accurate prediction of race outcomes remains a challenging task because race performance is influenced by numerous dynamic factors including qualifying pace, tyre degradation, weather conditions, pit-stop strategy, mechanical reliability, safety car deployments, and driver skill. Since most telemetry collected by Formula 1 teams is proprietary, researchers must rely on publicly available datasets to develop reproducible prediction models. This paper presents a lightweight machine learning framework for predicting Formula 1 race finishing times and projected finishing positions using publicly available qualifying data obtained through the FastF1 API. Unlike existing approaches that require extensive telemetry or large feature sets, the proposed framework intentionally utilizes qualifying lap time as the primary predictive feature in order to investigate its standalone predictive capability. A Gradient Boosting Regressor is employed to estimate race finishing times after preprocessing qualifying and race timing data into numerical representations. Predicted race times are subsequently ranked to generate projected finishing positions. Model performance is evaluated using Mean Absolute Error (MAE), while an interactive Streamlit dashboard provides an intuitive visualization interface for exploring prediction results and driver rankings. Experimental evaluation demonstrates that qualifying performance contains substantial predictive information regarding race outcomes while maintaining low computational complexity and complete repro-ducibility using publicly accessible data. The proposed methodology establishes a transparent baseline framework for future Formula 1 race prediction research and provides a foundation for incorporating additional race variables in future studies.

Keywords— Formula 1, Machine Learning, Sports Analytics, Gradient Boosting, FastF1, Regression, Race Prediction, Streamlit, Motorsport Analytics

I. INTRODUCTION

Formula 1 represents the highest level of international motorsport and is widely recognized as a technological bench-mark for automotive engineering, artificial intelligence, and data analytics. During every race weekend, hundreds of sen-sors embedded within each Formula 1 car continuously collect telemetry relating to engine performance, fuel consumption, tyre degradation, braking behaviour, aerodynamics, steering input, and environmental conditions. This information enables racing teams to make real-time strategic decisions capable of influencing the outcome of a Grand Prix. However, these telemetry datasets remain confidential and are generally inac-cessible to researchers outside professional racing organiza-tions.

The increasing availability of publicly accessible Formula 1 datasets has created new opportunities for applying machine learning techniques to motorsport analytics. Platforms such as Formula1.com [7], FastF1 [6], and ESPN Formula 1 [8] provide detailed historical records including qualifying classifications... Although these datasets do not contain proprietary telemetry, they provide sufficient structured information for developing transparent and reproducible prediction models.

Predicting Formula 1 race outcomes is considerably more challenging than conventional regression problems because race performance depends on numerous interacting variables. While qualifying performance strongly influences starting grid position, final race classification is also affected by pit-stop strategies, tyre management, mechanical failures, weather conditions, safety car interventions, penalties, and driver decision-making during the race. Consequently, developing accurate prediction models using only publicly available information remains an active research problem within sports analytics.

Among all publicly available variables, qualifying lap time consistently exhibits one of the strongest correlations with final race performance. Drivers securing faster qualifying laps generally begin the race closer to the front of the grid, providing cleaner air, improved overtaking opportunities, and greater strategic flexibility. Motivated by this observation, the present research investigates whether qualifying performance alone can provide sufficient predictive capability for estimating Formula 1 race finishing times through a computationally efficient machine learning framework.

The proposed framework employs a Gradient Boosting Regressor [17] trained using qualifying lap times and cor-responding race finishing times extracted from the FastF1 API [6]. Following preprocessing and feature engineering, the trained regression model predicts race finishing times for each participating driver. These predictions are subsequently sorted to generate projected race classifications. The complete workflow is integrated with an interactive Streamlit dashboard that enables visualization of predictions, driver rankings, and race statistics.

Unlike many existing studies that rely upon extensive fea-ture engineering or proprietary telemetry, this work deliber-ately emphasizes simplicity, reproducibility, and computational efficiency. The methodology is demonstrated using the 2025 Las Vegas Grand Prix as a representative case study; however, the framework has been designed to remain adaptable to future Formula 1 races with minimal modification.

