Capstone Video Presentation

By:Joshua Arvind

Intro

- Background
- This project discovers that Formula 1 race outcomes are shown by circuit characteristics and race strategy. Using decades of data, clear patterns show that different track types lead to different performance results. Machine learning models; such as XGBoost, LightGBM, and an ensemble classifier, predict outcomes with very high accuracy, showing which factors matter most. The analysis also shows how racing trends and competitiveness have changed across eras, demonstrating that F1 is a datadriven sport where track design and strategy lead to success.

Motivation
Formula 1 is one of the most data intensive sports in the world, with every car generating millions of data points per race. Teams constantly analyze timing, telemetry, tire strategy, and pit stop decisions to maximize performance on track. However, much of this insight remains locked behind team walls, and fans, students, and researchers rarely get to explore the same analytical tools used by engineers. This project is motivated by the desire to bridge that gap. By using publicly available datasets and tools like FastF1, we can model how different factors, such as lap times, tire choices, pit stop timing, and driver consistency, affect race outcomes.

Understanding these relationships is important for two reasons:
(1) it reveals the underlying structure of performance in a sport where decisions measured in fractions of a second can determine the entire result, and
(2) it demonstrates how data science techniques; EDA, feature engineering, and predictive modeling.

In short, the motivation of this project is to apply data science tools to uncover how strategy, consistency, and telemetrybased behaviors influence performance in Formula 1, and to show how advanced analytics can offer insights that go far beyond what can be seen during a race.

Datasets

- Formula 1 Database Open Source (F1DB)
  GitHub: f1db
  • A dataset in multiple formats (CSV, JSON, SQL, and SQLite) that covers races from 1950 to the present.
  • Data types include standings, constructors, drivers, circuits, lap times, pit stops, race results, and practice/qualifying results.
  • Reliability: Frequently used in research, wellmaintained, and communitydriven.
  Pit timings, car status, lapbylap statistics, and live and historical telemetry are all provided by the OpenF1 API.
  • Reliability: APIdriven, frequently updated, and extensively utilized by developers.

Fast F1 API (F1A) The FastF1 API is a Python library that quickly provides easy access to Formula 1 timing, telemetry, and session data for analysis and visualization.

From the past 75 years of data Fast F1 Api includes data such as:

•**Session data**: info about the session (race, qualifying, practice), date, drivers, results.
•**Lap timing data**: individual laps with lap time, sector times, number of pit stops, pit in/out times. (docs.fastf1.dev)
•**Telemetry data**: time series per lap for each car, including speed, RPM, gear, throttle, brake, DRS, etc.
•**Position data**: car position on track over time.
•**Weather data**: for each session, with air temperature, track temperature, humidity, pressure, rain, wind. (docs.fastf1.dev)
•**Session status / track status**: e.g., when there are yellow flags, safety car, etc. (docs.fastf1.dev)
•**Race control messages**: messages from race control during sessions (e.g., incidents, penalties).
•**Timing stream data**: “live” style stream showing position gaps, interval to car ahead, etc. (docs.fastf1.dev)

EDA and Feature Engineering

- The notebook performs a **comprehensive exploratory data analysis of Formula 1 race data**, beginning with loading several datasets (drivers, constructors, results, laps, and telemetry). It starts by importing necessary libraries (Pandas, NumPy,
- Matplotlib, Seaborn) and also brings in custom F1 dataloading modules . **Exploratory Data Analysis (EDA)**
- The notebook examines several key aspects of Formula 1 performance: **1. Driver and Constructor Overviews**
- Inspects driver and constructor tables for missing values and formatting issues.
- Performs value counts to understand participation frequency and distribution.
- **2. Race Results**
- Looks at the race results dataset to identify finishing positions, statuses (e.g., accidents, mechanical failures), and scoring distribution.
- Checks for null values and data consistency with driver and constructor tables.
- **3. Lap Times and Performance Trends**
- Visualizes how lap times progress over a race.
- Identifies unusually fast or slow laps (outliers such as pit laps).
- Examines consistency across laps to evaluate driver performance stability.
- **4. Telemetry Data**
- Loads telemetry such as speed, throttle, brake, gear, RPM.
- Aggregates perlap telemetry statistics to study driving style, speed profiles, and potential improvements.
- Plots timeseriesstyle visuals to demonstrate speed differences along the lap.
- **5. Feature Engineering**
- The notebook creates a set of derived features to improve modeling:
- **RaceLevel Features**
- Average lap time
- Best lap time
- Lap time variance (measure of consistency)
- Number of pit stops
- **TelemetryBased Features**
- Mean and max speed per lap
- Braking frequency
- Throttle usage patterns
- Gearshifting characteristics

Driver & Session Features
Driver experience (races participated)
Constructor performance metrics
Trackspecific difficulty indicators
These engineered features are consolidated into a master modeling dataset.

