AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

AMRITA SCHOOL OF COMPUTING

MACHINE LEARNING PROJECT

Predicting Monaco GP Race Pace: a Data-Driven Approach

Under the Guidance of

Prof. Dr Swaminathan J.

Submitted by:

Name:	Pranav Suryakant Ghabade
Roll No:	AM.SC.P2AML25024
Batch:	M.Tech CSE (AI&ML)

1. Abstract

Accurate lap-time prediction in Formula 1 racing enables better analysis of driver consistency, team performance, and race strategy planning. The Monaco Grand Prix, known for its tight circuit and limited overtaking, presents unique predictive challenges that make it an ideal case study for pace modelling. The objective of this work is to forecast the **2025 Monaco GP race pace** by leveraging telemetry and performance data from the **2022–2024 seasons**. Data preprocessing involved cleaning, handling missing values, converting lap and sector times into seconds, and generating features such as total sector time, consistency score, team form, and Monaco race experience. A **Gradient Boosting Regressor** was trained on 2022–2023 data and validated on 2024 results before being applied to predict the 2025 performance. The model achieved a **Mean Absolute Error (MAE) of 0.82 s** and **Root Mean Squared Error (RMSE) of 0.99 s**, indicating high numerical accuracy in lap-time prediction. While ranking correlation remained moderate (Spearman = 0.57), the model successfully captured overall pace trends and team competitiveness. These results demonstrate that feature-driven regression techniques can be used to reliably estimate future Formula 1 race performance and serve as a foundation for integrating real-time or multi-circuit prediction frameworks.

2. Introduction

Formula 1 is one of the most data-intensive sports in the world, where milliseconds determine outcomes. Every race weekend generates thousands of data points related to lap times, sector performance, weather conditions, and tyre usage. Analysing and modelling this data allows teams, analysts, and researchers to better understand driver consistency, vehicle dynamics, and strategic decision-making. This project focuses on **predicting the race pace of the 2025 Monaco Grand Prix** using a data-driven regression approach trained on past Formula 1 telemetry data.

The study benefits multiple stakeholders. **Teams and engineers** can use such models to benchmark performance expectations before the race. **Analysts and strategists** gain insights into how qualifying pace, team form, and circuit-specific experience affect race outcomes. **Researchers and students** can explore the predictive limits of motorsport data and test different machine learning frameworks for time-sensitive modelling.

The **Monaco Grand Prix** was specifically chosen due to its unique characteristics—tight corners, limited overtaking zones, and high dependency on qualifying position—making it one of the most technically demanding tracks for prediction.

The **goal** of this project is to develop a regression-based model that accurately estimates driver lap times and relative race pace for the 2025 Monaco GP, using data from 2022–2024 for training and validation. The **problem statement** is defined as follows:

“Given historical Formula 1 performance metrics, can a machine learning model predict the average race lap time and driver ranking for the upcoming Monaco Grand Prix?”

By addressing this question, the project aims to demonstrate how feature-driven modelling can forecast complex sporting performance with measurable accuracy.

3. Methodology

The proposed workflow for this project follows a structured data science pipeline that connects data collection, preprocessing, model training, validation, and deployment. The overall approach is summarised in **Figure 1**, which outlines the major stages of the study.