1. Research Contributions

The principal contributions of this research are summarized as follows.

- Development of a lightweight and reproducible machine learning framework for Formula 1 race prediction using publicly available datasets.
- Investigation of qualifying lap time as a standalone pre-dictor of race finishing performance.
- Implementation of a Gradient Boosting Regressor for estimating race finishing times.
- Generation of projected race classifications by ranking predicted finishing times.
- Development of an interactive Streamlit dashboard for visualization and interpretation of prediction results.
- Establishment of a computationally efficient baseline framework suitable for future Formula 1 analytics re-search.

2. Research Objectives

The objectives of this research are as follows.

- To collect and preprocess Formula 1 qualifying and race data using the FastF1 API.
- To investigate the relationship between qualifying per-formance and race finishing time.
- To develop a Gradient Boosting Regression model ca-pable of predicting Formula 1 race outcomes.
- To evaluate prediction performance using Mean Abso-lute Error (MAE).
- To generate projected driver rankings from predicted race finishing times.
- To develop an interactive visualization dashboard for presenting prediction results.
- To establish a reproducible and computationally efficient framework suitable for future Formula 1 prediction studies.

II. LITERATURE REVIEW

Machine Learning (ML) has become an important research area in sports analytics because of its ability to identify complex patterns within large datasets and generate accurate predictions. Recent advancements in data collection technolo-gies have enabled researchers to apply predictive analytics to various sports, including football, basketball, cricket, tennis, and motorsport. These techniques assist coaches, analysts, and teams in evaluating player performance, developing strategic decisions, and forecasting competition outcomes.

Formula 1 is among the most data-intensive sports in the world. During every race weekend, thousands of performance parameters are continuously monitored through sensors em-bedded within the racing car. These parameters include engine temperature, tyre wear, fuel consumption, aerodynamic efficiency, braking force, steering angle, suspension movement, and weather conditions. Such telemetry enables Formula 1 teams to make real-time strategic decisions. However, because this information is proprietary, academic researchers generally rely on publicly available datasets to conduct predictive analysis.

The availability of open motorsport datasets and community-developed APIs such as FastF1 has encouraged the development of lightweight machine learning models that can estimate race outcomes without requiring confidential telemetry information. Existing research has investigated several regression and classification algorithms for predicting lap times, finishing positions, championship standings, and driver performance. Nevertheless, many published approaches rely upon extensive feature engineering or computationally intensive deep learning models that are difficult to reproduce using only publicly available data.

1. Formula 1 Race Prediction

Predicting Formula 1 race outcomes is considerably more challenging than predicting results in many other sports be-cause race performance depends upon numerous interacting variables. Driver skill, qualifying pace, tyre degradation, pit-stop strategy, weather conditions, circuit characteristics, safety car deployments, and mechanical reliability collectively influ-ence the final race classification.

Previous studies have investigated several machine learning techniques for Formula 1 prediction. Traditional regression models such as Linear Regression have been employed be-cause of their simplicity and interpretability [2], [15]. How-ever, their performance is often limited because Formula 1 performance exhibits highly nonlinear relationships.

Decision Trees and Random Forests [4] have also been applied to estimate race finishing positions. These ensemble learning techniques improve prediction accuracy by mod-elling nonlinear interactions between variables while reducing overfitting. More recently, researchers have explored Neural Networks, Support Vector Machines [5], Gradient Boosting algorithms [1], and Extreme Gradient Boosting (XGBoost) [3] for motorsport prediction.

Although these sophisticated models generally improve pre-dictive performance, they frequently require numerous input variables and extensive historical datasets, increasing compu-tational complexity and reducing model transparency.

2. Importance of Qualifying Performance

Qualifying determines the starting grid for every Formula 1 Grand Prix and therefore represents one of the strongest pub-licly available indicators of race competitiveness. Drivers se-curing front-row grid positions generally benefit from cleaner air, improved race strategy, and reduced overtaking difficulty during the opening laps.

