# F1 Lap Time Prediction
# Regression vs. Time Series Models

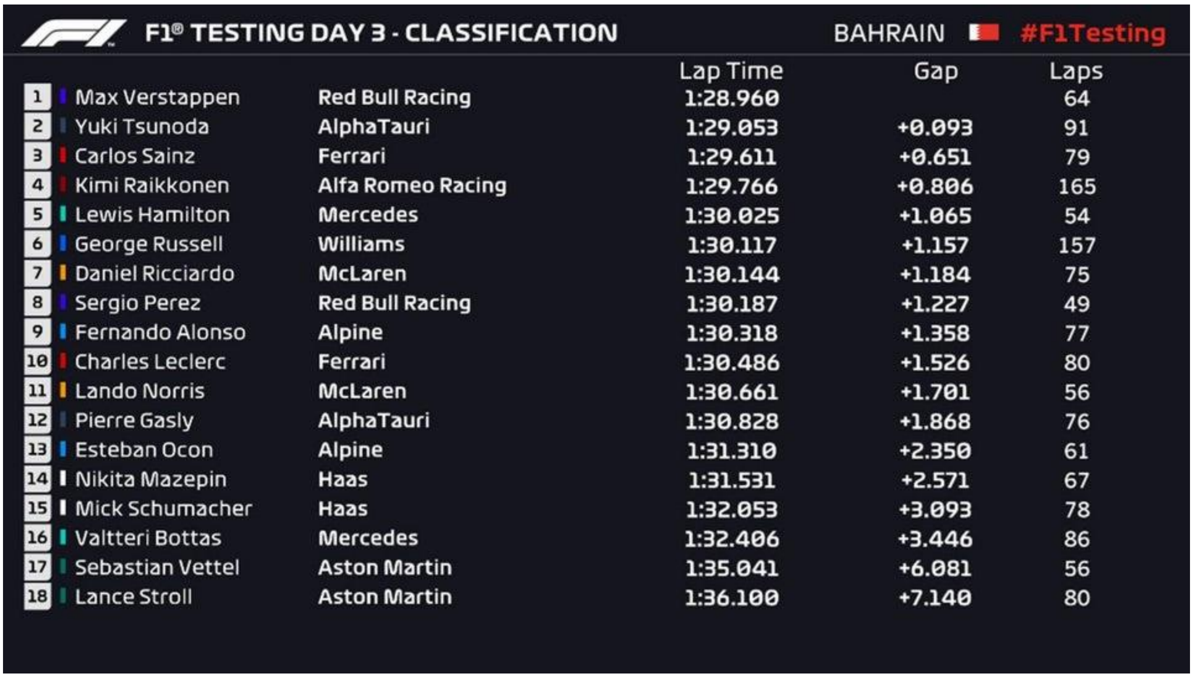
Henri **Eloy**
Data scientist

# F1 Lap Time Prediction

**Formula 1** is a highly competitive motorsport that involves complex strategies, engineering excellence, and cutting-edge technology.

Teams are always looking for ways to gain an advantage over their rivals, and one critical aspect of this is predicting lap times. Accurate lap time predictions can help teams devise better race strategies and make informed decisions during races.

In this report, we investigate the potential of various machine learning models to predict lap times in Formula 1 races. We start by collecting and preprocessing data, followed by feature engineering, exploratory data analysis, and modeling. We compare regression models such as RandomForestRegressor with time series models like Prophet and ARIMA, and finally evaluate the performance of these models to identify the best approach for predicting lap times in Formula 1.