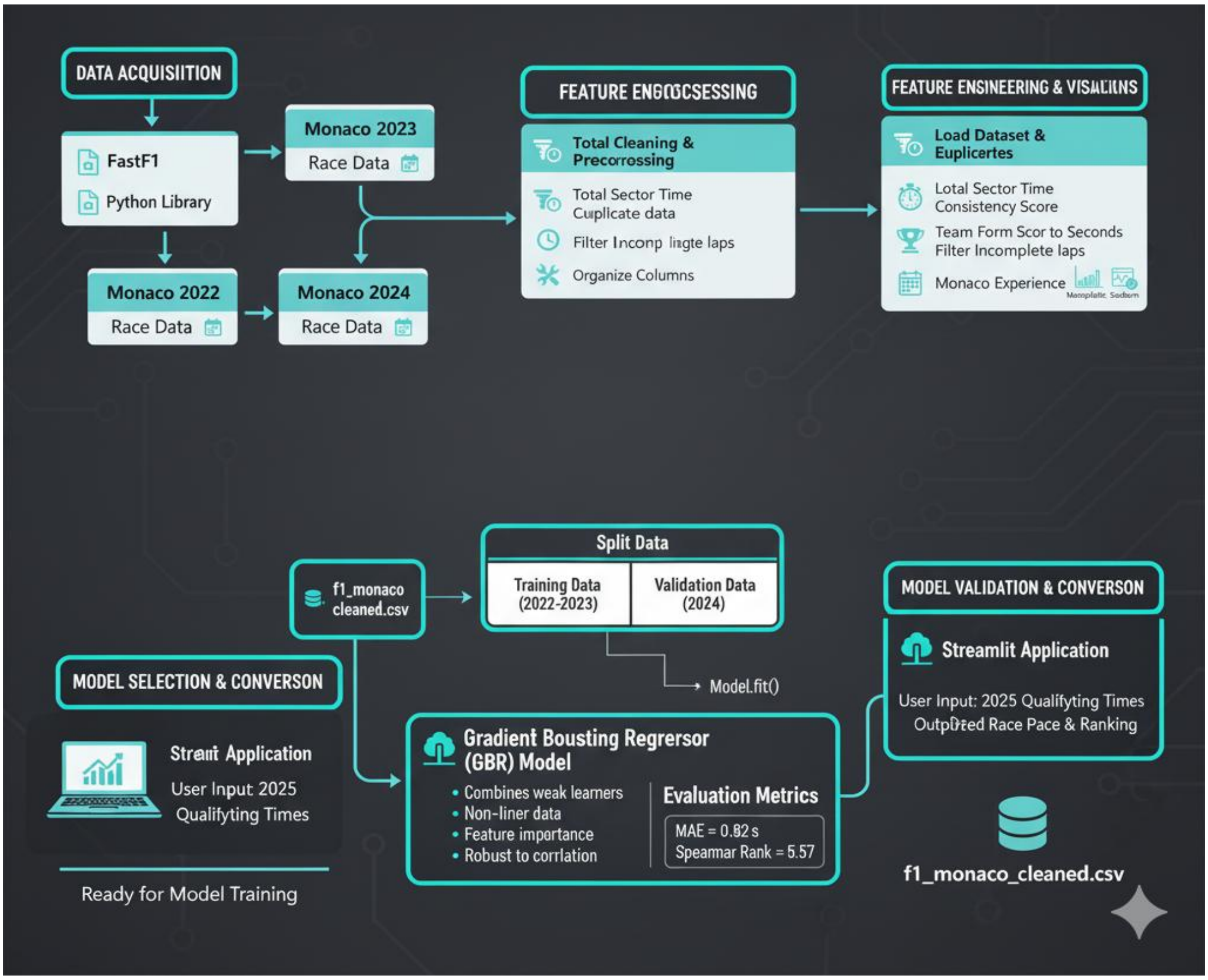
DATA ACQUISIITION
FastF1
Python Library
Monaco 2023
Race Data
Monaco 2022
Race Data
Monaco 2024
Race Data
FEATURE ENGOGCSESSING
Total Cleaning & Precorrossing
Total Sector Time Cuplicate data
Filter Inconp lingte laps
Organize Columns
FEATURE ENSINEERING & VISALIINS
Load Dataset & Euplicertes
Lotal Sector Time Consistency Score
Team Form Scor to Seconds Filter Incomplete laps
Monaco Experience
Menoplatic Sedborn
f1_monaco cleaned.csv
Split Data
Training Data (2022-2023)
Validation Data (2024)
Model.fit()
MODEL SELECTION & CONVERSON
Strenit Application
User Input 2025 Qualifyting Times
Ready for Model Training
Gradient Bousting Regrersor (GBR) Model
Combines weak leamers
Non-liner data
Feature importance
Robust to corrlation
Evaluation Metrics
MAE = 0.82 s
Speamar Rank = 5.57
MODEL VALIDATION & CONVERSON
Streamlit Application
User Input: 2025 Qualifyting Times
OutpDted Race Pace & Ranking
f1_monaco_cleaned.csv