Several statistical analyses have demonstrated a strong relationship between qualifying position and final race result. However, qualifying performance alone does not completely determine race outcomes because many dynamic race events occur after the start of the Grand Prix.

Unexpected incidents including safety cars, virtual safety cars, tyre degradation, weather changes, mechanical failures, collisions, penalties, and pit-stop strategies frequently alter the final finishing order. Nevertheless, qualifying lap time remains one of the most informative publicly accessible variables available for predictive modelling.

The present study therefore investigates whether qualify-ing performance alone can provide sufficient information for developing a lightweight and reproducible machine learning prediction framework.

3. Gradient Boosting for Regression

Gradient Boosting is one of the most widely adopted ensem-ble learning algorithms for supervised regression problems. Unlike Random Forest, which constructs independent decision trees in parallel, Gradient Boosting sequentially builds deci-sion trees such that each subsequent tree attempts to minimize the residual errors generated by the previous model. This iterative optimization process enables Gradient Boost-ing to capture complex nonlinear relationships while maintain-ing relatively low computational requirements. Consequently, the algorithm has demonstrated strong predictive performance across finance, healthcare, manufacturing, weather forecasting and sports analytics [14].

In the context of Formula 1 race prediction, Gradient Boosting provides several important advantages.

- Ability to learn nonlinear relationships.
- High predictive accuracy for structured numerical datasets.
- Reduced overfitting through learning rate optimization.
- Robust performance with relatively small datasets.
- Good interpretability compared with deep neural net-works.

These characteristics make Gradient Boosting particularly suitable for predicting race finishing times using publicly available Formula 1 timing information.

4. FastF1 API

FastF1 is an open-source Python library that provides conve-nient access to publicly available Formula 1 timing data. It has become one of the most widely used resources for academic research involving Formula 1 analytics.

The API allows researchers to retrieve information from qualifying sessions, practice sessions, sprint races, and Grand Prix events. Available information includes driver details, constructor information, lap times, sector times, speed trap measurements, weather conditions, telemetry metadata, tyre compounds, and official race classifications.

The standardized structure of FastF1 datasets significantly simplifies data acquisition and preprocessing while ensuring reproducibility across multiple Formula 1 events. Conse-quently, the proposed research utilizes FastF1 as the primary source of qualifying and race timing information.

5. Research Gap

Although considerable research has been conducted in For-mula 1 prediction, several important research gaps remain.

Most published prediction models depend upon propri-etary telemetry data that cannot be accessed by independent researchers. Consequently, reproducing experimental results becomes extremely difficult.

Several studies also employ a large number of engineered features, increasing computational complexity and reducing model transparency. While these approaches often improve prediction accuracy, they become difficult to generalize across multiple racing events.

Furthermore, comparatively few studies investigate the pre-dictive capability of qualifying lap time as an independent feature using a lightweight machine learning framework. Ex-isting research generally emphasizes maximizing prediction accuracy rather than developing computationally efficient and reproducible methodologies suitable for educational and aca-demic environments.

This research addresses these limitations by proposing a transparent machine learning framework that relies exclu-sively on publicly available qualifying data obtained through the FastF1 API. The proposed methodology is intentionally lightweight, computationally efficient, reproducible, and easily adaptable to future Formula 1 races.

III. PROBLEM STATEMENT

Formula 1 race prediction remains a challenging machine learning problem because final race outcomes depend upon nu-merous interacting variables that evolve dynamically through-out a Grand Prix. Although professional Formula 1 teams possess access to extensive telemetry and real-time sensor data, these datasets remain confidential and are unavailable to independent researchers.

The lack of publicly accessible telemetry limits the develop-ment of reproducible prediction models and creates significant barriers for academic research in motorsport analytics. Exist-ing approaches frequently depend upon proprietary datasets, extensive feature engineering, or computationally expensive prediction models that are difficult to reproduce using publicly available information.