F1® TESTING DAY 3 - CLASSIFICATION

BAHRAIN #F1Testing

| | | | Lap Time | Gap | Laps |
|---|---|---|---|---|---|
| 1 | Max Verstappen | Red Bull Racing | 1:28.960 | | 64 |
| 2 | Yuki Tsunoda | AlphaTauri | 1:29.053 | +0.093 | 91 |
| 3 | Carlos Sainz | Ferrari | 1:29.611 | +0.651 | 79 |
| 4 | Kimi Raikkonen | Alfa Romeo Racing | 1:29.766 | +0.806 | 165 |
| 5 | Lewis Hamilton | Mercedes | 1:30.025 | +1.065 | 54 |
| 6 | George Russell | Williams | 1:30.117 | +1.157 | 157 |
| 7 | Daniel Ricciardo | McLaren | 1:30.144 | +1.184 | 75 |
| 8 | Sergio Perez | Red Bull Racing | 1:30.187 | +1.227 | 49 |
| 9 | Fernando Alonso | Alpine | 1:30.318 | +1.358 | 77 |
| 10 | Charles Leclerc | Ferrari | 1:30.486 | +1.526 | 80 |
| 11 | Lando Norris | McLaren | 1:30.661 | +1.701 | 56 |
| 12 | Pierre Gasly | AlphaTauri | 1:30.828 | +1.868 | 76 |
| 13 | Esteban Ocon | Alpine | 1:31.310 | +2.350 | 61 |
| 14 | Nikita Mazepin | Haas | 1:31.531 | +2.571 | 67 |
| 15 | Mick Schumacher | Haas | 1:32.053 | +3.093 | 78 |
| 16 | Valtteri Bottas | Mercedes | 1:32.406 | +3.446 | 86 |
| 17 | Sebastian Vettel | Aston Martin | 1:35.041 | +6.081 | 56 |
| 18 | Lance Stroll | Aston Martin | 1:36.100 | +7.140 | 80 |

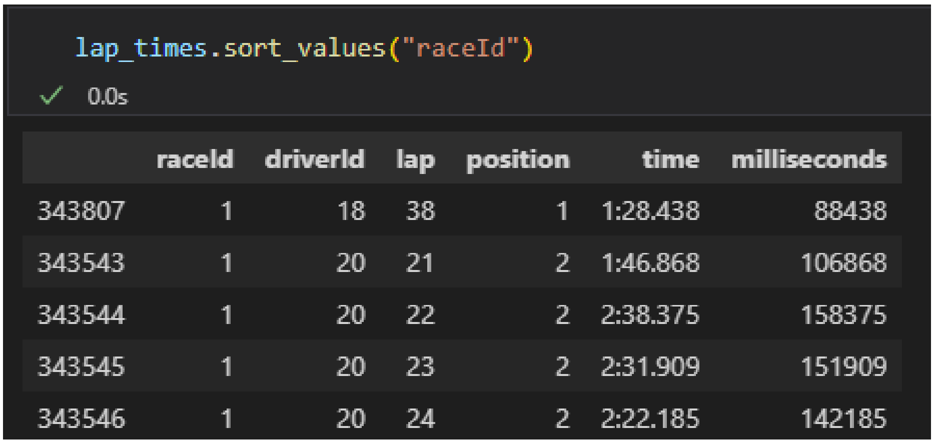
# I- Data Collection

The data used for this analysis is gathered from the "Formula 1 World Championship (1950 - 2023)" dataset available on Kaggle.

https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

The dataset contains information about lap times, circuit details, pit stops, and drivers, among other aspects of Formula 1 races. This dataset offers a comprehensive overview of Formula 1 racing data from 1950 to 2023, enabling us to analyze historical trends and make predictions about future races.

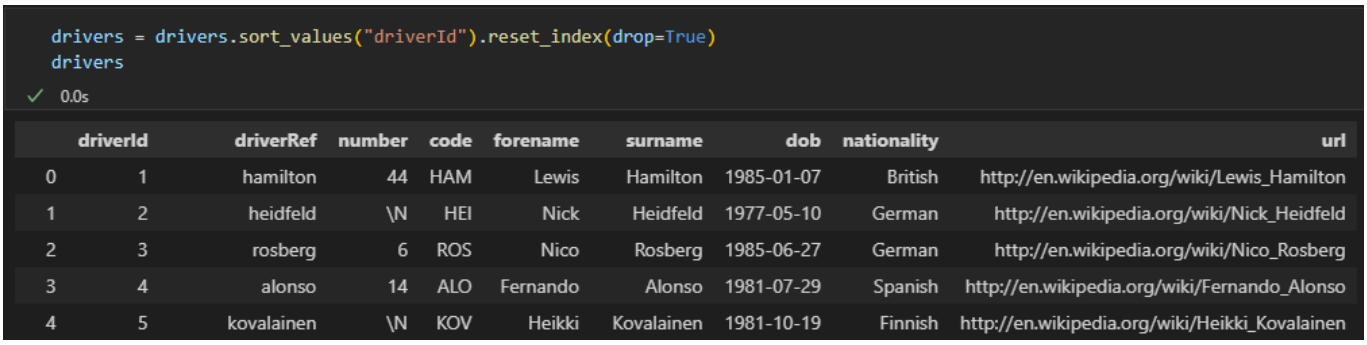
By using this dataset, we can create models that consider various factors such as circuit characteristics, driver performance, weather conditions, and more, which can impact lap times. The potential is amazing !



```
lap_times.sort_values("raceId")
```

