University of BRISTOL

DEPARTMENT OF COMPUTER SCIENCE

Weather Impacts on Formula1: Predicting Lap Times in Dynamic Racing Environments

Hyunho Kim

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering.

Friday 10th May, 2024

Contents

Abstract

Project Context and Objectives

This project aims to develop a robust machine learning framework that predicts lap times accurately in Formula 1, focusing on the influence of diverse weather conditions. Additionally, this study enhances understanding of environmental impacts on racing, suggests optimal race strategy planning, and contributes to dynamic sports analytics and data science fields.

Feature engineering was systematically constructed to recognize the interaction between varied weather conditions and race dynamics. Sophisticated data preprocessing steps were employed to address missing values of the dataset. Several advanced machine learning models were employed, including RandomForest and XGBoost, and hyperparameter tuning techniques were used to improve the accuracy of the prediction.

In summary, by utilizing data from 2019 to 2022 for model training and validation, and the latest year 2023 to predict the lap times which were unseen data to the model, this project has developed a machine learning framework that predicts the lap times with high accuracy, concentrates on the impact of the various weather conditions on race performance.

Dedication and Acknowledgements

This work is dedicated to all those who have supported and guided me throughout the journey of this project. I am deeply grateful to my family and friends for their sincere support and encouragement.

Special thanks to my supervisor, Dr.James Cussens, whose expertise and insightful guidance have been invaluable to my research. His patience and knowledge were crucial in overcoming the several challenges encountered during this research.

Ethics Statement

- This project did not require ethical review, as determined by my supervisor, Dr.James Cussens

Supporting Technologies

The project extensively employs the Python programming language with a variety of Machine Learning libraries:

- Scikit-Learn: Fundamental library for machine learning, including data preprocessing, model selection, and evaluation etc.
- Fastf1 library: Applied data collection via the library, which provides access to historical Formula 1 data.
- Lucid chart : Used to generate flow of the diagram of this project.

Chapter 1

Introduction

Formula 1 racing, known as the fastest race in the world, is characterized by its high speed, technologically sophisticated vehicles that could reach speeds up to 370 km/h and exhibit more than 50 percent thermal efficiency [6]. In this intense environment, drivers endure extreme g-forces and must make immediate decisions under demanding conditions. Every team build precise strategies to adapt the quickly changing race conditions, aiming to maximize driver focus and their performance. Each outcomes the races are impacted by complex factors, such as driver skills, vehicle aerodynamics, team strategy, and weather conditions.The unpredictable weather conditions present significant challenges for outstanding lap times and optimal strategies in the circumstance. For example, unexpected rainfall or changes in track surface temperature could significantly change race conditions, which forces teams to constantly adjust their strategies in real time.

Given the complexity of these dynamic environments, Formula 1 teams utilize advanced data analytics and machine learning to manage data effectively [14]. These technologies are not only beneficial in predicting race outcomes but also enable teams to make the best decisions under rapidly changing conditions.
The integration of technology into strategic planning highlights the growing importance of data science and the machine learning field in dynamic environments like Formula 1 racing [5].

Contributions to Knowledge

This dissertation enhances the field of dynamic analytics by introducing a detailed methodological framework designed to predict performance metrics in environments that are unpredictable and dynamic. The contribution of this dissertation lies in its novel integration of environment and historical data analysis with predictive modeling to support decision making processes during live Formula 1 races.
By utilizing machine learning models that capture weather variables such as rainfall, humidity, and wind, which significantly affect the aerodynamics of the cars and their performance, this research offers a robust tool for strategic race planning.

The integration of multiple data sources and the application of sophisticated machine learning techniques provide insights into the complex interactions between race conditions and performance. Thus, this work potentially guides future strategies not only in Formula 1 racing but also across various industries that require dynamic environmental feature handling. Furthermore, the application of machine learning also enhances the experience of individuals watching Formula 1 races. For instance, tools similar to Amazon's tyre performance estimator provide crucial decision support capabilities. Such tools could engage viewers by encouraging them to consider potential strategies, thereby deepening their understanding and involvement in the race [3].

Project Aims and Objectives

This study develops a robust machine learning framework to predict Formula 1 lap times, focusing on the impact of various weather conditions. In addition, aims to expand understanding of environmental impacts on racing, improve race strategy planning, and contribute to dynamic analytics and data science fields.

Chapter 2

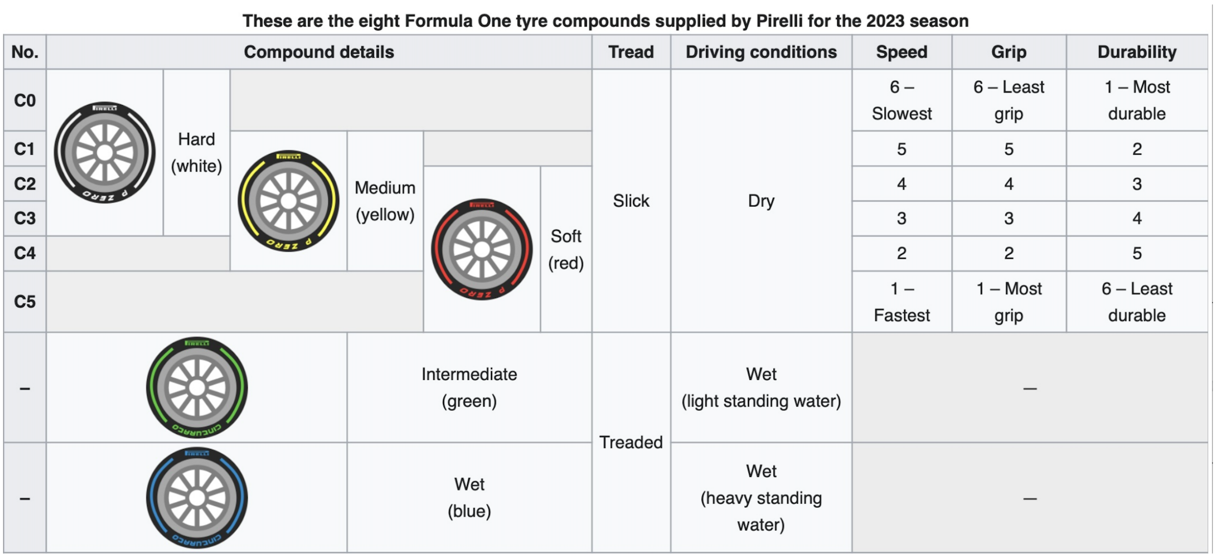
Background

2.1 Weather and Formula 1

Weather significantly influences the outcomes of Formula 1 races, affecting team strategies and car performance, driver conditions, and overall lap times. This section explores the impacts of these factors.

2.1.1 tyres in Formula 1

Outcomes of a Formula 1 race is determined by complex interactions between various factors, where the performance of the tyre is particularly important. Since these tyres are devised to maximize performance under certain conditions, proper strategic choices are required corresponding to weather conditions and track characteristics.



These are the eight Formula One tyre compounds supplied by Pirelli for the 2023 season

No.	Compound details			Tread	Driving conditions	Speed	Grip	Durability
C0	Hard (white)			Slick	Dry	6 – Slowest	6 – Least grip	1 – Most durable
C1		Medium (yellow)				5	5	2
C2			Soft (red)			4	4	3
C3						3	3	4
C4						2	2	5
C5						1 – Fastest	1 – Most grip	6 – Least durable
–	Intermediate (green)			Treaded	Wet (light standing water)	–		
–	Wet (blue)				Wet (heavy standing water)	–		

Figure 2.1: Types of tyres in Formula 1

In dry conditions, teams choose from three tyre compounds: Soft, Medium, and Hard, as illustrated in Figure 2.1. The detailed compounds range from C0 (hardest) to C5 (softest), Although specific compound data were not available, the terms Soft, Medium, and Hard are used for general reference [27]. Each compound offers distinct performance advantages; the hardest tyres provide durability, the softest offer increased higher speeds but the soft compounds foster tyre degradation than hard tyre, and the medium offers a balance between speed and durability. For rainy conditions, teams switch to tyres specifically designed to not slip with the wet track: intermediate tyres for light rain and wet tyres for heavy rain respectively [21].

2.1.2 Impact of Extreme Weather Conditions

Various weather conditions significantly influence Formula 1, impacting driving style, car performance, and team strategies. Proper adaption and strategies to these dynamic conditions are crucial to the teams during races.

Extremely High and Cold Temperature

Warm and dry temperatures are ideal circumstances for the Formula 1 race, which provide the best performance of aerodynamics and grip of the car as it is designed. However, if the temperature becomes extremely high, the hot environment cause overheating of the tyres which could leads to reduced tyre grip and tyres wearing out quickly, while cold temperatures prevent reaching the optimal operating conditions, impacting negatively their performance [15].

Humidity

In very humid circumstances, such as Singapore or Monaco Grand Prix, the environment causes a lack of grip between the tyres and track. Also well known for its very hot and humid circuit, the Singapore GP, typically drivers lose 3-4kg in body weight which is approximately four liters of sweat[15].

Altitude and Pressure

Racing at high altitudes, like in Mexico City, presents challenges due to lower air pressure and thinner air, which affects engine performance and aerodynamics. Teams need to adjust their cars to response with these conditions, which can include changes in strategies and aerodynamic setup to for reduced air density [7].

Rain

Racing in rain poses several tricky challenges to the drivers, reduced grip, increased risk of aquaplaning of a car, and reduced visibility, which are deadly dangerous problems to drive at extremely high speeds. These challenges and additionally required pit stop to change the tyres to intermediate or wet tyres could result in overall slow lap time [15].

Wind Speed and Direction

The wind is also an important feature of the car dynamic, especially in Formula 1, in which aerodynamic plays a significant role in the race results, also wind highly affects the cooling system.