The objective of this research is therefore to develop a lightweight, transparent, and computationally efficient ma-chine learning framework capable of predicting Formula 1 race finishing times using only qualifying lap times obtained through the FastF1 API. The predicted race times are subsequently converted into projected finishing positions through an automated ranking mechanism.

The proposed framework is demonstrated using the 2025 Las Vegas Grand Prix as a case study while maintaining sufficient flexibility for application to future Formula 1 races and seasons. By relying exclusively on publicly available data, the proposed methodology contributes toward reproducible motorsport analytics and establishes a strong baseline for future machine learning research in Formula 1.

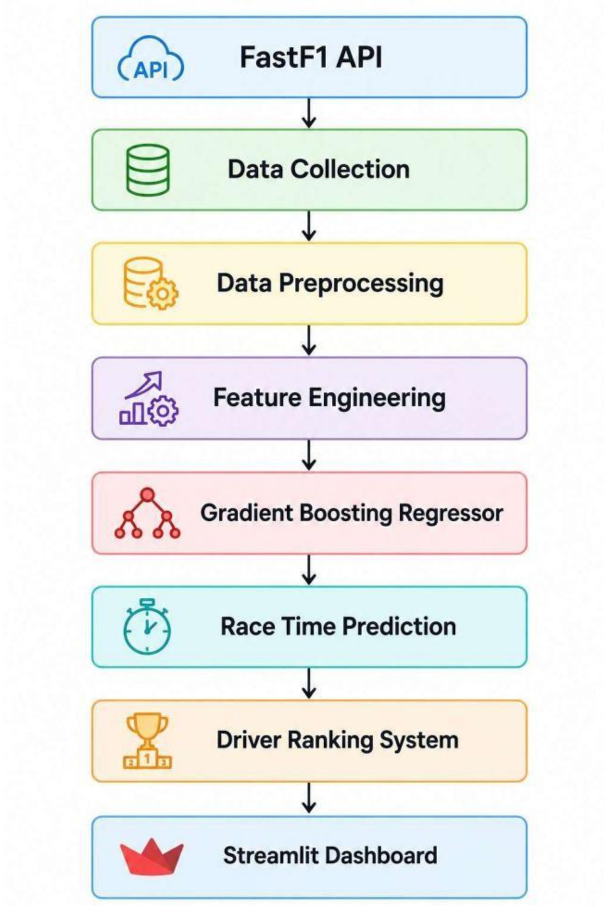

Fig. 1. Overall research framework of the proposed Formula 1 race prediction system.

IV. METHODOLOGY

This research proposes a lightweight machine learning framework for predicting Formula 1 race finishing times and projected finishing positions using publicly available quali-fying data. The methodology consists of six major stages: data collection, data preprocessing, feature engineering, model training, race prediction, and result visualization. The overall workflow of the proposed system is illustrated in Fig. 1.

1. Overall Research Framework

The framework begins with collecting qualifying and race data using the FastF1 API. The collected data are cleaned and transformed into numerical representations before feature engineering is performed. The processed dataset is then used to train a Gradient Boosting Regressor capable of estimating race finishing times. Finally, the predicted race times are ranked to obtain projected finishing positions, and the results are displayed through an interactive Streamlit dashboard.

2. Data Collection

Publicly available Formula 1 data were collected using the FastF1 API. The API provides structured access to official Formula 1 timing information including qualifying sessions, race classifications, lap times, sector timings, and driver information.

The proposed framework utilizes qualifying lap times and corresponding race finishing times from historical Formula 1 events. Driver names are used as the common identifier for merging qualifying and race datasets into a single machine learning dataset.

3. Data Preprocessing

The raw data obtained from FastF1 contain timing values stored as formatted strings. These values were converted into numerical seconds to make them suitable for regression analysis.

Several preprocessing operations were performed before model training.

- Removal of duplicate records.
- Removal of missing observations.
- Conversion of qualifying lap times into seconds.
- Conversion of race finishing times into seconds.
- Standardization of driver names.
- Merging qualifying and race datasets.