✓ 0.0s

| | raceId | driverId | lap | position | time | milliseconds |
|---|---|---|---|---|---|---|
| 343807 | 1 | 18 | 38 | 1 | 1:28.438 | 88438 |
| 343543 | 1 | 20 | 21 | 2 | 1:46.868 | 106868 |
| 343544 | 1 | 20 | 22 | 2 | 2:38.375 | 158375 |
| 343545 | 1 | 20 | 23 | 2 | 2:31.909 | 151909 |
| 343546 | 1 | 20 | 24 | 2 | 2:22.185 | 142185 |



```
drivers = drivers.sort_values("driverId").reset_index(drop=True)
drivers
```

✓ 0.0s

| | driverId | driverRef | number | code | forename | surname | dob | nationality | url |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | hamilton | 44 | HAM | Lewis | Hamilton | 1985-01-07 | British | http://en.wikipedia.org/wiki/Lewis_Hamilton |
| 1 | 2 | heidfeld | \N | HEI | Nick | Heidfeld | 1977-05-10 | German | http://en.wikipedia.org/wiki/Nick_Heidfeld |
| 2 | 3 | rosberg | 6 | ROS | Nico | Rosberg | 1985-06-27 | German | http://en.wikipedia.org/wiki/Nico_Rosberg |
| 3 | 4 | alonso | 14 | ALO | Fernando | Alonso | 1981-07-29 | Spanish | http://en.wikipedia.org/wiki/Fernando_Alonso |
| 4 | 5 | kovalainen | \N | KOV | Heikki | Kovalainen | 1981-10-19 | Finnish | http://en.wikipedia.org/wiki/Heikki_Kovalainen |

## II- Data Preprocessing

Data preprocessing is a crucial step in building any machine learning model, as it ensures that the data is cleaned, transformed, and formatted in a way that can be used efficiently by the models.

The data preprocessing steps for this project include:

- Sorting the data by year, race, lap, and position to ensure chronological order.
- Dropping rows with missing values (e.g., "\N" rows from the pit_stops dataset), ensuring that our models have complete information to make predictions.
- Resetting the index to avoid any potential issues when merging datasets or referencing specific data points.

These steps ensure that our data is ready for the subsequent feature engineering and analysis stages.