2.2 Overview of Machine Learning

> A computer program is said to learn from experience **E** with respect to some class of tasks ***T*** and performance measure **P**, if its performance at tasks in T, as measured by P, improves with experience E - Mitchell, 1997.

This is one of the well-known definitions of Machine learning [17]. In exploring the applications of machine learning to predict outcomes in dynamic environments like Formula 1 racing, it is crucial to align methods within established theoretical frameworks. This foundational definition not only implies the essence of machine learning but also provides insight into this project's approach to predicting lap times under various weather conditions:

Experience *E*, in the context of this project regarded to the collection of extensive of data from previous Formula 1 races. This data encompasses a variety of factors including weather features, race results, etc, comprehensive collection of data to train a model.

The Task *T*, this is the primary aim of this project: predicting lap times involves analyzing how changes in different weather conditions affect the performance and outcome of a race.

Performance Measure *P*, metrics are crucial for evaluating the accuracy of predictions of a model and for continuous improvement of the models, including metrics like the Root Mean Squared Error (RMSE), elaborated in Evaluation 3.2.

2.2.1 Types of Machine Learning

Machine learning models are mainly categorized into three main types, each suited to different kinds of data and learning objects.

Supervised Learning: Supervised learning models utilize labelled data, where each instance includes input features x (e.g., weather conditions, track temperature, humidity) and a response variable y ('LapTime' in this project). The primary goal is to minimize the error between the model's predictions $\hat{y}$ and the actual lap times y, which leads to enhanced prediction accuracy [11].

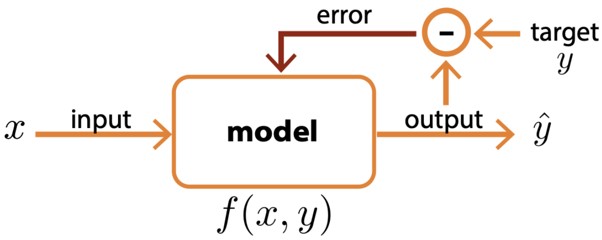


Figure 2.2: Supervised Learning Process

Unsupervised Learning: Unlike supervised learning, in unsupervised learning, models are provided only with input data x and they identify patterns within the given data. This method has strength for identifying latent relationships in the data [11].

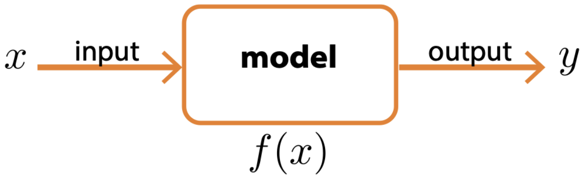


Figure 2.3: Unsupervised Learning Example

Reinforcement Learning: In reinforcement learning, models are not provided with explicit correct outputs. Instead, they receive rewards r or punishments based on the outcomes of their actions [11].

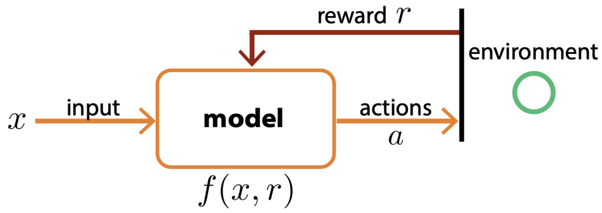


Figure 2.4: Reinforcement Learning

Using 'LapTime' as the Response Variable:

In this project, supervised learning is adopted to predict lap times, which is the response variable of this project; the term response variable is used when models predict a number.

2.2.2 Linear and Non-Linear Relationships in Machine Learning

The main objective of accurately predicting lap times in Formula 1 is driven by understanding how various racing variables interact with each other and their impact on performance outcomes distinguishing between linear and non-linear relationships, each of which has unique properties in the environment of the Formula 1 races.

A linear relationship implies that changes in features such as temperature or humidity result in proportional changes in lap times.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \tag{2.1}$$

Where y_i is the response variable, β_0 is the intercept, a constant variable that represents the prediction when all features x_i are zero, β_1 is the slope of the predictor x_i, and ϵ represents the error [1]. In other words, an increase in the features x_i, such as temperature or humidity would lead to a corresponding adjustment in lap times.
However, this assumption may fails to reflect the complex interaction on the track, where variables like track temperature and tyre performance interact in non linear way.

Non-linear relationships, introducing polynomial features allows transforming input features into polynomial terms, a model provides a multidimensional curve to fit the data points, which expands the model's capabilities beyond the limitations of a straight line. Unlike to the linear relationship, however, the relationship between track temperature and lap time is not straightforward (i.e., increasing temperature does not always mean simply decreasing or increasing lap time in a straight line). Initially, increasing temperature may reduce lap time (to the best performing point on the tyre), but above a certain temperature, the lap time may increase again as the tyre overheats and loss of grip.
The multiple linear regression model extends the simple linear regression by incorporating multiple predictor variables. It is expressed as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \epsilon \tag{2.2}$$

y represents the response variable for the i-th observation, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the predictor variables $x_{i,1}, x_i^2, \ldots, x_n^k$, and ϵ is a random error. [2].

The implementation of a sixth-degree polynomial model with this project features can be reproduced as follows:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6 + \epsilon \tag{2.3}$$

The following linear and polynomial model shows the phenomenons.

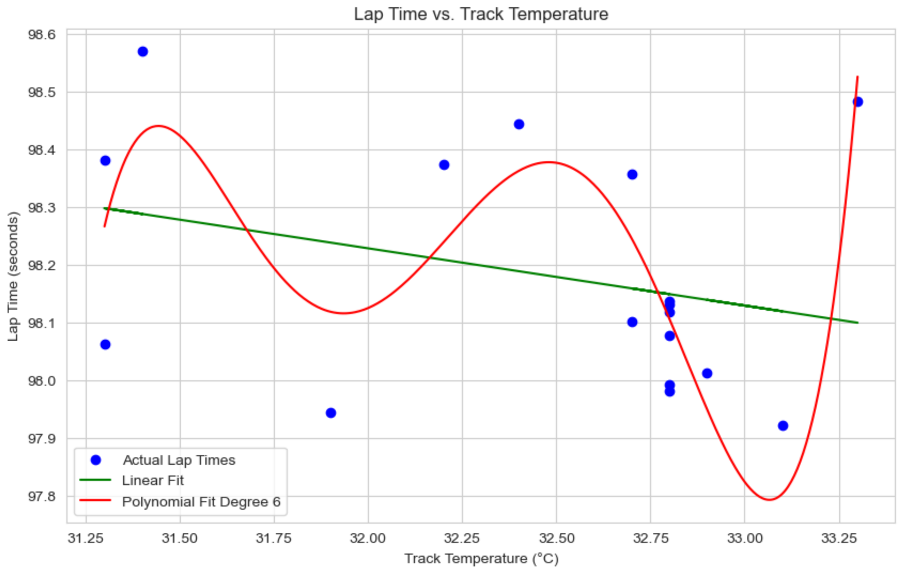


Figure 2.5: Comparison of linear and non linear fitting

As an example, figure 2.5 generated from one specific driver Carlos Sainz Jr in 2018 Bahrain Grand Prix, clearly visualize the each phenomenons.
When actual lap times are plotted against track temperature, the limitations of the linear model become apparent at the extremes of the temperature range. the linear model, shown by a green line, fails to adequately capture the complexities of the dataset.
In contrast, a polynomial fit (six degrees), denoted by a red curve, shows an enhanced alignment with the actual data points, accurately reflecting the complex impact of track temperature variations on lap times.

The formula 2.3 demonstrates how each incremental power of x (track temperature) contributes to the model's ability to follow the nuanced changes in lap times, with ϵ representing the error term.
However, selecting the degree of the polynomial requires careful consideration. Although higher degree polynomials are able to model more complex curves, they also introduce a significant challenge in machine learning known as *Overfitting*. It could occurs when a model training some noise as significant data, which undermines ability of models to generalize to new, unseen datasets.

To capture and measure these non linear interactions, advanced machine learning models are essential. Random forests and gradient boosting machines like XGBoost are representative capable of understanding non linear relationships through their unique structures. This enables them to interpret interactions between multiple features more effectively than linear models.

2.2.3 Advanced models: Ensemble Methods

Ensemble methods enhance machine learning results by connecting multiple models. This idea uses the strengths rather than relying solely on a single model and leverages the capabilities of multiple algorithms. The fundamental idea of ensemble methods is that a collection of "weak learners" collectively form a "strong learner". By averaging the predictions from the multiple models, the ensemble improves the overall performance.

$$f(y|x) = \frac{1}{|M|} \sum_{m \in M} f_m(y|x) \tag{2.4}$$

This expression illustrates the main concept of an ensemble method: $f(y|x)$ represents the prediction made by the ensemble, where $f_m(y|x)$ are the predictions from the individual models within the ensemble, and M represents all models in the ensemble[18].

Bagging and boosting are two representative types of ensemble methods that show how these concepts are implemented in practice :

Bagging and Random Forest

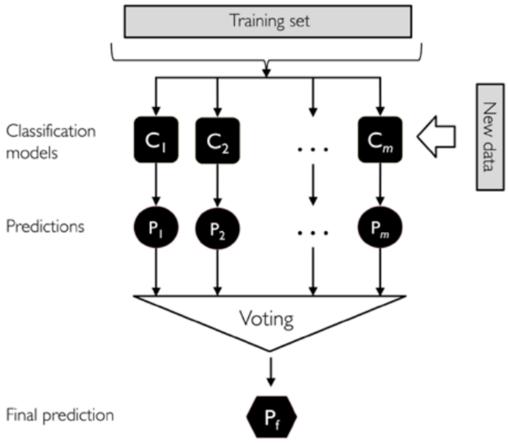


Figure 2.6: bagging

Bagging, or bootstrap aggregating, is a key technique in ensemble learning that utilizes multiple classifiers to improve prediction accuracy. Each classifier is trained on a unique bootstrap sample, a randomly drawn subset with replacement from the original training dataset. Unlike other ensemble methods that use the same data to train all classifiers, this variation in training sets improves the accuracy and reliability of the predictions.

