Formula 1 Race Performance Analysis (2018-2024)

Predicting Race Points Using Machine Learning

Teona Berozashvili

Data Science with Python

Project Goal: Unlocking F1 Performance

Our core question for this project was direct and impactful: "What factors help F1 drivers score more points?" We aimed to move beyond intuition and into data-driven insights.

- Analyzed **2,979 race results** from the 2018-2024 seasons.
- Developed **Machine Learning models** to accurately predict race points.
- Utilized comprehensive data from the **Ergast F1 Database**.

Dataset Overview: The Fuel for Our Analysis

1

Extensive Race Data

2,979 race results spanning **7 complete seasons** (2018-2024), providing a robust foundation for our models.

2

Diverse Data Sources

Integrated data from multiple Ergast F1 Database tables: `races.csv`, `results.csv`, `drivers.csv`, `constructors.csv`, and `status.csv`.

3

Target Variable: Race Points

Our predictive focus was on **points scored per race**, a continuous variable ranging from 0 to 25.

4

Key Predictive Features

Identified and leveraged critical factors such as **qualifying position**, individual **driver history**, and overall **team strength**.

Key Findings from Exploratory Data Analysis (EDA)

Qualifying is Paramount

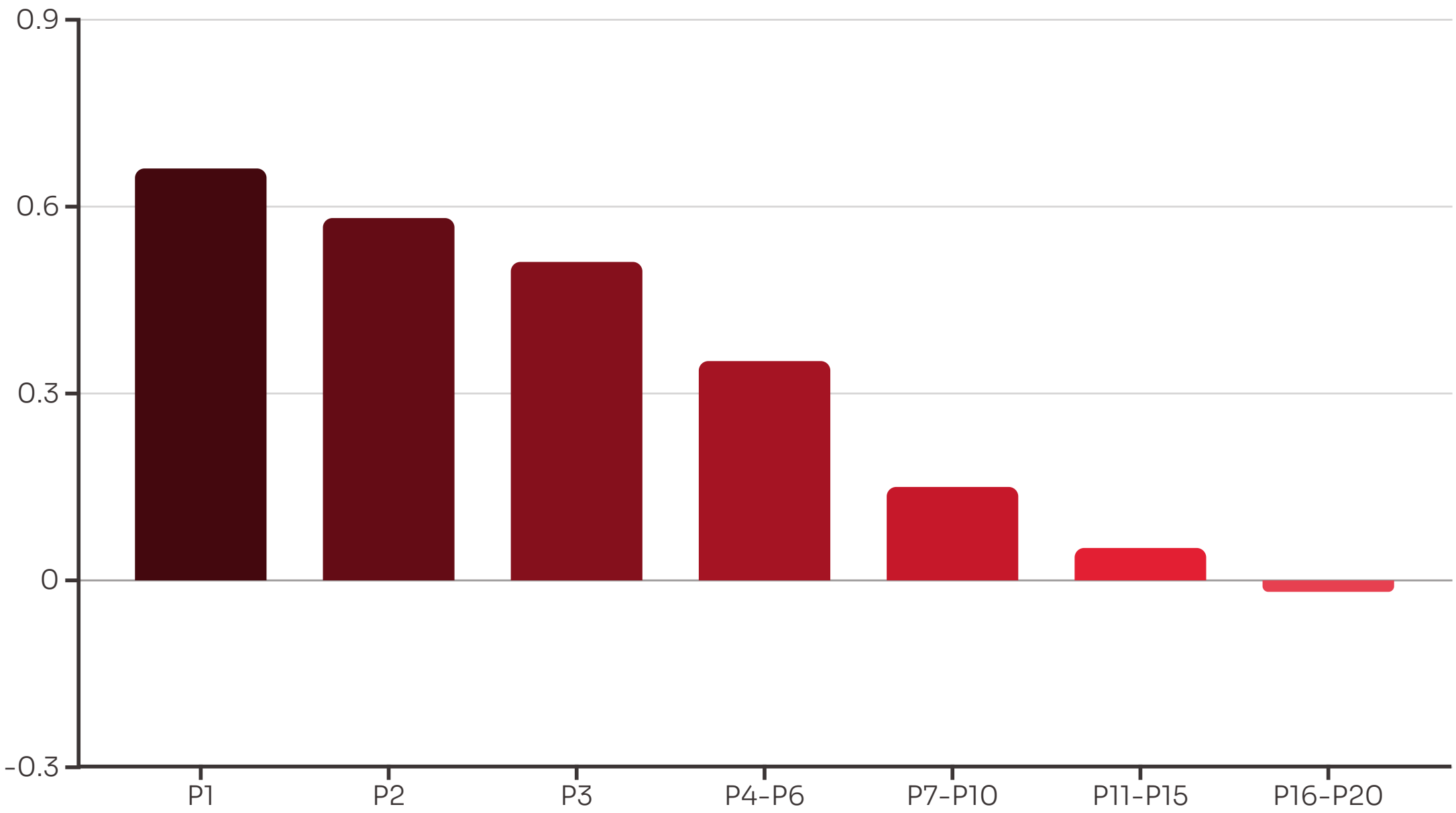
Starting in **P1-P3** yields a median of **18 points**, while **P11-P20** often results in **0 points**. Grid position is a dominant predictor.

