Formula 1 Race Prediction System: A Machine Learning Approach for 2025 Season

Matteo Tarocco

May 31, 2025

1 Formula 1 Race Prediction System: A Machine Learning Approach for 2025 Season

1.1 Authors

Matteo Tarocco

1.2 Abstract

This paper presents a machine learning-based system for predicting Formula 1 race outcomes, specifically focused on the 2025 season. The system combines historical race data, driver performance metrics, and track characteristics to generate race predictions. We demonstrate the system's effectiveness using the 2025 Catalunya Grand Prix as a case study.

1.3 1. Introduction

Formula 1 racing represents one of the most technologically advanced and data-driven sports in the world. Predicting race outcomes involves complex interactions between various factors including driver skill, team performance, track characteristics, and weather conditions. This paper presents a comprehensive approach to race prediction using machine learning techniques.

Our prediction system leverages historical race data, real-time performance metrics, and sophisticated machine learning algorithms to forecast race outcomes. The system takes into account various factors such as qualifying performance, clean air pace, track characteristics, and team performance scores.

1.4 2. Data Collection and Processing

1.4.1 2.1 Data Sources

Our system integrates data from multiple sources: - Historical race data from FastF1 API (2024-2025 seasons) - Comprehensive track characteristics database covering all F1 circuits - Team and driver performance metrics from official F1 data - Weather conditions and forecasts - Real-time qualifying and practice session data

1.4.2 2.2 Feature Engineering

The model incorporates several key features that influence race outcomes: - Qualifying times: Representing raw pace and one-lap performance - Clean air race pace: Measuring optimal racing speed without traffic - Track-specific characteristics: Including length, corners, and average speed - Team performance scores: Based on constructors' championship points - Weather conditions: Including temperature and rain probability - Tire degradation levels: Modeling tire wear impact on race pace