The process of bagging is as follows :

1. Bootstrap Sampling: Each classifier in the ensemble is trained on a unique bootstrap sample drawn from the original dataset. These samples are created by randomly selecting instances from the dataset with replacements, meaning the same instance can appear more than once in a sample.

2. Training Classifiers: The classifiers are trained independently on their subsets respectively. They are allowed to grow fully without being pruned(removed) to capture complex patterns and interaction in the dataset.

3. After the classifiers are trained, the final prediction for a specific input is determined by the consensus among the classifiers. This is achieved by amalgamating the predictions from each classifier through a method commonly referred to as majority voting[22].

Overview of Random Forests:

CART, Classification, and Regression Trees, also known as decision tree is the primary component of random forests, that operate by recursively dividing the input space into distinct regions. To perform this steps, the data is divided at each tree node according to feature thresholds. After more splitting, each leaf that results reflects a distinct prediction outcome for the inputs grouped into that zone.

Random forests are ensemble learning models that combine the predictions of multiple decision trees to generate more accurate and reliable prediction than a single tree could provide. In these models, each tree is trained from a different subset of the data, which helps reduce the risk of overfitting; for regression tasks, predictions from individual trees are averaged to enhance accuracy, while a weighted majority vote is typically used for classification task. By aggregating the outcomes of various trees, random forests improve its performance significantly [18]. This unique system allows random forests to manage high dimensional spaces and large non linear interactions between features efficiently. With these strengths, they are widely used in various applications from predictive analytics to feature selection in complex dataset [22].

Boosting and Gradient Boosting

This process, as seen in the figure 2.7 displays that each classifier, indicated by $y_m(x)$, is adjusted depending on how its predecessor performed by changing the weights of the training set (indicated as

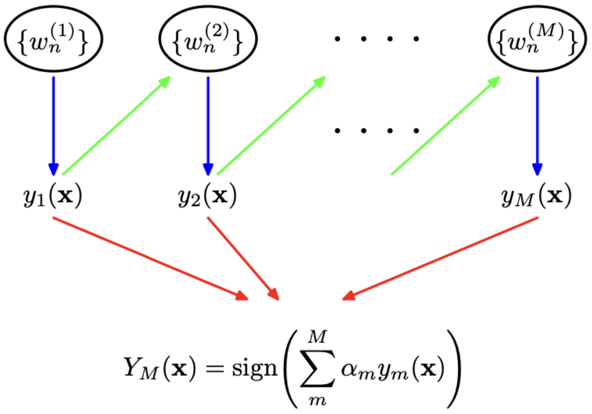


Figure 2.7: framework of a boosting algorithm

blue arrows). In later stages, the model focusing more attention to incorrect predictions, focusing more on the more challenging examples.

AdaBoost is one of the popular boosting algorithm designed to sequentially improve a predictive model by focusing on the most challenging data points. The following steps describe the AdaBoost procedure:

1. Begin with a set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and corresponding target variables $t_1, \ldots, t_N$, initializing each data point's weight to $w_n^{(1)} = \frac{1}{N}$.
2. Iteratively perform the following for $m = 1, \ldots, M$:
   (a) Fit a weak classifier $y_m(\mathbf{x})$ to the weighted training data by minimizing the weighted error function:
   $$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \tag{2.5}$$
   (b) Compute the weighted error rate ε_m of $y_m(\mathbf{x})$:
   $$\varepsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{2.6}$$
   (c) Calculate the classifier weight α_m based on the calculated ε_m:
   $$\alpha_m = \ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right) \tag{2.7}$$
   (d) Adjust the weights for the following round:
   $$w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\} \tag{2.8}$$
   where the weights are updated focusing on incorrect predictions for each data point, $y_m(\mathbf{x}_n) \neq t_n$.
3. Construct the final model as:
   $$Y_M(\mathbf{x}) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x})\right) \tag{2.9}$$

This algorithm sequentially enhances the model by assigning more weight to the data points where incorrectly predicted in previous iterations, ensuring that each new classifier is adjusted to address these points more effectively [9].

Gradient Boosting in XGBoost

As discussed above, boosting, unlike bagging, improves predictions sequentially, by focusing especially on the errors made by earlier models. One improved methodology is gradient boosting, notably implemented in XGBoost, where the term gradient in gradient boosting refers to its use of the gradient descent algorithm to minimize the loss function. In this approach, trees are constructed sequentially, with each new tree specifically focus correcting the residual errors (differences between predicted and actual values) left by the previous trees. This approach of models aims to minimize a loss function, effectively reducing both bias and variance, thereby enhancing accuracy with each step.[18].

Figure 2.8 illustrates the operational flow of XGBoost, showing how each component of the model interacts to enhance predictive performance. While it also provides a visual representation of the progressive building of trees, showing its iterative error correction mechanism[12].

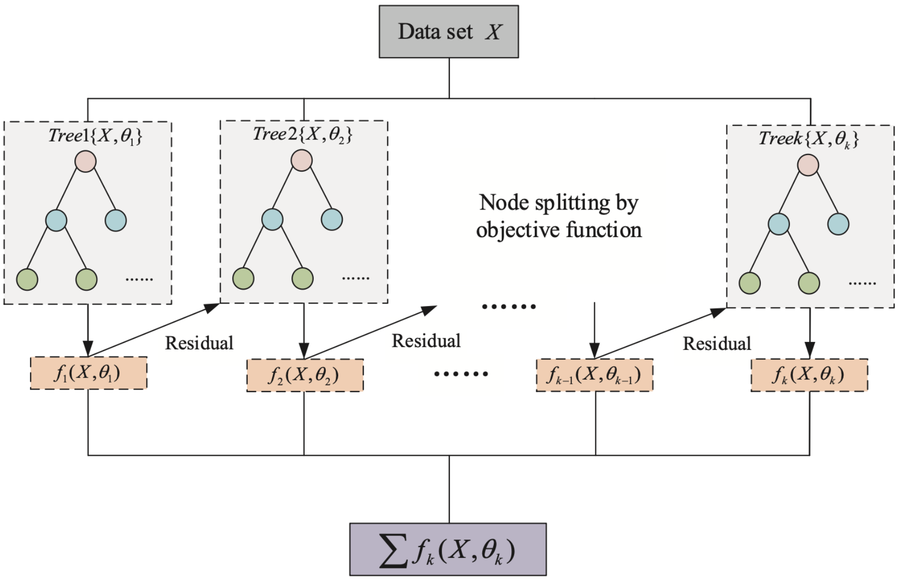


Figure 2.8: Flow of XGboost

XGBoost, or Extreme Gradient Boosting, is recognized for its high optimization and robust performance within the gradient boosting framework. XGBoost applies machine learning algorithms to provide a strong, scalable, and fast solution for boosting trees. This model has become very popular due to its excellent performance; as an illustrative example in 2015, it was used in 17 of 29 winning entries in Kaggle competitions.[8].

2.2.4 Practical Application in Formula 1 Lap Time Prediction

In the context of predicting Formula 1 lap times, these ensemble techniques are invaluable. By leveraging bagging and boosting, the models could offer enhanced performance with the complex interactions of complex variables like weather conditions, track characteristics, and car performance.

The workflow of this project is visually represented below. This diagram illustrates the overall process from data collection to model application, inspired by a machine learning text book[22]. The structured approach ensures that each phase of the project, from initial data acquisition to the final model prediction.

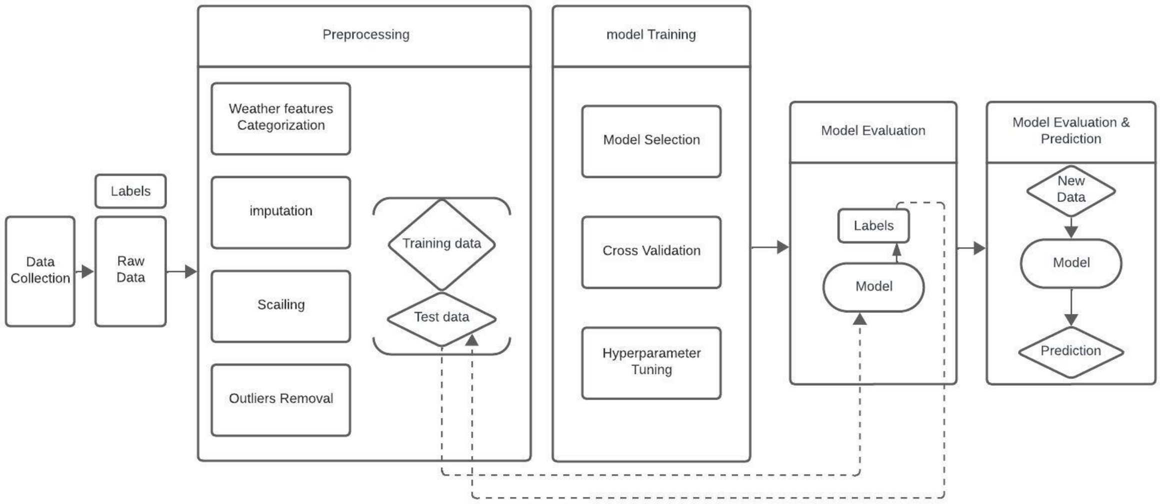


Figure 2.9: Flow of this project

Chapter 3

Project Execution

3.1 Data collection and preprocessing

Introduction to Data
This section explains the foundational elements of the data collection process. The dataset covers various telemetry and lap data sourced from the Fastf1 Python library, powered by the Ergast API (https://docs.fastf1.dev/#), which provides detailed race data for each Grand Prix since the year 1950. However, data from 2019 to 2023 was selected to secure data integrity and consistency,particularly due to significant regulation changes in Formula 1 during the past period(such as engine changes).
Selecting a limited dataset within recent years, this selection decreases the risk that the model learns from irrelevant and noise. Subsequently mitigating overfitting, a critical issue when models only trained well on an extensive number of datasets and aligned with irrelevant data points, fail to perform well on new, unseen data.