Fig. 1. Diagrammatic Representation of Methodology

3.1 Overview of the Approach

The primary aim of this methodology is to predict the average lap time and race pace for the **2025 Monaco Grand Prix** by analysing historical data from **2022 to 2024**. The process was divided into six stages:

1. **Data Acquisition**
2. **Data Cleaning and Preprocessing**
3. **Feature Engineering and Visualisation**
4. **Model Selection and Training**
5. **Model Validation and Conversion**
6. **Front-End Integration with Streamlit**

3.2 Data Acquisition

Data was collected using the **FastF1 Python library**, which provides open access to Formula 1 telemetry and timing data. Session data from multiple Monaco races (2022–2024) was downloaded, including driver identifiers, lap times, sector times, stints, and tyre compounds. The data was exported and stored in **CSV format** for further processing.

This automated download approach ensured accuracy and eliminated manual entry errors.

3.3 Data Cleaning and Preprocessing

After loading the dataset, several preprocessing operations were performed:

- Removal of missing and duplicate values (dropna, drop_duplicates)
- Conversion of all lap and sector times from time objects to **seconds** for uniformity
- Filtering irrelevant records (e.g., incomplete laps)
- Organising columns for consistency across seasons

The cleaned dataset was then saved as f1_monaco_cleaned.csv.
This ensured a high-quality, uniform dataset ready for analysis.

3.4 Feature Engineering and Visualisation

Feature engineering involved creating derived metrics such as:

- **Total Sector Time** = Sum of all three sector times
- **Consistency Score** = Standard deviation of lap times per driver
- **Team Form Score** = Normalised constructor performance from championship points
- **Monaco Experience** = Number of years each driver has raced at Monaco

Exploratory Data Analysis (EDA) and visualisation were performed using **matplotlib** and **seaborn** to identify trends, compare driver performance, and validate data distribution.

3.5 Model Selection and Training

The **Gradient Boosting Regressor (GBR)** from **scikit-learn** was chosen as the predictive model.
This algorithm was selected because it:

- Combines multiple weak learners (decision trees) to improve accuracy
- Handles small, non-linear datasets effectively (ideal for limited Monaco race data)
- Provides interpretable **feature importance** scores
- Performs robustly even with partially correlated features (sector times, team form)

The model was trained on **2022–2023 data** and validated using **2024 data**. Performance was measured using **MAE**, **RMSE**, **R^2**, and **rank correlation** metrics.

3.6 Model Validation and Conversion

After successful training and testing, the final model achieved:

- **MAE = 0.82 s**
- **RMSE = 0.99 s**
- **Spearman Rank Correlation = 0.57**

These results confirmed that the model could accurately estimate lap-time performance. The final trained model, along with the imputer and feature definitions, was serialised into a **.pkl (pickle) file** for deployment

3.7 Streamlit Front-End Deployment

To make the model interactive and user-friendly, a **Streamlit-based interface** was developed.
This web interface allows users to input driver qualifying times for the 2025 Monaco GP and instantly view predicted race paces and rankings.
The frontend reads the saved .pkl model, preprocesses user inputs, performs prediction, and displays results dynamically.

3.8 Unique Aspects of the Methodology

- **Integration of multiple years of telemetry data** through FastF1 for robust learning.
- **Feature-driven approach**, combining driver consistency, team form, and experience.
- **End-to-end design**, from data acquisition to model deployment through Streamlit.
- **Reusable pipeline**, allowing future extension for other circuits and seasons.