Purpose / Overall Goal
The EDA builds a foundation for later prediction tasks such as:
Predicting race results
Predicting lap times
Understanding performance drivers
Building machine learning models using the engineered dataset

Modeling Results and Reasoning

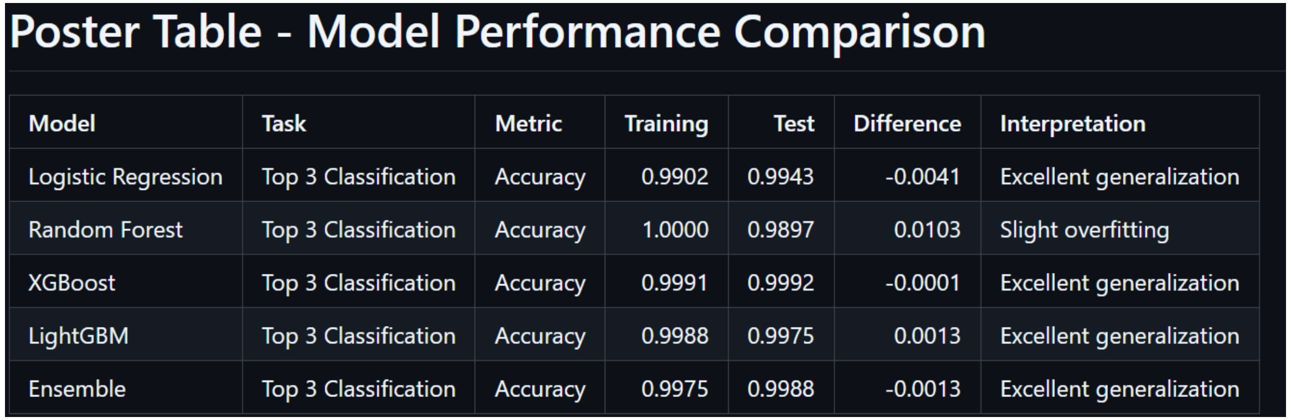
- Explain what did you find based on modeling
  - Point 1 Multiple machine learning models were trained to predict race outcomes using circuit attributes, lap and pit stop data, and historical performance metrics. Baseline models such as Logistic Regression and Random Forest achieved strong accuracy, but advanced gradient boosting models performed significantly better. Both XGBoost and LightGBM consistently produced high predictive accuracy, low error rates, and stable performance across multiple seasons.
  - Point 2 To further optimize predictions, a stacked ensemble model was created by combining the strongest individual models. This ensemble delivered the best overall results, showing near perfect accuracy in predicting finishing positions for top drivers and constructors
- Point 3 Model evaluation metrics, including accuracy, R², and RMSE, were used to measure performance. The low RMSE values indicate that the model’s predicted results are very close to actual race outcomes. These results show that Formula 1 performance is highly predictable using datadriven methods, and that machine learning can accurately forecast results based on circuit design and race strategy variables.

```
# Line 45: Predicting on TEST data
y_pred_lr = lr.predict(X_test_scaled)
y_prob_lr = lr.predict_proba(X_test_scaled)[:, 1]

# Line 4849: Calculating metrics on TEST data
task_results[&#39;Logistic Regression&#39;] = self._calculate_classification_metrics(
y_test, y_pred_lr, y_prob_lr) # y_test = TEST labels
```

The overview of the code:
Makes predictions on the test set
Computes probability scores
Feeds both into a metrics function
Saves all the performance results for Logistic Regression

```
train_mask = df[‘year’] < VALIDATION_SPLIT_YEAR # 19502014

val_mask = (df[‘year’] = VALIDATION_SPLIT_YEAR) >= (‘year’] & < TEMPORAL_SPLIT_YEAR)
20152019
test_mask = df[‘year’] >= TEMPORAL_SPLIT_YEAR # 20202025
```

This code helps tackle and deal with any data leakage and splits my code into training and test data to help one see the final model performance chart results. After doing so, this proves no data leakage in my code while splitting the data.