3.1.1 Data preparation

Data preparation is a fundamental step of a machine learning process which is performed in the initial phase, where the raw data is cleaned and transformed properly into a suitable format to analyze and train for machine learning models. Several steps were involved in this project:

Encoding String Features

Boolean variables, such as *Rainfall*, in the dataset provided as string format (True or False). Transforming these strings to numerical value is an essential step in machine learning since most machine learning algorithms require numerical input to understand. Consequently, the Boolean features were encoded to binary integers (1 or 0). A detailed description of the raw data features can be found in Appendix A(in Section A.1).

One hot encoding

In Formula 1, there are 20 drivers and over 20 circuits worldwide. Each driver and circuit has unique characteristics that can significantly influence lap times. To address the individual characteristics and impacts of categorical variables such as Driver and Circuit, one-hot encoding was applied.
One-hot encoding, also known as dummy encoding, transforms these categorical features into a numerical scale. This method creates binary columns, one for each category, ensuring that the model recognizes and evaluates the distinct impact of each driver and circuit [18].

The following example illustrates the one-hot encoding process applied to the categorical variable Circuit, as an example, the *British Grand Prix.* In this method, each category value is transformed into a binary column and assigned a 1 or 0, depending on the presence of the category:

$$\text{British Grand Prix} = \begin{cases} 1 & \text{if the race is at the British Grand Prix,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

As represented in Equation 3.1, the one-hot encoding technique assigns binary values according to the location of each circuit. Below is an example that visualizes this technique for selected circuits in the dataset: British Grand Prix, *Monaco Grand Prix*, *Italian Grand Prix*, *Japanese Grand Prix*, and *Canadian Grand Prix.*

Data Frame After One-Hot Encoding

British Grand Prix	Monaco Grand Prix	Italian Grand Prix	Japanese Grand Prix	Canadian Grand Prix
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Table 3.1: Data after one hot encoding

The above table 3.1.1 illustrates the result of applying one-hot encoding, where each category within the *circuit* variables is transformed into newly created binary columns by indicating as 1 if the row is the corresponding circuit. This transformation is crucial as it ensures that each driver's and circuit's unique impact on lap times. Additionally, irrelevant or duplicated features, such as *LapStartDate*, and *LapStartTime*, are removed.

3.1.2 Handling Outliers

Handling outliers is one of the significant processes in developing machine learning models. Outliers could skew data, which could lead to inaccurate mean and variance estimates. This could significantly impact on the performance of predictive models, by failing to capture the underlying trends properly. As visualized in Fig 3.1, the distribution of lap times, the distribution of lap times shows a positive skew.

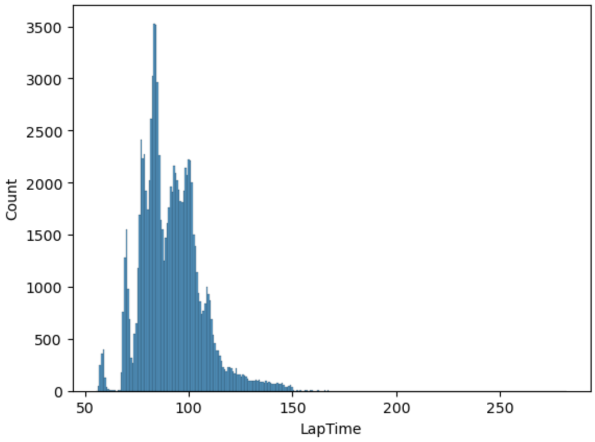


Figure 3.1: Distribution of lap time

Figure 3.2: The histogram showing positive skew which means the tail on the right side than the left side

In this project, the Interquartile Range (IQR) is utilized to manage outliers effectively. The IQR method is particularly advantageous because it does not require the data to follow a normal distribution, aligning well with the distribution of lap times as shown in Fig 3.1.

Fig 3.3 shows detailed description of IQR:

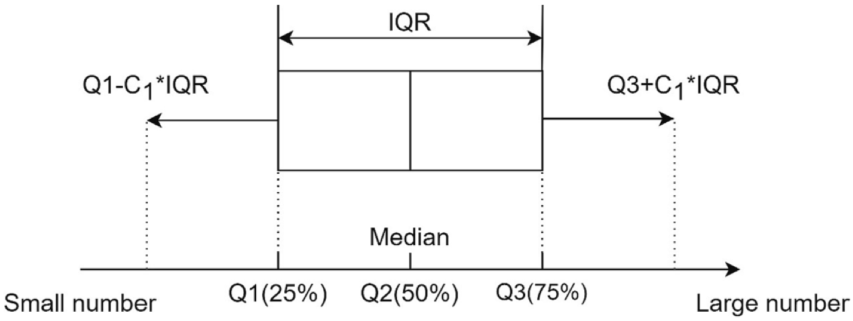


Figure 3.3: Interquartile Range

- **Minimum** and **Maximum Values**: The minimum and maximum values of the data set are represented at the far left and right of a box plot data visualization respectively.
- **Median**: The median is located in the center of the box.
- **Q1 (First Quartile):** This represents the 25th percentile of the data, meaning 25% of data points fall below this value. It captures the median of the lower half of the data.
- **Q3 (Third Quartile)**: This marks the 75th percentile, meaning 75% of data points are below this value. It represents the median of the upper half of the data.
- **Interquartile Range (IQR)**: Defined as Q3 minus Q1 (IQR = Q3 - Q1), the IQR measures the middle 50% of the data, providing a statistical measure of data spread that is less sensitive to outliers.
- **Outliers**: Values that fall below Q1-1.5×IQR *or* above Q3+1.5×IQR are typically considered outliers.

The IQR method eliminates these extreme values by concentrating on the statistical spread of the data, thus preserving the dataset's integrity for more accurate results. This technique could be highly beneficial when dealing with skewed distributions, such as lap times as shown, where different factors could have significant effects on the performance while preventing skewed outcomes caused by uncleaned outliers.[28].

3.1.3 Understanding Informative outliers

Understanding and managing outliers in Formula 1 is a complex task where outliers; mainly slower lap times could be caused by numerous factors including accidents, technical failures (e.g., engine failure), or abnormal weather conditions like rainy days. These exceptional events are indeed outliers, however, they may contain important information about the race dynamics that affect racing performance. Therefore, it is required to distinguish between "noise" and informative outliers in the dataset.
In order to handle the informative outliers, *TrackStatus* features were utilized as indicators, the detailed conditions listed in the below table 3.2 were categorized and treated as informative outliers based on their status code and the required actions :

TrackStatus	Condition	Impact on Racing Conditions
2	Yellow Flag	Indicates caution to the drivers, leading to slower lap times due to reduced speed limits.
4	Safety Car	Significantly impacts lap times by reducing speeds and bunching up cars, crucial for predicting lap variability.
5	Red Flag	Red Flag not only require a reduction of the speed but also halts the race entyrely.
6	Virtual Safety Car	Deployed similarly to a safety car but with less reduction in speed.

Table 3.2: TrackStatus and their impacts on racing conditions

Balancing Rainy Days and Outliers

Furthermore, lap times on rainy days are also isolated from the noise. Under wet conditions, lap times usually increase because of reduced tyre grip. Moreover, additional pit stops to change from slick tyres to wet or intermediate tyres generally add 20 seconds to a lap time [16]. In order to prevent lap times from rainy days being flagged as anomalies, the feature *Rainfall* was used to categorize lap times into binary classifications for both dry and rainy conditions; recognizing their distinct influences on racing performance. Subsequently, the Interquartile Range was applied to the differentiated data for dry and wet conditions respectively. After refining the data for both conditions independently, the datasets were merged into a single, well-structured dataset.

This comprehensive dataset is crucial for training a machine learning model, enabling it to perform effectively across different weather conditions and provide precise predictions of lap times. By retaining these informative outliers and removing true noise, the models could perform effectively across various weather conditions and capture the expanded range of racing dynamics.

3.1.4 Data imputation and Scaling

Imputation data set

It is crucial that maintain integrity of the time series data. Inaccurate prediction might caused by several missing data which could lead to inaccurate predictions and biased results in trend analysis. For this project, the forward-fill method was selected for imputation with the dataset. This technique is selected because of its simpleness and effectiveness for time-series data where the sequence of observations is important.In order to secure the integrity of the time series data, this method forwards the last observed data point forward until a new data point is recorded [19].

Scaling Data set

To efficiently train models with a range of Formula 1 data metrics such as temperature, humidity, and other numerical values, scaling the values, the different value ranges of several variables have to be adjusted to a common scale. This step is crucial since variables with larger numerical ranges could skew the results, potentially causing the model to fit well into irrelevant patterns.

To address this issue, the RobustScaler is employed as a standardization method. While traditional approaches rely on the mean and standard deviation, which could be heavily influenced by outliers; the RobustScaler focuses on the median value and the interquartile range to mitigate the problem by ignoring the mean and standard deviation, and aligning the data with the median and adjusting it using the interquartile range. In this way, new data could be updated consistently [4].

3.2 Exploratory Data Analysis: Weather Impact on Formula 1 Performance

In Formula 1 racing, the interaction between weather conditions and car performance is crucial, significantly influencing team strategies and overall race dynamics. As discussed earlier, unpredictable weather conditions could affect the outcome of a race by altering track conditions. This section provides a detailed analysis of how various weather variables interact with each other and impact lap times.
This Exploratory Data Analysis aims to enhance understanding of how weather significantly influences outcomes in Formula 1 racing.