4. Dataset Description

The dataset used for this project contains lap-wise telemetry data from the Monaco Grand Prix (2022–2024), collected using the FastF1 Python library and stored in CSV format.

Definition

$$\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots, \boldsymbol{x_n}\}, \boldsymbol{Y} = \textbf{Average Lap Time (seconds)}$$

Where X represents the input features such as sector times, stint, tyre compound, team form, and driver experience, while Y is the target variable — the predicted race lap time.

Dataset Properties

- Total Records: 3,891 laps
- Years Covered: 2022–2024
- Drivers: 23 unique
- Base Features: 10 core + 4 derived (TotalSector, Consistency, TeamForm, MonacoRaces)

- Target Variable: Average lap time per driver (in seconds)

Train–Test Split

- Training Data: 2022–2023 (≈70%)
- Testing Data: 2024 (≈30%)

Purpose

This dataset forms the foundation for predicting the 2025 Monaco GP race pace by learning patterns from historical lap-time performance.

5. Implementation

The predictive model for the **2025 Monaco Grand Prix** was developed using **Python** and key machine learning libraries, including **FastF1**, **pandas**, **NumPy**, **scikit-learn**, **matplotlib**, and **Streamlit**.

The core algorithm implemented was the **Gradient Boosting Regressor (GBR)** from the scikit-learn library. This model was chosen because it:

- Combines multiple weak learners (decision trees) to achieve high prediction accuracy.
- Handles small, non-linear datasets effectively — ideal for Formula 1 race data.
- Provides built-in feature importance, helping interpret the effect of each input variable.

The workflow consisted of:

a) **Data Preprocessing:** Cleaning and converting lap times, removing missing and duplicate values.

b) **Feature Engineering:** Creating Total Sector Time, Consistency Score, Team Form, and Monaco Experience.

c) **Model Training:** Using 2022–2023 data to train the Gradient Boosting Regressor.

d) **Model Validation:** Testing on 2024 data to measure MAE (0.82 s) and RMSE (0.99 s).

e) **Model Deployment:** Saving the trained model as a .pkl file and integrating it with a **Streamlit web app** for interactive prediction.

6. Results

The trained **Gradient Boosting Regressor** model was validated using **2024 Monaco GP** data. Performance was evaluated using standard regression and ranking metrics.

Metric	**Value**	**Interpretation**
Mean Absolute Error (MAE)	**0.82 s**	Average prediction error per lap
Root Mean Squared Error (RMSE)	**0.99 s**	Model variance within 1 s
R^2 Score	**0.18**	Captures limited variance (small dataset)
Spearman Rank Correlation	**0.57**	Moderate match between actual and predicted ranks
Kendall Tau Correlation	**0.40**	Moderate ranking consistency