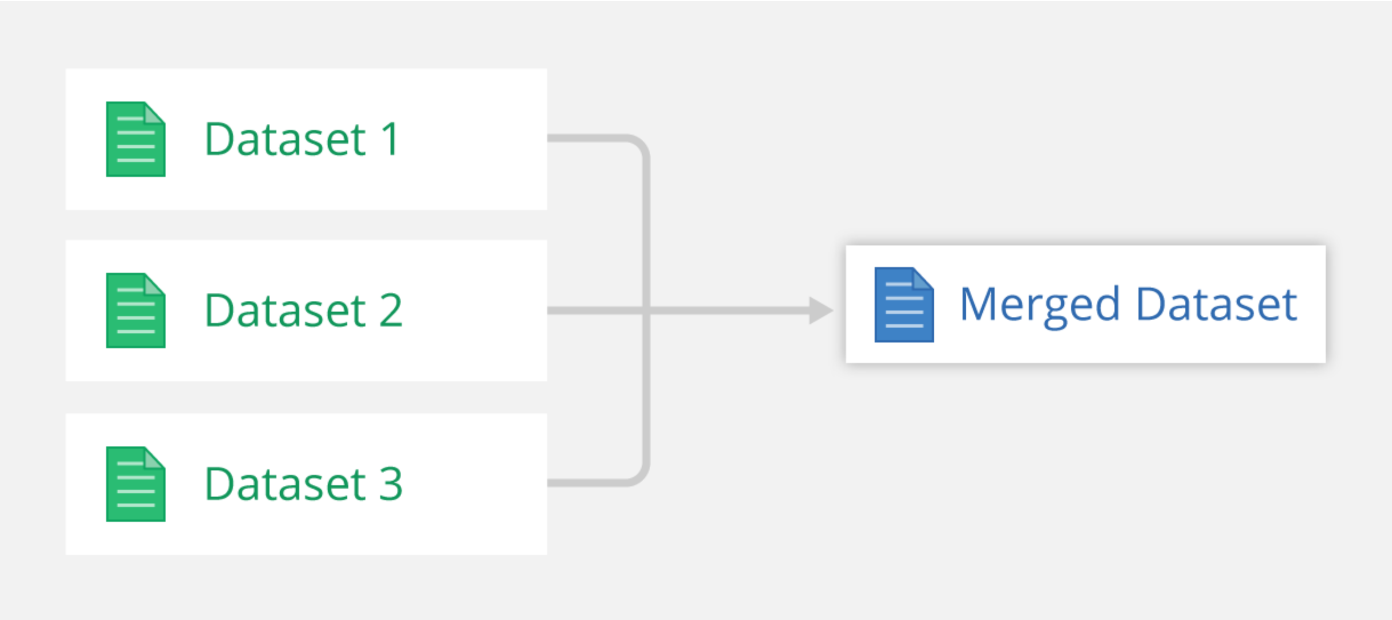
# III- Feature Engineering

Feature engineering is the process of transforming raw data into features that can be used by machine learning algorithms to make predictions. In this project, we focus on creating meaningful features that capture the nuances of Formula 1 racing and can help us make accurate lap time predictions. To achieve this, we perform the following steps:

- Join the *lap times*, *races*, *pit stops*, and *drivers* datasets to create a comprehensive dataset containing information about every aspect of each race.

- Calculate the number of laps since the last pit stop for each driver, as this information can be critical in understanding the impact of tire degradation and race strategy on lap times.

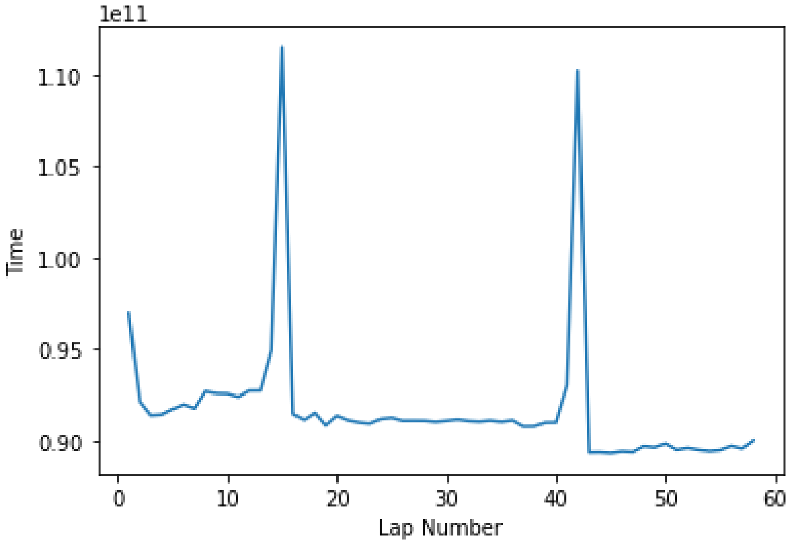
By incorporating these features, our models can better understand the underlying factors that influence lap times and make more accurate predictions.



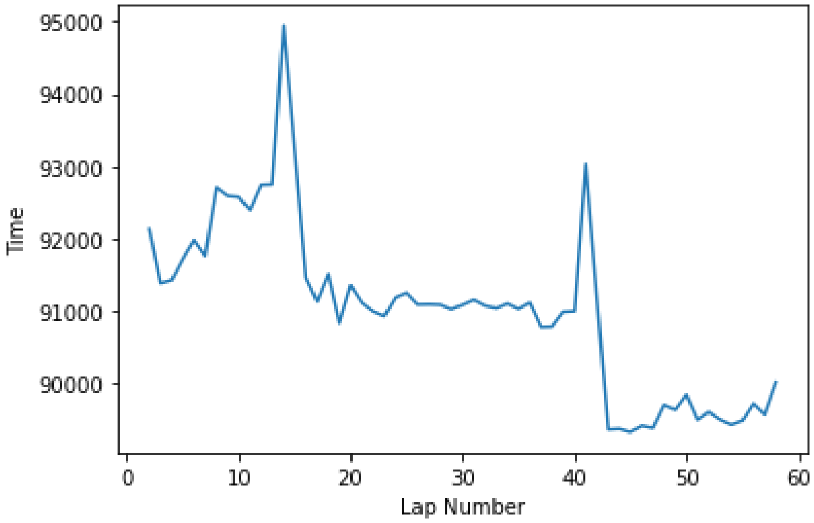
# IV- Exploratory Data Analysis

Before diving into modeling, we perform exploratory data analysis (EDA) to gain insights into the data and identify any trends or patterns that could help us in our modeling efforts. In this project, we focus on analyzing lap times for Esteban Ocon's 2022 Abu Dhabi Grand Prix race, as it provides a recent example of how different factors can impact lap times.