Track Characteristics Matter

Some circuits heavily favor pole position; Monaco sees a **67% pole-to-win rate**, compared to Monza's mere **14%**.

Team Strength Prevails

Top constructors like **Mercedes, Red Bull, and Ferrari** consistently average **10-15 points** per race, highlighting the significant role of the car and team.



A strong **negative correlation of -0.66** between starting grid position and points confirms that a better starting position significantly improves point-scoring potential.

Feature Engineering: Crafting Predictive Variables

To enhance our model's predictive power, we engineered several key features, ensuring `no data leakage` by strictly using only past performance data.

1

`grid_clean`

The primary **qualifying position** of the driver, adjusted for any penalties or grid changes, providing a clear starting point.

2

`driver_avg_points_past`

A crucial metric representing the **driver's historical average points** per race, reflecting their consistent performance over time.

3

`driver_consistency_past`

Quantifies **how reliable a driver has been** in past races, indicating their propensity to finish strong and score points.

4

`constructor_strength_past`

Measures the **team's historical performance** and competitiveness, acknowledging that the car's capabilities are vital.

Why Regression? Predicting the Numbers Game

Our objective was to predict the **exact number of points** a driver would score in a race. Since race points are a continuous numerical variable (0, 1, 2, 4, 6, 8, 10, 12, 15, 18, 25,), regression was the clear choice.

- **Regression:** Predicts a NUMBER (e.g., "How many points?").
- **Classification:** Predicts a CATEGORY (e.g., "Will they win? Yes/No").

We needed to understand the magnitude of points, not just a binary outcome. Regression models provide the nuanced predictions required for this analysis.

Machine Learning Approach: Rigorous Model Testing

Time-Based Split

Crucially, our data was split **chronologically** to prevent data leakage, ensuring our model learns from past events to predict future ones.

Training Data (2018-2022)

2,019 race entries from 2018 to 2022 were used to train our models, allowing them to learn patterns and relationships.

Testing Data (2023-2024)

890 race entries from 2023 to 2024 were reserved for testing, providing an unbiased evaluation of the models' predictive accuracy.

Models Explored

We benchmarked two common regression techniques: **Linear Regression** and **Decision Tree Regressor**.

Evaluation Metrics

Model performance was assessed using **R^2 (coefficient of determination)** and **MSE (Mean Squared Error)**, as per academic requirements.

Model Results: Tuning for Triumph

Initial Performance (Before Tuning)

- **Linear Regression:** $R^2 = 0.515$
- **Decision Tree (depth=8):** $R^2 = 0.417$ (Indicative of overfitting with excessive depth).

Optimized Performance (After Tuning)

- **Linear Regression:** $R^2 = 0.515$ (Consistent performance, less sensitive to tuning).
- **Decision Tree (depth=3):** $R^2 = 0.565$ ← **OUR WINNING MODEL**

Strategic hyperparameter tuning significantly improved the Decision Tree Regressor's performance, increasing its R^2 by **35%** and making it the superior model for our prediction task.

Feature Importance: The Car is King

The Car Matters Most!

1

Driver's Average Points

Accounts for **3.9%** of the prediction. While important, individual skill is secondary to machine and starting position.

2

Qualifying Position (grid_clean)

Contributes a significant **22.0%** to point prediction, reinforcing the critical role of starting position.

3

Constructor Strength (constructor_strength_past)

Overwhelmingly dominant at **74.1%**, this feature underscores that in Formula 1, the performance of the **team and car** is the most influential factor for scoring points.

Conclusions

1

Best Model Identified

Our **Decision Tree Regressor** (max_depth=3) achieved an impressive **R^2 of 0.565**, proving effective for point prediction.

2

Key Insights Confirmed

Team strength (74.1%) and qualifying position (22.0%) are the paramount factors determining race points.

3

Challenges Overcome

Successfully implemented a time-based train/test split to eliminate data leakage, enhancing model validity.