3.2.1 Correlation matrix with *Laptime*

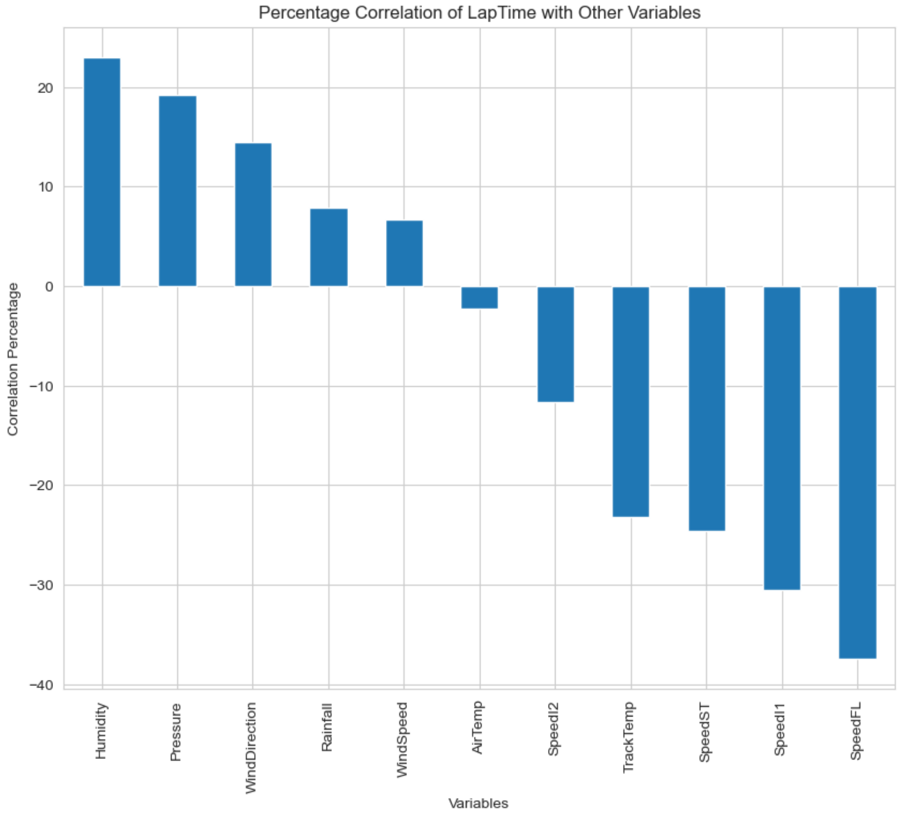


Figure 3.4: visualized correlation values to *LapTime*

The overall correlation between numerical values and lap times, as illustrated in Figure 3.4, demonstrates the significant influence of various weather conditions on racing performance. This analysis highlights how each weather conditions affect lap times. Additionally, it emphasizes the complex interaction between these factors and their impact on vehicle dynamics. As discussed in the Background section, these findings indicate how weather conditions could affect racing dynamics. The detailed results are the following:

- **Humidity** (22.99%) and **Pressure** (19.27%): Both exhibit significant correlations with lap times, suggesting that changes in humidity and pressure levels could highly affect lap times.
- ***WindDirection*** (14.46%) and ***WindSpeed*** (6.72%): These factors show a big correlation with increased lap times, indicating that higher wind speeds and less favorable wind directions may negatively affect vehicle stability and aerodynamics.
- **Rainfall** (7.87%): A moderate positive correlation suggests that rain increases lap times, likely due to reduced grip and visibility.
- **AirTemp** (-2.23%): The negative correlation here is relatively small, indicating a slight tendency for cooler air temperatures to be associated with faster lap times, possibly due to better engine performance and denser air aiding in aerodynamics.

- **TrackTemp** (-23.17%): The analysis shows a strong negative correlation of -23.17% between track temperature and lap times. This suggests that as track temperatures increase, lap times decrease, indicating improved performance.
- **Speed Metrics**:
  - **SpeedI1 (float)**: Sector 1 speedtrap [km/h]
  - **SpeedI2 (float)**: Sector 2 speedtrap [km/h]
  - **SpeedFL (float)**: Finish line speedtrap [km/h]
  - **SpeedST (float)**: Longest straight speedtrap [km/h]

  These speed measures negatively correlated with lap times which collected from multiple tracks,showing the observed values are consistent across various environments; with values ranging from -11.66% to -37.43%, showing that higher speeds at these points generally lead to faster lap times.

Further research was conducted with the top three conditions that significantly impact lap times to understand their effects in detail.

3.2.2 Track temperature

Track temperature is one of the key factors that influence lap time in Formula 1, especially tyre performance, subsequent pit stop changes, and brake efficiency.

Low Temperature

Formula 1 tyres are optimized to operate within a specific temperature range to reach the best tyre temperature balancing between grip and durability.
Cold temperatures, due to the insufficient heating of the tyres, low temperatures have the potential to considerably decrease tyre grip to address the problems, all of the teams employed tyre warmers to preheat the tyres to a desired temperature of 70 degrees Celsius as they much could stay on track before starting the race. Even with these efforts, the cold could still have a major effect on performance during the race, particularly in situations when driving intensity is reduced, such as the deployment of a safety car tyres rapidly lose their ideal heat and traction, leading drivers to zig-zagging on the track to maintain tyre temperature.

Similarly, brake performance is also highly dependent on the temperature of the race environment. Brakes need to be within a 'Goldilocks' zone to function optimally. In particular circuits such as the Las Vegas Grand Prix, teams often use the smallest brake ducts to maintain higher temperatures within the braking system, which transfers more heat to the wheel rims and, indirectly, to the tyres. Significantly low temperatures may cause insufficient friction to decelerate the car effectively. This breaking problem issues may lead to increased lap time and safety problems[13].

High Temperature

Conversely, higher track temperatures typically enhance tyre performance by maintaining the tyres within their optimal operating temperature range, which improves grip and reduces lap times.

However, this advantage is offset by the risk of overheating. Which could cause rapid tyre wear, subsequently, may require extra pit stops. The cooling breaks also could face a huge problem due to high temperatures that force the brake to heat up to 1000 degrees Celsius [15].

Moreover, beyond the pressure on the car, drivers are also put in extreme circumstances; racing in such extreme heat often leads to dehydration to the drivers. In a typical 90-minute race, a driver could lose about 3 kilograms of body fluid, which is roughly 5% of their body weight [26].

Furthermore, Each pit stop takes about 20 seconds, a lot of time in races where teams put a lot of effort into reducing even milliseconds by adjusting the number of rotations of a car's bolt [16]. Understanding the critical importance of accurate weather data, especially concerning variables like high temperatures, proper utilization of the data could be a determinative factor in developing strategies that determine success or failure.

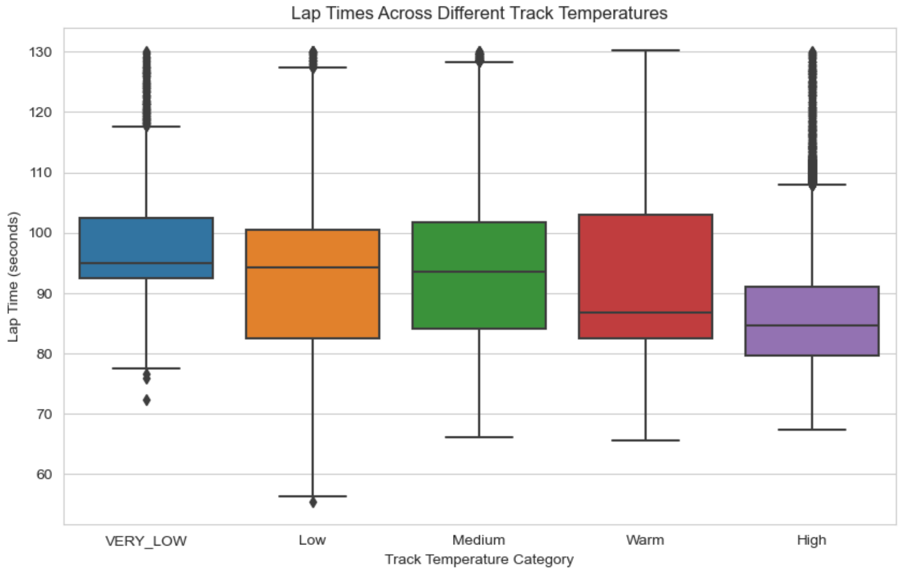


Figure 3.5: Lap Times under variable Track Temperature

The box plot 3.5, generated from track temperatures categorized via k-means clustering (Detailed in the feature engineering section), illustrates how lap times vary with temperature:

- **VERY LOW**: This category shows the highest IQR, indicating that low lap times could be achieved if proper strategies are optimized at low temperatures, despite the presence of top outliers. The longer upper whisker and higher median lap time suggest that colder conditions generally reduce performance, some drivers or teams may achieve good lap times under proper strategies optimized for low temperatures.

- **LOW**: Similar to the very low category, this category also shows a visibly low interquartile range in lap time but also has a slightly lower median. This suggests that while the consistency of lap performance improves as the temperature rises, the tyre and brake temperatures are sub-optimal and the challenge remains

- **MEDIUM**: This category has a slightly lower median and shows a balanced IQR, suggesting more consistent and improved performance as conditions approach an optimal temperature range for tyre operation.

- **WARM**: Lab time shows a large IQR. However, a notably lower median and a few outliers imply consistent performance close to the optimal tyre temperature, suggesting that the drivers are achieving near-optimal conditions.

- **HIGH**: This category has the lowest median and small IQR, indicating consistent and fast lap times at high temperatures. However, the presence of many outliers in the upper whisker also suggests the negative effects of excessive temperature such as overheating of tyres and brakes as discussed earlier.

3.2.3 Pressure and Altitude

Atmospheric pressure and altitude are critical factors in Formula 1 racing, which directly influence air density, and also affect engine performance, aerodynamics, and vehicle cooling systems, each essential for optimal race performance.