We plot Ocon's lap times throughout the race, highlighting the impact of pit stops on lap times. From the plot, it is evident that pit stops have a significant impact on lap times, causing one of them to increase dramatically and then a stabilization at different height depending on the tire compounds.



Next, we remove the pit stop laps from the dataset and create a new plot showing Ocon's lap times without the influence of pit stops. This plot reveals more subtle trends in lap times, such as tire degradation and fuel consumption, which gradually decrease lap times as the race progresses.



# V- Modeling

With the data preprocessed, features engineered, and insights gained from the EDA, we proceed to train different machine learning models to predict lap times. We start by comparing regression models, including *RandomForestRegressor,* with time series models like Prophet and ARIMA.

## 1. Regressions Models

We first train a *RandomForestRegressor* model, which is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the mean prediction of the individual trees. We one-hot encode categorical variables such as *raceId*, *year*, *driverId*, and *circuitId* and normalize the continuous variables. However, the Mean Squared Error (MSE) for the *RandomForestRegressor* is quite high, indicating that this model may not be suitable for predicting lap times.

## 2. Time Series Models

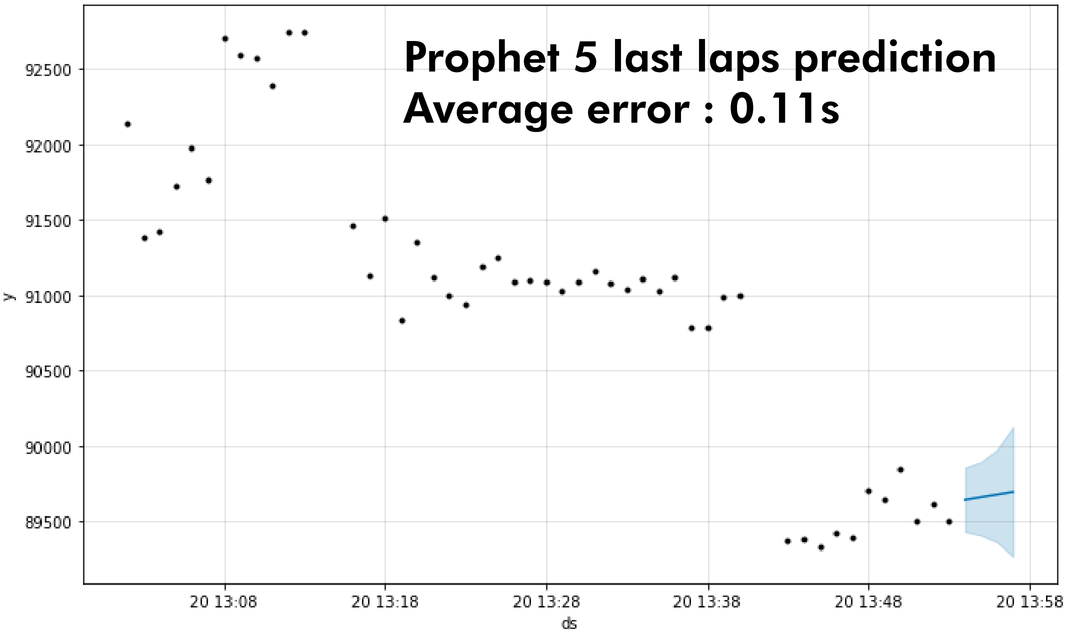
Given the limitations of regression models, we explore time series models, which are specifically designed to handle data with a temporal component. Two common time series models are the Prophet model, developed by Facebook, and the ARIMA model, which stands for Autoregressive Integrated Moving Average.