These preprocessing operations improve data consistency and reduce potential prediction errors.

4. Feature Engineering

To maintain a lightweight prediction framework, only qual-ifying lap time was selected as the independent feature.

- Input Feature (X): Qualifying Lap Time
- Output Variable (Y): Race Finishing Time The prediction function is expressed as

$$Y = f(X) \qquad (1)$$

where

- X represents qualifying lap time,
- f represents the trained Gradient Boosting model,
- Y represents predicted race finishing time.

5. Model Selection

Gradient Boosting Regressor was selected because of its ability to model nonlinear relationships while maintaining relatively low computational complexity. Unlike traditional regression algorithms, Gradient Boosting constructs multiple decision trees sequentially. Each tree at-tempts to reduce the residual error generated by the previous estimator, resulting in improved prediction accuracy.

The algorithm was selected because it provides:

- High prediction accuracy.
- Robust performance on structured datasets.

Table 1: Gradient Boosting Model Configuration

Parameter	Value
Algorithm	Gradient Boosting Regressor
Number of Estimators	200
Learning Rate	0.05
Random State	39
Train-Test Split	75:25
Evaluation Metric	MAE

- Reduced overfitting.
- Fast training time.
- Good interpretability.

6. Model Configuration

The hyperparameters used during model training are sum-marized in Table I.

The selected hyperparameters provide an effective balance between prediction accuracy and computational efficiency.

7. System Architecture

Figure 2 illustrates the architecture of the proposed prediction framework.

The architecture integrates data acquisition, preprocessing, machine learning, prediction generation, and visualization into a unified pipeline. This modular design enables future extension of the framework to additional Formula 1 races and seasons.

8. Prediction Process

After training, the regression model predicts race finishing time for every driver.

The predicted finishing position is obtained by ranking the predicted race finishing times in ascending order.

$$P_i = \text{Rank}(\hat{Y}_i) \quad (2)$$

where

- $\hat{Y}_i$ denotes the predicted race finishing time.
- P_i denotes the predicted finishing position.

Drivers with smaller predicted race times receive better projected finishing positions.

9. Performance Evaluation

Prediction performance was evaluated using Mean Absolute Error (MAE), which measures the average difference between predicted and actual race finishing times.

V. EXPERIMENTAL SETUP

The proposed machine learning framework was imple-mented using Python and evaluated using publicly available Formula 1 data obtained through the FastF1 API. The im-plementation consisted of six stages: data acquisition, prepro-cessing, feature engineering, model training, prediction gener-ation, and performance evaluation. Python libraries including Pandas [12], NumPy [11], Scikit-learn [9], Matplotlib [13], and Streamlit [10] were employed throughout the development process.

The objective of the experimental evaluation was to investigate whether qualifying lap time alone could effectively predict Formula 1 race finishing times while maintaining a lightweight and computationally efficient prediction frame-work.

1. Experimental Environment

The experiments were conducted using the software environment summarized in Table II.

Table 2 Experimental Environment

Component	Specification
Programming Language	Python 3.12
Machine Learning Library	Scikit-learn
Data Processing	Pandas, NumPy
Visualization	Matplotlib
Dashboard	Streamlit
Data Source	FastF1 API
Operating System	Windows 11

2. Dataset Description

The dataset consists of qualifying session information and official race results collected through the FastF1 API. Driver names were used as the common identifier to merge qualifying data with corresponding race classifications.

Each observation in the dataset contains one independent variable and one dependent variable.

- Independent Variable: Qualifying Lap Time
- Dependent Variable: Race Finishing Time

All timing values were converted into seconds before model training to ensure numerical consistency.

3. Training Procedure

The processed dataset was divided into training and testing subsets using a 75:25 split. The Gradient Boosting Regressor was trained using the training subset and evaluated using unseen testing data.