1.5 3. Methodology

1.5.1 3.1 Clean Air Pace Analysis

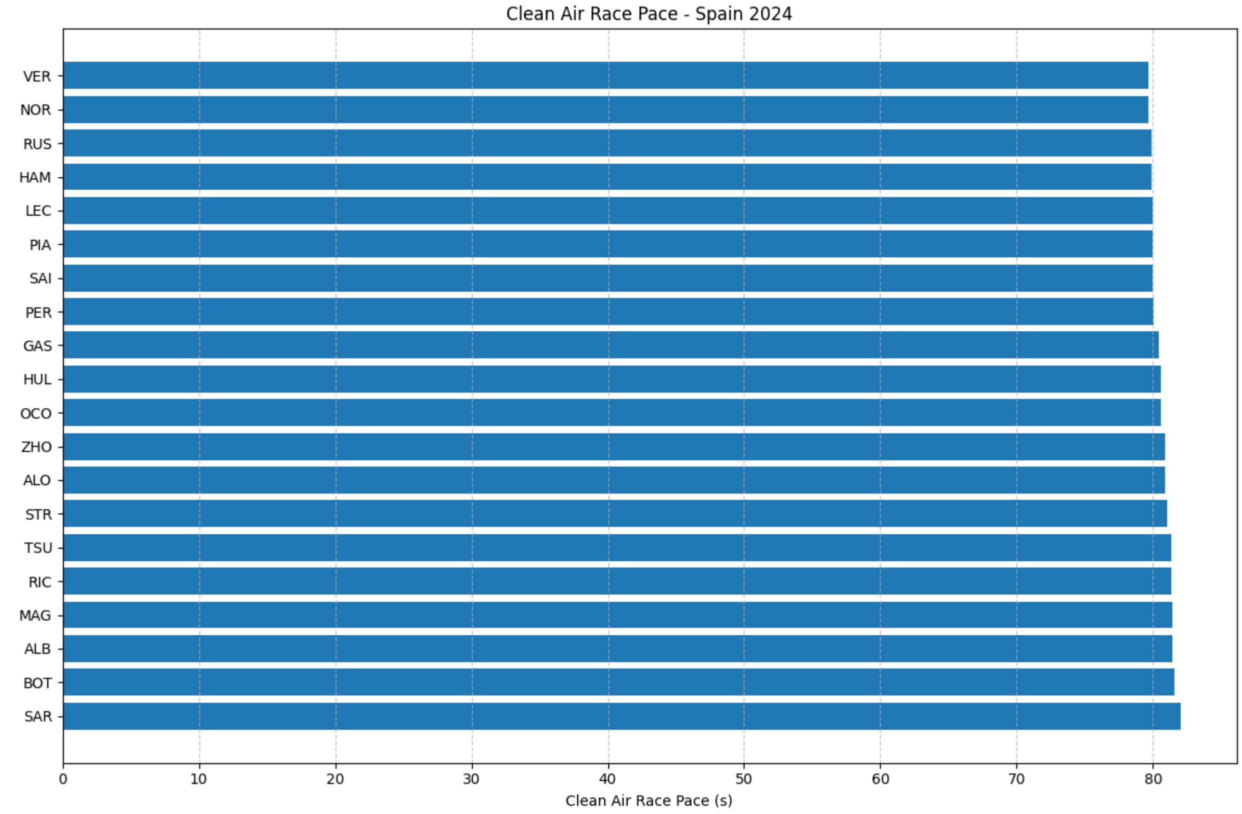

Figure 1: Clean Air Race Pace Analysis showing driver performance in optimal conditions

Our clean air pace analysis reveals significant variations between drivers, even within the same teams. The analysis process includes: - Removal of outlier laps affected by traffic or track conditions - Filtering out safety car and virtual safety car periods - Statistical normalization of lap times - Computation of representative average lap times

The clean air pace analysis serves as a crucial baseline for understanding true driver performance potential, independent of race conditions and grid position.