Relationship between Altitude and Pressure

Figure 3.7 illustrate the inverse relationship between altitude and atmospheric pressure; as altitude increases, atmospheric pressure decreases which leads to a reduction in air density. This phenomenon not only affects the oxygen availability for engine combustion but also impacts aerodynamic forces due to reduced air resistance, both of which are crucial for the performance of a race car [20].

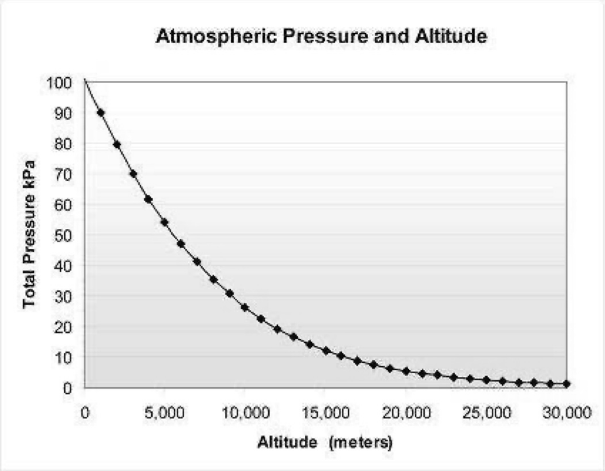


Figure 3.6: Relationship between Altitude and Pressure

Understanding the relationship between altitude and atmospheric pressure is foundational to analyzing their impact on Formula 1 race performance. Subsequently, to understand the impact of atmospheric pressure on lap times at different Formula 1 circuits, an analysis was conducted focusing on comparing high altitude and sea-level circuits. As a representative case, The Mexico Grand Prix, located at an altitude of 2,285 meters with an ambient pressure of only about 780 hPa, approximately 20% lower than at sea level, this is the most representative case study for high altitude effects. In the unique conditions in Mexico, with thin air reducing engine power by up to 25% and significant downforce reduction due to the altitude, demand innovative engineering solutions to maintain competitive performance. The altitude not only affects aerodynamics, with approximately 25% less downforce generated but also impacts the Power Unit's efficiency and cooling mechanisms due to reduced air density [7] [15].

Data from the 2019 to 2023 Formula 1 seasons at the Mexican and German Grand Prix circuits were used, where the German circuit has a mean pressure value of 992 hPa, providing a baseline as it is close to standard atmospheric pressure at sea level (1,013.25 hPa)[10].
To ensure the reliability of the analysis, both circuits were selected for their similar lengths, elevation changes, and number of corners, as detailed in Table 3.3. This analysis focuses on the effect of significant changes in atmospheric pressure at high altitudes on lap time performance, comparing a circuit at sea-level, also has similar characteristics.To enhance data accuracy, extreme values were removed using the Interquartile Range (IQR) method.

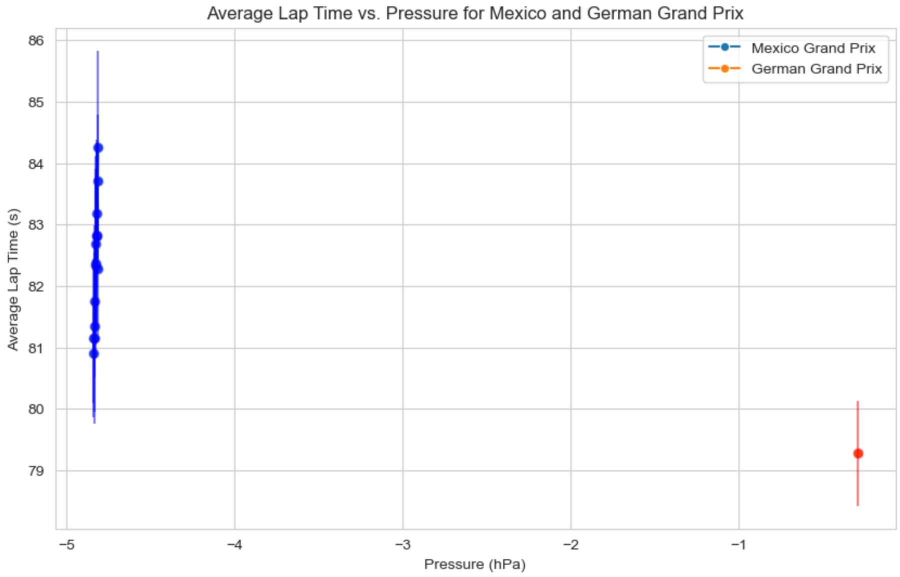


Figure 3.7: Relationship between Altitude and Pressure

Circuit	Length (km)	Mean Lap Time (s)	Mean Pressure (hPa)	Elevation changes (m)	Turns
Mexico GP	*4.421 km*	*82.1 s*	*780 hPa*	2.8 m	17
German GP	*4.574 km*	*79.3 s*	*992 hPa*	4.3 m	17

Table 3.3: Mean Lap Time comparison aligns with Atmospheric Pressure at Mexico and German Grand Prix Circuits

As shown in the figure 3.7 and table 3.3, the high altitude, in other words, low pressure presents unique challenges to the car dynamic due to its significantly reduced atmospheric pressure. This environment impacts various aspects of F1 car performance, requiring teams to make specific adaptations as lower air density affects several aspects of car performance including engine power problems and aerodynamic efficiency [7]. In contrast, the German Grand Prix Analysis, above 148,2m sea-level, located at a more low altitude, the German Grand Prix offers more stable environmental conditions. However, the results show that even though the Mexican circuit is slightly shorter than Mexico City's, the longer circuit achieves faster lap times, which may imply the impact of altitude on lap times.

The visible differences between the Mexican and German circuits highlight the necessity for Formula 1 teams to adapt their strategies to local atmospheric conditions to achieve a satisfying result in a race.

3.2.4 Humidity

According to the correlation matrix, Humidity showed the highest correlation with the response variable. High humidity cause loss of the grip with the tyre and the track.

The relationship between atmospheric humidity and car performance, specifically focusing on lap times during the Monaco Grand Prix. This Grand Prix, renowned for its challenging track and unique environmental conditions due to its proximity to the Mediterranean Sea, provides a robust case study for understanding weather-related impacts on Formula 1 racing.

Lap time data was collected from historical race performances at the Monaco circuit from 2019-2023, with corresponding local humidity levels. In addition, regression is employed to analyze and identify trends and variability in performance relative to changes in humidity. Data visualization techniques were used to support the analytical approach, providing a clear representation of the trends.

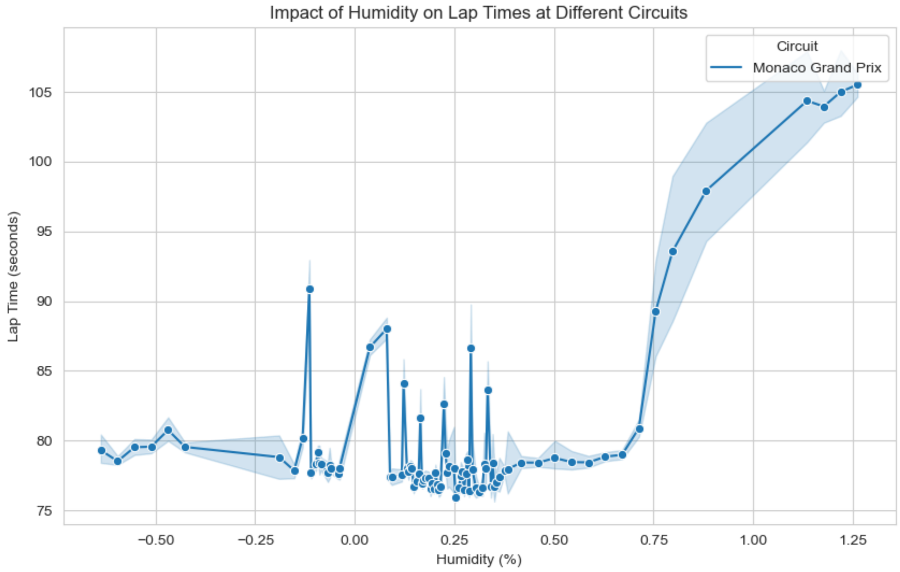


Figure 3.8: LapTime trend corresponding to Humidity

The analysis 3.8 showed a significant upward trend in lap times as humidity levels increased, especially when humidity exceeded 75% of the normalized point. Furthermore, oscillating in lap times at mid range humidity levels was also captured. Regression analysis prediction intervals are shown in the graph as blue envelopes around the trend line which indicating the expected variability in lap times for given humidity levels.

The results support the hypothesis that higher humidity affects race performance because of the effect on tyre grip and engine efficiency. In addition, the variability in lap time may indicate that teams adjust their strategies corresponding to the changing conditions. the coastal environment of Monaco might causing rapid changes in humidity, which presents challenges to predictive modeling in race conditions.

3.3 Feature Engineering

This section specifically focuses on the application of feature engineering to enhance predictive modeling through systematic classification of temperature data. Feature engineering plays a key role in an accurate predictive model by converting raw data into a structured dataset that more accurately captures the underlying dynamics of the study.

K-means clustering

K-means is a clustering algorithm used to partition a set of data points into K distinct non-overlapping subgroups (clusters), where each data point belongs to the cluster with the nearest mean. The algorithm aims to minimize the within-cluster variances (squared Euclidean distances), optimizing the homogeneity within each cluster [22].

Steps of K-means Clustering:

1. **Initialization**: Select K initial centroids randomly from the dataset.
2. **Assignment**: Assign each data point to the nearest centroid, based on the Euclidean distance.
3. **Update**: Recalculate the centroids as the mean of the data points assigned to each cluster.
4. **Iteration**: Repeat the assignment and update steps until the centroids no longer change significantly, indicating convergence.