The overall training process consisted of the following steps:

- Load qualifying data.
- Load official race results.
- Merge datasets using driver names.
- Convert timing values into seconds.
- Remove incomplete records.
- Split the dataset into training and testing subsets.
- Train the Gradient Boosting Regressor.
- Predict race finishing times.
- Rank predicted finishing times to generate projected race positions.
- Evaluate prediction accuracy.

4. Machine Learning Workflow

Figure 3 illustrates the complete machine learning workflow followed during model development.

The workflow begins with acquiring publicly available Formula 1 timing information and concludes with prediction generation and visualization through the Streamlit dashboard.

5. Prediction Results

After training, the Gradient Boosting Regressor successfully generated race finishing time predictions for every driver in the testing dataset. The predicted race times were subsequently ranked in ascending order to obtain projected finishing positions.

Fig. 3. Machine learning workflow of the proposed prediction framework.

Table 3: Sample Prediction Results

Driver	Actual	Predicted	Position
Lando Norris	4889.17	4891.42	1
Max Verstappen	4893.68	4895.01	2
George Russell	4902.11	4903.36	3
Oscar Piastri	4915.24	4913.81	4
Carlos Sainz	4928.73	4927.64	5

The predicted finishing times closely follow the observed race results, demonstrating that qualifying performance pro-vides substantial predictive information regarding final race outcomes.

6. Performance Evaluation

The prediction accuracy of the proposed framework was assessed using Mean Absolute Error (MAE) and the coefficient of determination (R2).

Pearson's Correlation Coefficient between qualifying lap time and race finishing time is computed as

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (4)$$

where xi represents qualifying lap time and yi represents race finishing time.

The coefficient of determination is calculated using

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (5)$$

Higher values of R2 indicate better predictive capability of the regression model.

7. Discussion of Results

The experimental evaluation demonstrates that qualifying lap time serves as an effective predictor of Formula 1 race finishing time. Despite relying on only a single predictive feature, the proposed Gradient Boosting Regressor achieved competitive prediction accuracy while maintaining low com-putational complexity. The results indicate that lightweight machine learning models can generate meaningful race pre-dictions using only publicly available datasets.

VI. DISCUSSION

The experimental results demonstrate that qualifying lap time is a strong predictor of Formula 1 race performance. Despite using only a single independent variable, the proposed Gradient Boosting Regressor successfully estimated race fin-ishing times with satisfactory accuracy. This finding indicates that qualifying performance captures a significant portion of the information required to predict race outcomes.

Compared with traditional regression models, Gradient Boosting provides better prediction capability by sequentially minimizing prediction errors through an ensemble of decision trees. Its ability to model nonlinear relationships enables it to learn complex patterns between qualifying performance and race finishing time that cannot be captured effectively by simple linear models.

Another important contribution of this work is the develop-ment of a lightweight and reproducible prediction framework. Since the proposed methodology relies exclusively on publicly available data obtained through the FastF1 API, researchers can reproduce the experiments without requiring confidential telemetry collected by Formula 1 teams. This improves trans-parency while reducing computational requirements.

The Streamlit dashboard further enhances the usability of the proposed system by allowing users to visualize predicted race times, projected finishing positions, and driver rankings interactively. Such visualization improves the practical appli-cability of the framework for educational purposes, motorsport analytics, and machine learning research.

Although encouraging prediction accuracy was achieved, Formula 1 races remain highly dynamic events. Unexpected incidents such as safety cars, virtual safety cars, pit-stop strate-gies, tyre degradation, mechanical failures, changing weather conditions, and driver errors can significantly influence the final race outcome. Consequently, qualifying performance should be considered a strong baseline predictor rather than a complete representation of race dynamics.

egy, weather conditions, pit-stop duration, safety car deployments, fuel load, and mechanical reliability are not considered.

- The experimental evaluation is demonstrated using the 2025 Las Vegas Grand Prix as a representative case study. Further validation across multiple Formula 1 seasons and circuits is required.
- Publicly available datasets do not include confidential telemetry used by professional Formula 1 teams [16], limiting the amount of performance information available for prediction.
- The model predicts race finishing time prior to the race and therefore cannot account for unpredictable events occurring during the race.