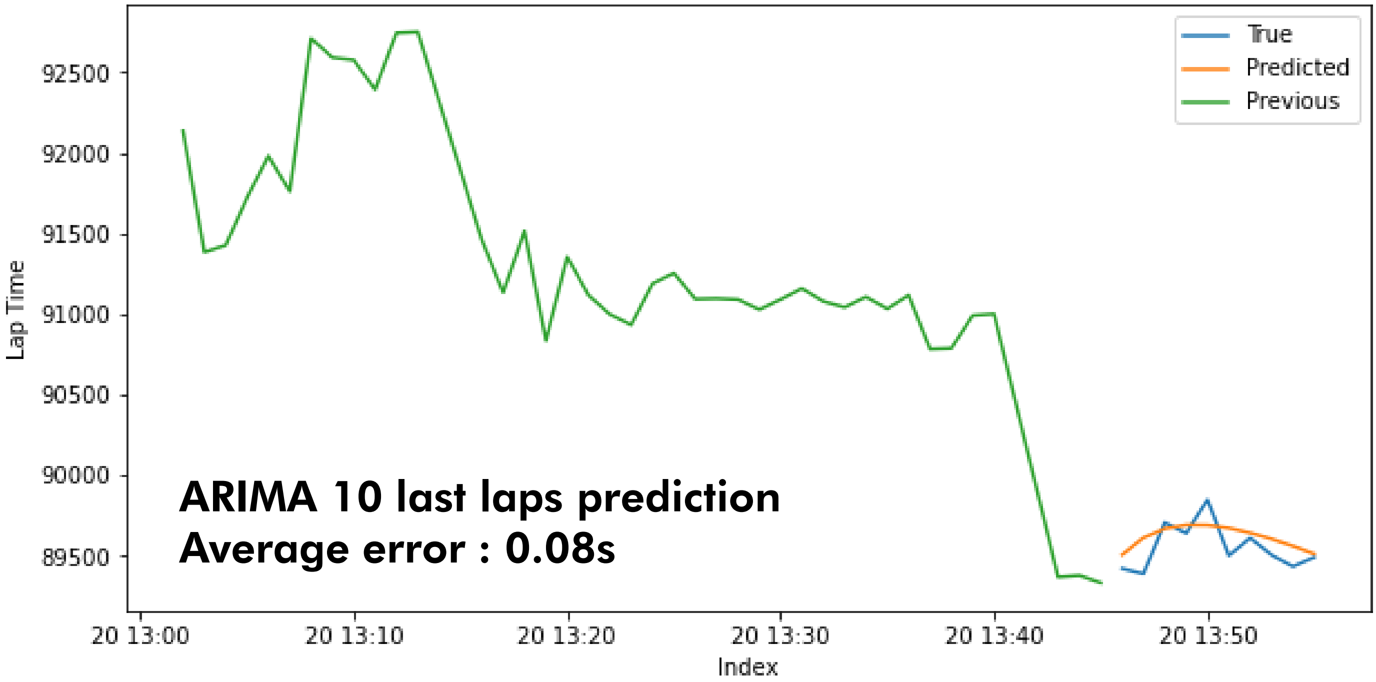
Prophet is a flexible time series forecasting model that can handle multiple seasonalities, holidays, and other special events. ARIMA, on the other hand, is a linear model that combines three components: autoregression (AR), differencing (I), and moving average (MA).

# VI- Model Evaluation and Comparison

After training the models, we compare their performance to determine which approach is best suited for predicting lap times in Formula 1 races. In this case, we evaluate the models based on their Mean Squared Error (MSE) and visual inspection of their predictions.

The *RandomForestRegressor*, as mentioned earlier, yields a high MSE, indicating that it may not be the best choice for predicting lap times. In contrast, the time series models, such as Prophet and ARIMA, demonstrate better performance, both in terms of their MSE and the visual inspection of their predictions.





# What can we remember ?

In this report, we explored the potential of various machine learning models to predict lap times in Formula 1 races. Our analysis indicates that time series models, such as Prophet and ARIMA, are more suitable for this task compared to regression models like *RandomForestRegressor*.

These models can capture the temporal patterns and underlying trends in the data more effectively, leading to more accurate predictions of lap times.

As a result, Formula 1 teams can leverage time series models to devise better race strategies and make informed decisions during races, ultimately leading to improved performance on the track.

Further research could involve incorporating additional features, such as weather conditions, tire compound choices, and telemetry data, to enhance the predictive capabilities of the models. Additionally, advanced time series models, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, could be explored to further improve lap time predictions.

Never stop pushing !