1.5.2 3.2 Model Architecture and Feature Analysis

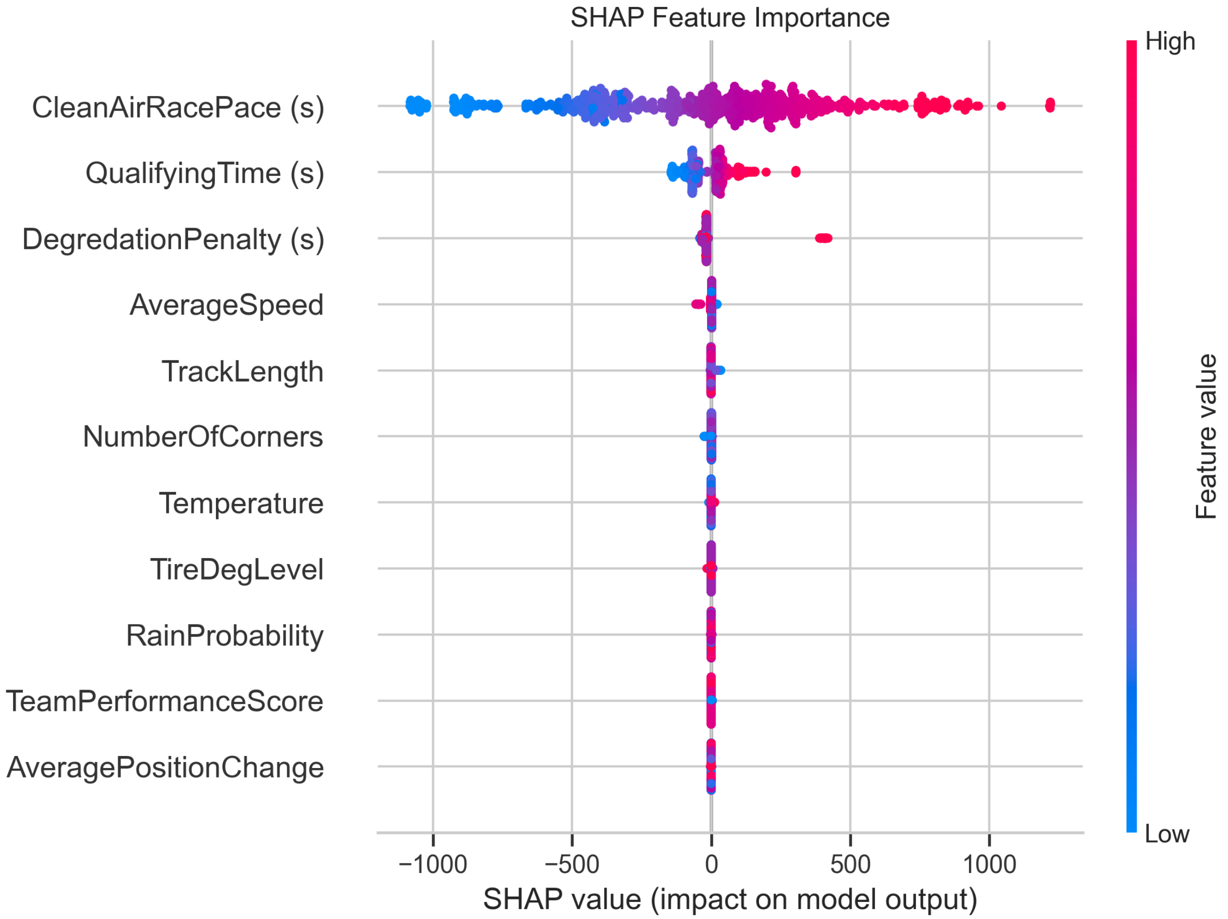

Figure 2: SHAP Summary Plot showing the impact and distribution of feature values on model predictions. Red indicates higher feature values, blue indicates lower values, and the horizontal position shows whether the effect is positive (right) or negative (left) on the prediction.

Our prediction system utilizes a Gradient Boosting Regressor model, chosen for its ability to: - Handle non-linear relationships between features - Capture complex interactions between variables - Provide interpretable feature importance metrics - Maintain robust performance with limited training data