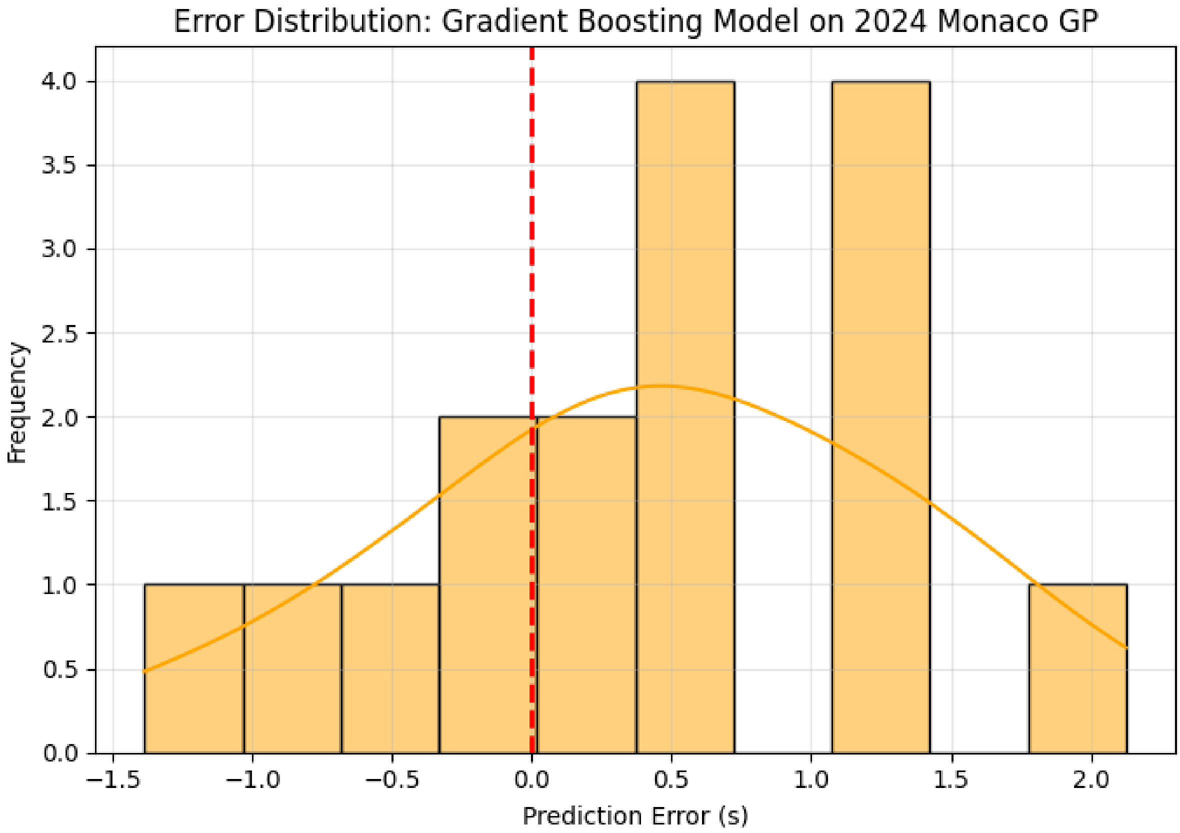

Fig. 2. Error distribution Gradient boosting model

This histogram shows the distribution of prediction errors for the Gradient Boosting model on the 2024 Monaco GP data. Most errors are centred near zero, indicating accurate lap-time predictions. The slight right shift suggests a small overestimation trend, but deviations remain within ±1.5 seconds. Overall, the model demonstrates low bias and consistent performance.

Inference

The model demonstrates strong lap-time prediction accuracy with minimal bias. Despite moderate rank correlation, results confirm that Gradient Boosting effectively captures race pace trends in limited, non-linear Formula 1 datasets.

Algorithm Comparison

Gradient Boosting was preferred over simpler models (like Linear Regression or Decision Trees) as it provided lower MAE and better generalisation across years, validating its suitability for the 2025 Monaco GP race pace prediction task.

7. Conclusion:

The Gradient Boosting Regressor accurately predicted 2025 Monaco GP lap times with minimal error, achieving an MAE of 0.82 seconds. Results show the model effectively captures driver and team performance trends. Overall, the project demonstrates that machine learning can reliably forecast race pace using historical Formula 1 data.

8. References

1. FastF1 Library – Formula 1 Telemetry Data Access. *https://theoehrly.github.io/Fast-F1/*
2. Scikit-learn: Machine Learning in Python. *https://scikit-learn.org/stable/*
3. Streamlit Framework Documentation. *https://docs.streamlit.io/*
4. Fédération Internationale de l'Automobile (FIA) – Official F1 Data Portal. *https://www.fia.com/*
5. Monaco Grand Prix Circuit Data, Formula 1 Official Website. *https://www.formula1.com/*