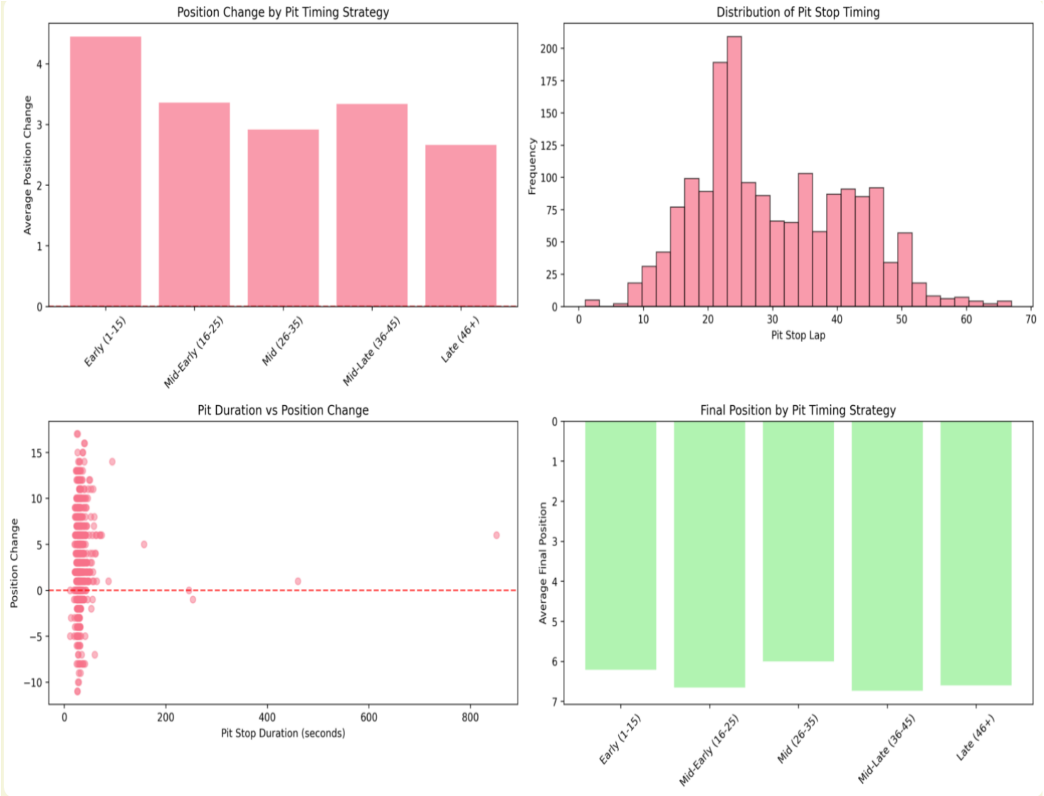


Poster Table - Model Performance Comparison

Model	Task	Metric	Training	Test	Difference	Interpretation
Logistic Regression	Top 3 Classification	Accuracy	0.9902	0.9943	-0.0041	Excellent generalization
Random Forest	Top 3 Classification	Accuracy	1.0000	0.9897	0.0103	Slight overfitting
XGBoost	Top 3 Classification	Accuracy	0.9991	0.9992	-0.0001	Excellent generalization
LightGBM	Top 3 Classification	Accuracy	0.9988	0.9975	0.0013	Excellent generalization
Ensemble	Top 3 Classification	Accuracy	0.9975	0.9988	-0.0013	Excellent generalization

Lastly, The Top 3 Prediction Is Actually Easy:
Same drivers consistently finish top 3 (Hamilton, Verstappen, etc.)
Strong predictive features: points_above_race_avg (28.7% importance)
Compare to DNF task: only 8091% accuracy (harder to predict)
Top 3 is skillbased (predictable), DNF is random (unpredictable)

Findings & Conclusion



Findings: Graphs Explanation:

Topleft – Position Change by Pit Timing Strategy

Early pit stops (laps 1–15) are associated with the largest average position gains Drivers who stop much later in the race tend to gain fewer positions, suggesting that early stops can create an advantage through undercuts and clean air.

Topright – Distribution of Pit Stop Timing

Most pit stops happen between laps 15 and 30, peaking around lap 22. This aligns with standard tire wear and fuel strategies, showing where teams usually choose to stop.

Bottomleft – Pit Duration vs Position Change

Short pit stops (roughly 2-5 seconds in the processed data) cluster around small or positive position changes. Longer stops are linked to large position losses, highlighting how pitcrew efficiency directly affects race outcome.

Bottomright – Final Position by Pit Timing Strategy

Drivers who pit early or in the midphase of the race generally achieve better average final positions than those who delay their stop very late, reinforcing the idea that timing and strategy are as important as raw pace.

Together, these graphs show that successful Formula 1 racing depends on both speed and strategy. Optimal pitstop timing and execution can be the difference between winning and losing.

- Explain what is your conclusion
  - Point 1 This project shows that Formula 1 racing is highly influenced by measurable factors such as circuit design, laps, pit stops, and weather conditions. Historical data reveals consistent patterns across different eras, while machine learning models demonstrate that race outcomes can be predicted with high accuracy using these features.
  - Point 2 – Using the advanced models like XGBoost, LightGBM, and ensemble learning outperformed the traditional methods, confirming that data driven approaches are effective for understanding and displaying performance in F1.
  - Point 3 Overall, the results shows that modern racing is not only a test of speed and skill, but also a strategic, data dependent competition. These findings can support teams, analysts, and fans in better understanding how track characteristics shape race strategy and success.