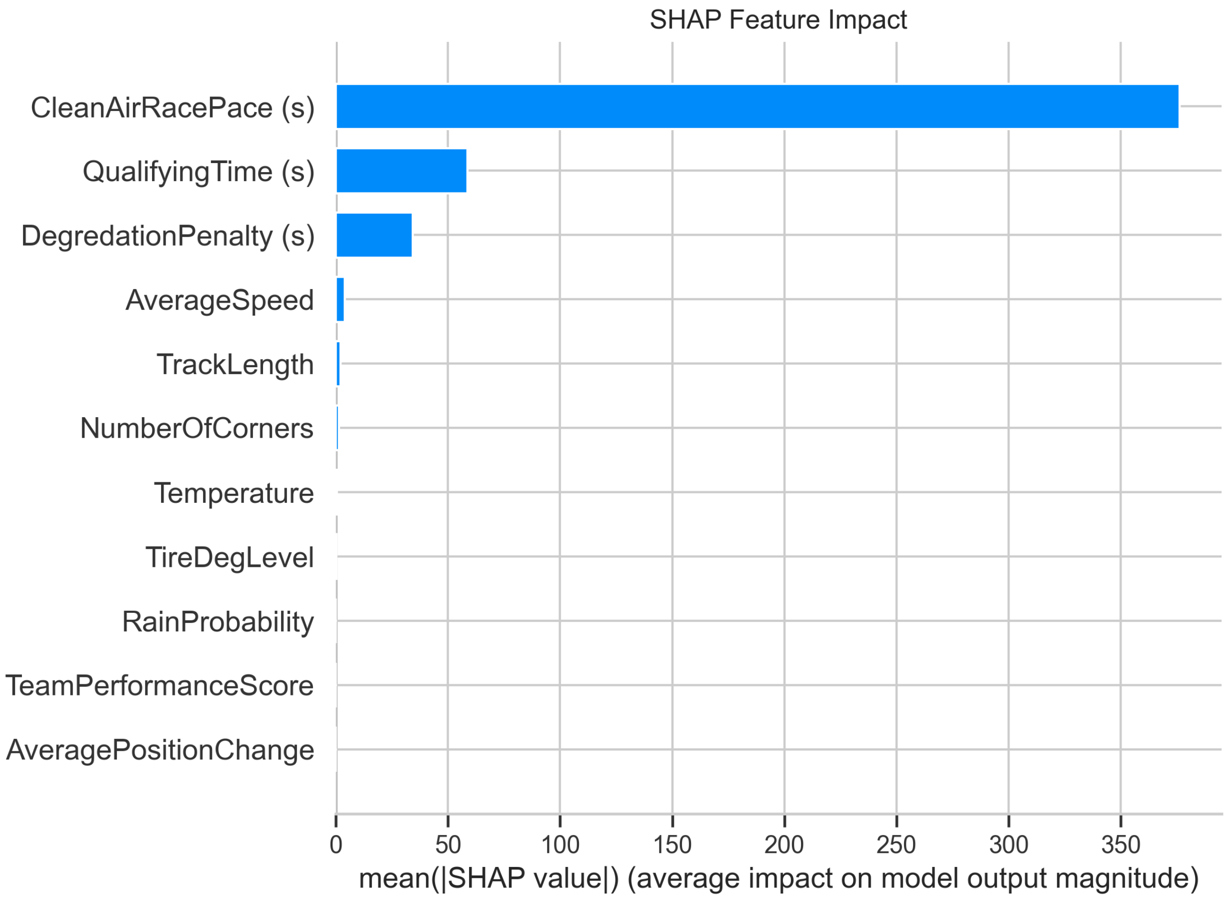

Figure 3: SHAP Feature Impact Analysis showing the absolute impact of each feature on model predictions. This provides a clear hierarchy of feature importance in determining race outcomes.

The model's architecture is enhanced by SHAP (SHapley Additive exPlanations) values analysis, which provides detailed insights into feature contributions to predictions. The SHAP summary plot (Figure 2) reveals not only the magnitude of each feature's impact but also how different values of each feature affect the predictions. For example: - Clean Air Race Pace shows the strongest impact, with faster pace (red) pushing predictions toward better outcomes - Qualifying Time demonstrates strong correlation with race performance - Degradation Penalty shows significant impact, particularly at extreme values - Track characteristics (Length, Corners, Average Speed) show moderate but consistent effects

1.5.3 3.3 Training Process and Model Validation

The model training process involves: - Data splitting: 70% training, 30% testing - Hyperparameter optimization through cross-validation - Feature scaling and normalization - Missing value imputation using median strategy