Data Categorization Based on K-means Clustering:

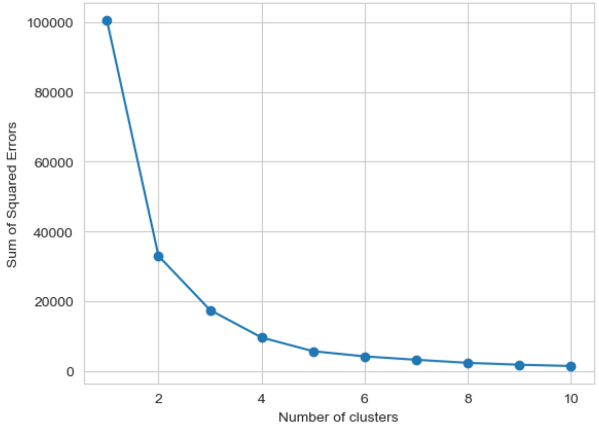


Figure 3.9: Distortion for different number of clusters

Using centroids generated from K-means clustering,the each centroid of clusters are effectively categorized into five distinct track conditions labels: *Very low, Low, Medium, Warm, High.*
Track temperature data is segmented into predefined bins that were determined as optimal divisions through K-means clustering. The decision to categorize track temperatures into five ranges using k-means clustering was based on the results of the elbow method analysis. This analysis indicated that four clusters optimally balance model simplicity, as demonstrated by a rapid decrease in SSE from 1 to 4-5 clusters, with minimal changes observed beyond five clusters [22].

Number of the clusters were selected for its efficiency in minimizing within-cluster variance SSE (distortion), a measure of how closely each data point in a cluster aligns with the centroid of that cluster, effectively distinguishing temperature regimes that impact racing performance. In addition, temperatures can be more easily interpreted by dividing them into clusters, which makes it easier to compare differences in lap times between different temperature circumstances and to build specialised racing strategy.
The application of K-means clustering offers a strong technique for categorizing weather conditions in a way that is both adaptive and based on the natural distribution of the data. This method ensures that each category is significant and reflective of different weather scenarios, boosting the accuracy and relevance of weather-related analyses in this study.

3.4 Model Development

3.4.1 Model Training and Testing

In machine learning, dividing dataset into training and testing sets is a fundamental practice for evaluating the performance of models. This split helps ensure that the model effectively applies what it has learned to new, unseen data, subsequently avoiding issues, such as overfitting where a model performs well on its training data but poorly on any new data.

For this project, 2019 to 2022 was used as training set, and 2023 used as a test set, this split was chosen based on several factors:

Using a time based split ensures that the training data (2019 to 2022) reflect past conditions and that the model's predictive accuracy is tested against the most current year data (2023) as unseen data. This approach handles a real-world scenario where models must perform well on new, unseen data reflective of the latest race conditions and regulations. Moreover, the use of the most recent year for testing aligns with the practical needs of Formula 1 teams and strategists, who require models that could adapt to and accurately predict performance under the newest, unseen conditions. Testing the model on 2023 data provides insights into its effectiveness and robustness when applied to actual race strategy planning and performance analysis.

3.4.2 Results and Evaluation

To evaluate the performance of the predictive models developed in this study, the Root Mean Squared Error (RMSE) was chosen as the primary metric.
RMSE is particularly suitable for regression models as it provides a straightforward interpretation of error magnitudes by measuring the average magnitude of the prediction errors. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.2}$$

where y_i is the response variable(actual lap times), $\hat{y}_i$ is the predicted(predicted lap time) value and n is the number of samples. One of the most advantages of RMSE is it calculates in the same units as the variable it is predicting. In this case, the response variable is *lap time*, which is measured in seconds. then subsequently RMSE also provides the results measured in seconds, it directly reflects the average magnitude of the errors in the predictions [25].

3.4.3 Cross-validation and Hyperparameter Tuning:

Cross Validation

Developing models that perform effectively on new, unseen data is crucial in predictive modeling. K-fold cross-validation is an enhanced statistical method used frequently to evaluate the performance of machine learning models. This technique is crucial to preventing overfitting and ensuring that the model performs consistently across different data subsets. The process involves:

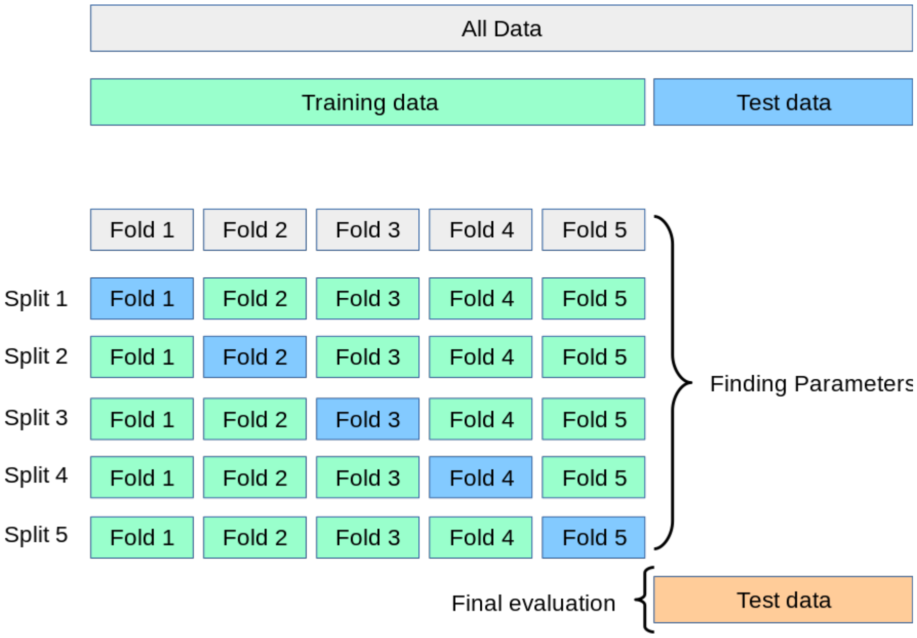


Figure 3.10: diagram of K-fold cross -validation

- **Model Training and Validation**: The validation process involves training the model on *k-1* folds of the data while using the remaining fold as the test set. This training and validation cycle is repeated *k* times, with each of the folds used exactly once as the validation data.

- **Performance Aggregation**: After training, the performance of the model is assessed in each iteration using a predefined metric. The results from each fold are then averaged to obtain a single estimation, which provides a comprehensive view of the model's performance [24].

Hyperparameter Tuning

Grid Search

Grid Search is a methodical technique for optimizing hyperparameters that finds the optimal combination of parameters by exhaustively investigating a predefined subset of hyperparameters. Using Grid Search in the context of XGBoost lap time prediction would enable a comprehensive analysis of important factors including tree depth, number of estimators, and learning rate. The optimal settings are crucial to minimize prediction errors.

Random Search

In contrast, Random Search provides a probabilistic approach to hyperparameter optimization. Instead of testing all possible combinations, Random Search samples parameter values from a defined statistical distribution for a fixed number of iterations. Random Search could significantly reduce the computational burden by focusing on randomly selected points in the parameter space, potentially discovering high performing configurations more efficiently than Grid Search. In this project, Randomized search is mainly used given the constraint of available hardware resources [23].

Chapter 4

Analysis and Evaluation:

This chapter aims to critically evaluate the success of the project in achieving its primary goal; developing a robust machine learning framework for predicting Formula 1 lap times, with an emphasis on diverse weather conditions.All model were trained using data from 2019 to 2022 and tested on 2023 data:

1. Linear Regression Model:
The Linear Regression model initially exhibited an RMSE of 4.735, which was reduced to a cross-validated mean RMSE of 3.313, this initial RMSE indicates its limitations in capturing the non linear relationships that are crucial for accurately modeling complex race dynamics. This initial result suggests potential weaknesses in handling nonlinear complexities such as Formula 1 racing environments. This highlights the requirement of choice a proper model given the characteristics of the data set.

2. Random Forest Regression:
This model demonstrated a significant improvement over the baseline, with an RMSE of 8.070 and a cross-validated mean RMSE of 1.737. The Random Forest efficiently managed the dataset's complexity by implicitly modeling nonlinearities, making it a robust choice for this application. Its strength lies in integrating multiple decision trees to reduce overfitting and increase predictive accuracy, crucial for dynamic environments like Formula 1 racing.

3. XGBoost:
Initially showed an RMSE of 6.224 with a cross-validated mean RMSE of 1.499, indicating good predictive capability but also highlighting the need for further optimization.

4. Optimized XGBoost with Random Search:
Optimization with Random Search showed the best performance between all tested models, with RMSE of 1.353. This strategic optimization highlights the importance of hyperparameter tuning in enhancing a model's predictive accuracy, particularly under varying weather conditions as seen in Formula 1 races.

4.0.1 Critical Evaluation and Limitations

The model training incorporated data, from the 2019 to 2022 Formula 1 seasons, including periods disrupted by the COVID-19 pandemic. This inclusion aimed to improve the robustness of the model under various conditions, but further analysis is needed to better incorporate and differentiate the effects of circuit and vehicle dynamics under weather conditions.

The project identified a critical insight: while weather conditions do affect lap times, their influence is less decisive than that of circuit and sector characteristics. These results point to a potential overestimation of the weather's impact on the initial project hypothesis. In addition, the models were also numerously influenced by other dynamic race elements, suggesting a need for a sophisticated approach in future studies to enhance the integration and impact of weather variables.

4.0.2 limitation

This section discusses the faced challenges and constraints encountered in this project, highlighting the challenges, and providing the way of enhancement:

High complexity of the lap time

Predicting lap time in Formula 1 involves the complex interaction of variables including the driver's skill, car performance, driver conditions, and investment of the each team, etc. The intrinsic probabilistic nature of the race environment, beyond the effect of various weather conditions and race accidents, adds more complexity to creating an accurate model.