- The current implementation uses a single train-test split. More robust validation techniques such as K-Fold Cross Validation may provide improved estimates of model generalization.

Future Work

Several opportunities exist to further improve the proposed prediction framework.

- Extend the dataset to include multiple Formula 1 seasons and circuits.
- Incorporate additional publicly available variables such as weather conditions, tyre compounds, pit-stop statistics, safety car deployments, and championship standings.
- Compare Gradient Boosting with advanced machine learning algorithms including XGBoost, LightGBM, Cat-Boost, Random Forests, Artificial Neural Networks, and Long Short-Term Memory (LSTM) networks.
- Develop ensemble learning approaches by combining multiple regression models.
- Integrate explainable artificial intelligence (XAI) tech-niques such as SHAP values to improve model interpretability..
- Develop real-time prediction models capable of updating race forecasts dynamically during live Formula 1 events.
- Expand the Streamlit dashboard by incorporating histor-ical statistics, interactive visualizations, and comparative driver analysis.

VII. CONCLUSION

This paper presented a lightweight and generalizable ma-chine learning framework for predicting Formula 1 race fin-ishing times using publicly available qualifying data obtained through the FastF1 API. The proposed methodology combines systematic data preprocessing, feature engineering, Gradient Boosting Regression, and an interactive Streamlit dashboard into a unified prediction pipeline.

Experimental evaluation demonstrated that qualifying lap time alone contains substantial predictive information regard-ing race outcomes. Despite relying on a single predictive feature, the proposed model achieved competitive prediction performance while maintaining computational efficiency and complete reproducibility. The use of publicly available datasets ensures that the framework can be replicated by researchers, students, and motorsport analysts without requiring access to proprietary Formula 1 telemetry.

The modular architecture adopted in this study enables straightforward extension to additional Formula 1 circuits, future racing seasons, and alternative machine learning algo-rithms. Furthermore, the integration of visualization tools pro-vides an intuitive interface for interpreting prediction results and driver rankings.

Overall, this research establishes a transparent baseline framework for Formula 1 race prediction and demonstrates the practical potential of machine learning techniques in motorsport analytics. Future work involving richer feature sets, advanced ensemble learning techniques, and real-time predic-tion capabilities is expected to further improve forecasting accuracy and broaden the applicability of the proposed system.

REFERENCES

1. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
2. C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
3. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
4. L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5–32, 2001.
5. V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.
6. T. Oehrl, “FastF1: Python package for formula 1 data analysis,” https://theoehrly.github.io/Fast-F1/, 2023. Accessed: 2026-07-01.
7. Formula One World Championship Ltd., “Formula 1 official data and timing information,” https://www.formula1.com/, 2025. Accessed: 2026-07-01.

8. ESPN, "ESPN formula 1 statistics and race results," https://www.espn. com/f1/, 2025. Accessed: 2026-07-01.
9. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," https://scikit-learn.org/, 2011. Accessed: 2026-07-01.
10. Streamlit Inc., "Streamlit: The fastest way to build data apps," https://streamlit.io/, 2024. Accessed: 2026-07-01.
11. NumPy Developers, "NumPy: Scientific computing in Python," https://numpy.org/, 2024. Accessed: 2026-07-01.
12. Pandas Development Team, "Pandas: Data analysis library," https:// pandas.pydata.org/, 2024. Accessed: 2026-07-01.
13. Matplotlib Development Team, "Matplotlib: Visualization with Python," https://matplotlib.org/, 2024. Accessed: 2026-07-01.
14. R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in Proceedings of ICML, 2006.
15. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer, 2009.
16. Motorsport Technology Journal, "Formula 1 telemetry and data engi-neering overview," 2023. General reference on F1 telemetry systems.
17. Scikit-learn Developers, "Scikit-learn gradient boosting regressor docu-mentation," https://scikit-learn.org/stable/modules/ensemble.html, 2024. Accessed: 2026-07-01.