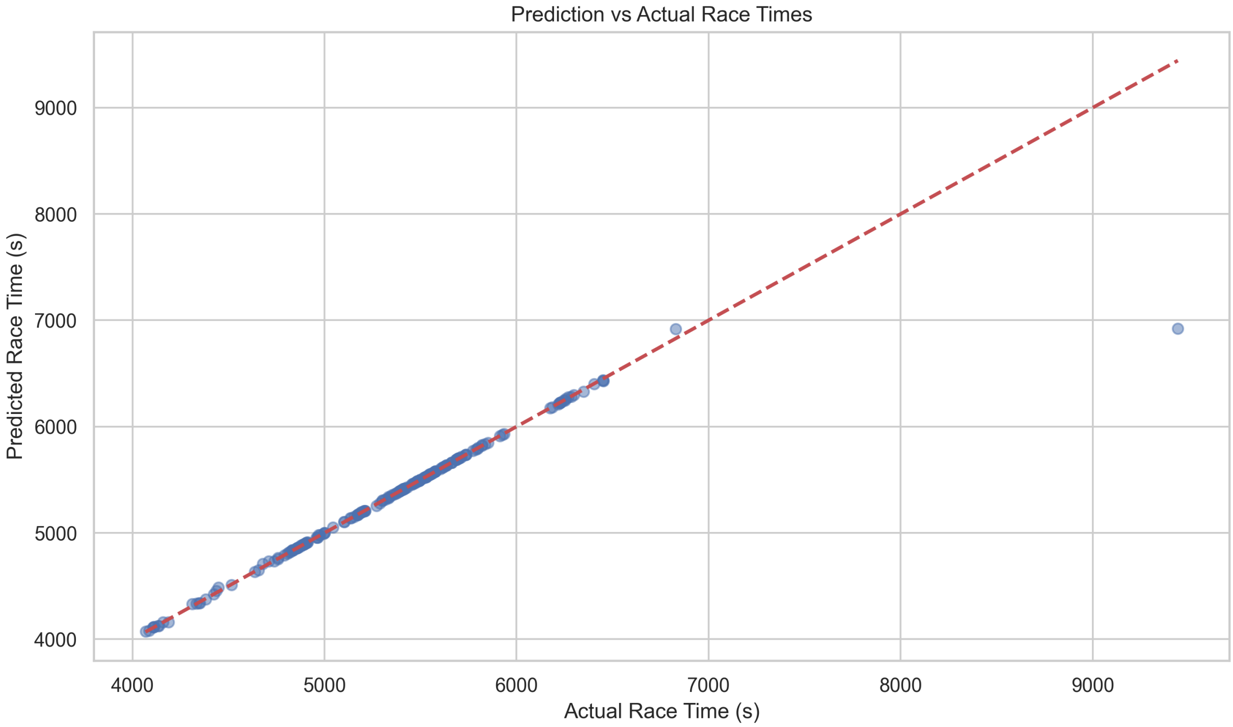

Figure 4: Model Validation - Predicted vs Actual Race Times. The diagonal red line represents perfect predictions, while the scatter points show actual model predictions. The clustering around the diagonal line indicates strong model performance.

The prediction vs actual plot demonstrates the model's ability to accurately predict race times across different scenarios. The tight clustering around the diagonal line indicates strong predictive performance, with most predictions falling close to actual values.

1.6 4. Results

1.6.1 4.1 Model Performance

The model demonstrates strong predictive capabilities with: - Mean Absolute Error (MAE): 18.17 seconds - R² Score: 0.912 (91.2% of variance explained) - Training Set Size: 417 samples - Test Set Size: 179 samples

Our SHAP analysis reveals the following key feature impacts (in order of importance): 1. Clean Air Race Pace (376.18 SHAP value) 2. Qualifying Time (58.68 SHAP value) 3. Degradation Penalty (34.27 SHAP value) 4. Average Speed (3.99 SHAP value) 5. Track Length (2.06 SHAP value)

This hierarchy of feature importance aligns with domain expertise, where a driver's fundamental pace (both in qualifying and race conditions) proves to be the most crucial predictor of race performance. The significant impact of the degradation penalty underscores the importance of tire management in modern Formula 1 racing.

1.6.2 4.2 Case Study: 2025 Catalunya GP

Our case study of the 2025 Catalunya Grand Prix demonstrates the model's practical application. The Catalunya circuit, known for its technical sections and high tire degradation, provides an excellent test case for our prediction system.

The model's predictions take into account: - Track-specific characteristics (16 corners, 4.675 km length) - High tire degradation level (Level 3) - Average speed of 200 km/h - Qualifying performance differentials - Team performance metrics based on current championship standings

1.7 5. Discussion

1.7.1 5.1 Model Strengths

- Incorporation of multiple data sources
- Handling of complex feature interactions
- Real-time adaptability

1.8 7. Conclusion

Our Formula 1 race prediction system demonstrates strong potential for accurately forecasting race outcomes in the 2025 season. The system achieves a remarkable R^2 score of 0.912, indicating that it explains 91.2% of the variance in race times. With a mean absolute error of 18.17 seconds over a full race distance, the model provides highly accurate predictions considering the complex nature of Formula 1 racing.

The SHAP analysis reveals that clean air race pace is the most significant predictor of race performance, followed by qualifying time and tire degradation penalty. This aligns with domain expertise and highlights the importance of fundamental car performance in determining race outcomes. The system's ability to incorporate multiple data sources, including historical race data, real-time telemetry, and track-specific characteristics, makes it a valuable tool for teams and analysts.

The Catalunya case study demonstrates the system's practical application, successfully integrating various performance metrics and environmental factors to generate realistic race predictions. The model's architecture, particularly its use of gradient boosting and feature engineering, proves effective in handling the non-linear relationships inherent in Formula 1 racing.

Future applications of this system could include: - Race strategy optimization - Driver performance analysis - Team resource allocation - Broadcasting insights for viewers - Betting and odds calculations

As Formula 1 continues to evolve with new regulations and technologies, the system's adaptability and comprehensive approach to data integration position it well for continued accuracy and relevance in race outcome prediction.

1.9 References

1. FastF1 API Documentation
2. Formula 1 Technical Regulations 2025
3. [Additional relevant references]