Confidential data

The Formula 1 team exclusively stores their strategic information, including detailed telemetry data, tyre usage strategies, and aerodynamic adjustments. This classification limits access to data that would more comprehensively understand the factors affecting lap time, that also associated with the various weather featurs, this issue limits the depth of possible analysis in external research projects such as ours.

DRS, ERS and Fuel consumption

In addition, important features that significantly affect vehicle performance, such as fuel consumption, Energy Recovery System (ERS), and drag reduction system (DRS), are not disclosed. ERS and DRS are particularly important as they directly affect speed and efficiency, DRS can increase the speed of a car by about 15% over certain sections of the track, and ERS manages energy storage and distribution, which is critical for optimizing lap time. Due to the unavailability of such data, it was difficult to investigate the implications on race dynamics.

4.0.3 Future Directions and Open Problems

Given these results, future research should aim to enhance sensitivity of models to weather variables, by developing sophisticated models which accurately capture the complex effects of weather. This may include integrating relevant features such as pit stops and accident data, ensuring they are indirectly correlated with weather conditions.

This research establishes the foundation for future studies that could be refined and expanded with the current findings:

1. **Deepening Weather Impact Analysis**: Further research could investigate the implicit impacts of specific climatic conditions on particular segments of race tracks to gain a more localized understanding.

2. **Enhancement of Predictive Models**: There is potential for incorporating more complex machine learning models or deep learning models such as transformers that could more effectively capture the complex interactions of an expanded set of variables influencing lap times, and also efficiently capture the time-series data.

3. **Broader Application Scope**: this modeling approach could be adapted for use in other dynamic environments where environmental factors significantly influence outcomes, therefore expanding the range of applications for which the results could be used.

Chapter 5

Conclusion

Summary of Main Contributions and Achievements

This study established a comprehensive machine learning framework aimed at predicting Formula 1 lap times, especially focusing on the impact of diverse weather conditions on race outcomes. By utilizing the datasets from the 2019 to 2022 Formula 1 seasons, and implementing advanced analytical methodologies and models, this study expanded our understanding of environmental influences on race dynamics.

Evaluation of Project Status and Objectives

The project has achieved the primary aim that constructing a predictive model that accurately forecasts lap times; mainly considering various weather conditions. The results showed not the best predictive accuracy but also suggested positive further enhancement with a more sophisticated approach.
Furthermore, the project has deepened our understanding of the environmental impacts on racing dynamics which is the secondary objective of this project. Additionally, the results highlighted the substantial influence of non-weather factors, which have emerged as more significant than initially expected.

Concluding Remarks

The project achieved its aimed goals but has also established a robust foundation for research in dynamic sports analytics. By studying the complex factors affecting Formula 1 lap times, this dissertation contributes significantly to the wide field of performance analytics in dynamic fields, providing strategies for enhancing competitive performance under diverse environmental conditions.

Bibliography

[1] 2.3 - the simple linear regression model. Online Resource. Accessed: insert-date-here. URL: https://online.stat.psu.edu/stat462/node/93/.

[2] 5.3 - the multiple linear regression model. Online Resource. Accessed: insert-date-here. URL: https://online.stat.psu.edu/stat462/node/131/.

[3] Accelerating the fan experience. Accessed: 20 April 2024. URL: https://pages.awscloud.com/rs/112-TZM-766/images/AWS_Formula_1_eBook_Accelerating_the_Fan_Experience_Final.pdf.

[4] sklearn.preprocessing.robustscaler — scikit-learn 0.24.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html.

[5] Everything you need to know about f1, 2023. URL: https://www.formula1.com/en/latest/article/drivers-teams-cars-circuits-and-more-everything-you-need-to-know-about.7iQfL3Rivf1comzdqV5jwc.

[6] How f1 technology has supercharged the world, 2023. URL: https://www.formula1.com/en/latest/article/how-f1-technology-has-supercharged-the-world.6Gtk3hBxGyUGbNH0q8vDQK.

[7] What impact does high altitude have on an f1 car?, 2024. URL: https://www.mercedesamgf1.com/news/what-impact-does-high-altitude-have-on-an-f1-car.

[8] Ron Bekkerman. The present and the future of the kdd cup competition: an outsider's perspective, Aug 2015. URL: https://www.linkedin.com/pulse/present-future-kdd-cup-competition-outsiders-ron-bekkerman/.

[9] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, 2016. URL: https://books.google.co.uk/books?id=kOXDtAEACAAJ.

[10] Wikipedia Contributors. Atmospheric pressure, Feb 2019. URL: https://en.wikipedia.org/wiki/Atmospheric_pressure.

[11] R. Ponte Costa. COMS30035, Machine learning: Key ML concepts, 2022. Accessed: Apr. 25, 2024. [Online]. URL: https://uob-coms30035.github.io/RuiLectures/Lec2-handout.pdf.

[12] Rui Guo, Zhiqian Zhao, Tao Wang, Guangheng Liu, Jingyi Zhao, and Dianrong Gao. Degradation state recognition of piston pump based on iceemdan and xgboost. *Applied Sciences*, 10(18):6593, 2020.

[13] Justin Hynes. How the cold desert temperatures will impact the las vegas grand prix, Nov 2023. URL: https://www.redbull.com/us-en/theredbulletin/cold-temperature-f1-las-vegas-grand-prix.

[14] M Keertish Kumar and N Preethi. Formula one race analysis using machine learning. In *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022*, pages 533–540. Springer.

[15] Paul Keith. How does the weather impact f1 teams and drivers?, Mar 2024. URL: https://www.redbull.com/in-en/how-weather-impacts-f1-racing.

[16] Romain Mathon. How long does a pit stop in formula 1 last?, Sep 2023. URL: https://www.motorsinside.com/en/f1/news/30983-how-long-does-pit-stop-in-formula-1-last.html#:~:text=During%20a%20Formula%201%20race.

[17] T.M. Mitchell. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997. URL: https://books.google.co.uk/books?id=EoYBngEACAAJ.

[18] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

[19] Zhuofu Pan, Yalin Wang, Kai Wang, Hongtian Chen, Chunhua Yang, and Weihua Gui. Imputation of missing values in time series using an adaptive-learned median-filled deep autoencoder. *IEEE Transactions on Cybernetics*, 53(2):695–706, 2023. doi:10.1109/TCYB.2022.3167995.

[20] Anna-Lisa Paul and Robert J Ferl. The biology of low atmospheric pressure–implications for exploration mission design and advanced life support. *Gravitational and Space Biology*, 19(2):3–18, 2006.

[21] Pirelli. F1 tires, 2019. URL: https://www.pirelli.com/tires/en-us/motorsport/f1/tires.

[22] S. Raschka and V. Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2, 3rd Edition*. Expert insight. Packt Publishing, 2019. URL: https://books.google.co.uk/books?id=n1cjyAEACAAJ.

[23] scikit learn. 3.2. tuning the hyper-parameters of an estimator. URL: https://scikit-learn.org/stable/modules/grid_search.html#exhaustive-grid-search.

[24] SciKit-Learn. 3.1. cross-validation: evaluating estimator performance — scikit-learn 0.21.3 documentation, 2009. URL: https://scikit-learn.org/stable/modules/cross_validation.html.

[25] Scikit-Learn. 3.3 metrics and scoring: Quantifying the quality of predictions. Scikit-learn.org, 2013. URL: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics.

[26] News Team. How weather affects formula 1, Jan 2022. URL: https://news.williamhill.com/motor-racing/weather-effect-formula-one/.

[27] Wikipedia Contributors. Formula one tyres, Apr 2024. [Online; accessed 22-April-2024]. URL: https://en.wikipedia.org/wiki/Formula_One_tyres#cite_note-36.

[28] Wei Xie, Guanwen Huang, Wenju Fu, Bao Shu, Bobin Cui, Mengyuan Li, and Fan Yue. A quality control method based on improved iqr for estimating multi-gnss real-time satellite clock offset. *Measurement*, 201:111695, 2022.

Appendix A

Appendix

A.1 Appendix : Detailed DataFrame Information

```
RangeIndex: 82418 entries, 0 to 82417
Data columns (total 40 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
0   Time                82418 non-null  object
1   Driver              82418 non-null  object
2   DriverNumber        82418 non-null  int64
3   LapTime             80808 non-null  object
4   LapNumber           82418 non-null  float64
5   Stint               82418 non-null  float64
6   PitOutTime          2660 non-null   object
7   PitInTime           2758 non-null   object
8   Sector1Time         80613 non-null  object
9   Sector2Time         82252 non-null  object
10  Sector3Time         82186 non-null  object
11  Sector1SessionTime  80415 non-null  object
12  Sector2SessionTime  82252 non-null  object
13  Sector3SessionTime  82186 non-null  object
14  SpeedI1             70615 non-null  float64
15  SpeedI2             82206 non-null  float64
16  SpeedFL             79518 non-null  float64
17  SpeedST             75935 non-null  float64
18  IsPersonalBest      82322 non-null  object
19  Compound            82418 non-null  object
20  TyreLife            82418 non-null  float64
21  FreshTyre           82418 non-null  bool
22  Team                82418 non-null  object
23  LapStartTime        82418 non-null  object
24  LapStartDate        0 non-null      float64
25  TrackStatus         82322 non-null  float64
26  Position            82322 non-null  float64
27  Deleted             82418 non-null  bool
28  DeletedReason       558 non-null    object
29  FastF1Generated     82418 non-null  bool
30  IsAccurate          82418 non-null  bool
31  AirTemp             82418 non-null  float64
32  Humidity            82418 non-null  float64
33  Pressure            82418 non-null  float64
34  Rainfall            82418 non-null  bool
35  TrackTemp           82418 non-null  float64
36  WindDirection       82418 non-null  int64
```

```
37  WindSpeed           82418 non-null  float64
38  Circuit             82418 non-null  object
39  Year                82418 non-null  int64
dtypes: bool(5), float64(15), int64(3), object(17)
```