Aston Business School

Creating Machine Learning Models to Predict Formula 1 Tyre Life

Submitted in partial fulfilment of the requirements for the degree of MSc Business Analytics at Aston University, Birmingham

Vedant Modani (210208830)

September, 2023

Declaration

I declare that I have personally prepared this report and that it has not in whole or in part been submitted for any other degree or qualification. Nor has it appeared in whole or in part in any textbook, journal or any other document previously published or produced for any purpose. The work described here is my own, carried out personally unless otherwise stated. All sources of information, including quotations, are acknowledged by means of reference, both in the final reference section and at the point where they occur in the text.

Acknowledgements

I would like to express my gratitude to my supervisor Mr. Ammar Al-Bazi for his guidance and encouragement for the duration of the project. I am grateful to the faculty and my lecturers for MSc Business Analytics for providing me with the knowledge and skills that have made me capable of completing this project.

I am grateful to the developers of FastF1 library and Ergast API for making F1 data publically available and accessible and the wonderful community members that make F1 data analysis interesting and insightful.

I am grateful to my family and friends for encouraging me throughout the duration of my course. Without their support this project would not have been what it is. Their guidance and endless motivation have helped me complete the course and take maximum advantage of my opportunities.

Abstract

The aim of the project is to predict Formula One tyre life with publicly available data. With F1 being a very technical sport, the core of the competition is in the technical data. Much of the live data is proprietary to the teams as they collect it and analyse it to create better and faster cars. Some live data is available to the public to make the viewing experience better and keeping the fans engaged. Such data is available to access and analyse through open-source APIs and libraries. The project was completed in python using common libraries such as pandas, scikit-learn, and seaborn.

The data consisted of records of race stints from the 2019 F1 season till the summer break of the 2023 season. 3 data sources were used for different parameters – Ergast API, FastF1 library, and Pirelli Weekend Preview Press Releases. The data was then joined and prepared using SQL before loading it to the Python script for the machine learning models.

The project uses this data to create machine learning models that predict tyre life using parameters that represent tyre, weather, and circuit characteristics. Two tree-based ensemble algorithms were considered for the project – Random Forest and Gradient Boosting Trees. The two models were created after analysing and cross validating several hyperparameter settings using Grid Search CV. The Random Forest Regressor and Gradient Boosting Regressor had different results for the data.

Gradient Boosting Regressor was a more accurate model with a RMSE of 6.11 as opposed to RMSE of Random Forest at 8.01. The Gradient Boosting Regressor also had a better fit with a R2 score of 0.67, whereas Random Forest Regressor's R2 score was 0.44. Although while Gradient Boosting Regressor was a better model, it had a higher overfitting than the Random Forest Regressor.

The models were then used to make predictions on new records that were collected from the same sources, but after the models were trained. The records were selected to simulate real life situations to make predictions. These predictions were used to justify the use of the models in a race scenario.

Key Words: Supervised Machine Learning Models, Ensemble Methods, Hyperparameter Tuning, Overfitting, FastF1, Ergast API, Formula 1, Pirelli, Motorsports, F1 Tyre Life, F1 Race Strategy, F1 Tyre Compounds, Circuit and Weather Characteristics

Table of Contents

List of Tables

List of Figures

List of Appendices

1. Introduction

Formula One is one of the most popular sports in the world. With hundreds of thousands of fans attending each race and millions watching from home, Formula One has captivated its audience for decades. Fans flock to F1 for the speed, skill, technology, jeopardy, and strategic planning involved in each race weekend. Since 1950, when the championship was formally established, it has been dubbed the pinnacle of motorsport and has been the most sought-after championship for drivers and teams to participate in and win. Every big automotive company has at some stage participated in F1 in some capacity. Brands like Ferrari and McLaren are born out of F1 and use the automotive business as a secondary operation. These teams invest heavily in the development of the cars and the automotive technologies of required to compete and win and create some of the most technologically advanced race cars in the world. The fast and precise nature of the sport with minimum margin for error, real-time analytics has emerged to the forefront of innovation.

1.1 Background for the Problem

This section will talk about the historical context of the problem, addressing the issues and regulations that face the tyres in Formula One and the root of the problem, with an example. The next section will discuss the problem itself in more detail and explain why this study is important to solve the problem.

Over the last 73 years, Formula 1 has evolved, not just in the technology used but also in the way the drivers race. No longer is it about the fastest car or the most skilful driver, real time strategy decisions play a key role in the success of a team. Often, strategy is the only point of difference between teammates. In the hard-core fan clubs, mathematicians and strategists are celebrated equally with the drivers, for pulling off great strategies and victories. This new era of strategy driven racing has transformed the way even fans view the races.

In 2017, F1 was acquired by Liberty Media, and they have worked hard to improve the access fans have to the drivers and teams. This has coincided with the rise of social media as well. But most importantly, F1 has worked hard engage fans with the technical aspects of the sport. F1, in partnership with AWS (Amazon Web Services) has made data and insights available to fans during the live broadcast and after the races. This partnership has provided live insights for the viewers of the grands prix to engage the audience more and make it more exciting for the viewers. The insights display race strategy and performance for the drivers and teams. Most of the graphics compare the performance of 2 competing drivers. (F1, 2023)

With publicly available data and widespread knowledge of tools like Python, there has been a growth in online communities that study, interpret, and share race strategies and driver performance to the tiniest detail. Tools like Ergast API, FastF1 Python library and online communities like r/F1Technical on Reddit have grown exponentially in recent years. This is where part of the motivation for this study is based. To participate in a community that uses open-source data and cutting-edge tools to produce informative insights about a major sport. These predictions can be used in healthy discussions about the sport and about each driver's strategy whenever fans gather and watch or talk about F1.

When we talk about strategy in F1, we primarily talk about pitstops. Historically at pitstops, teams would refuel the car and change the tyre. This created a strategic decision on 2 fronts. But refuelling was banned for the 2010 season, and since then the tyres have been the only aspect of the pitstop strategy decision. The tyres are arguably the most important part of a race car as they are the only parts that are in contact with the road. This becomes especially important when the cars are constantly doing speeds of over 200mph and pulling 4-5Gs in corners.

Since 2011, Pirelli & C. S.p.A., the Italian tyre manufacturer has been Formula 1's sole tyre supplier. This means, all the teams use identical tyres produced by the same company. It was not always the case, and in the past, multiple companies produced tyres and supplied to different teams. But since 2007, first Bridgestone and now Pirelli is the sole supplier. This means that the tyres are some of the few parts that are common across all teams, making them even important. Even if a team has a slower engine or weaker downforce package, if they unlock the tyre performance better than their rivals, they can be competitive.

To conclude and fully establish the context of the problem it is important to understand the conditions and regulations regarding the F1 tyres. Pirelli creates multiple compounds of tyres for the F1 season. Each compound is a unique chemical construction formula. The various compounds mean that the tyres are on a spectrum. The main characteristics of this spectrum classify the tyres on their wear rate and grip. On each race weekend, Pirelli assigns 3 dry weather tyres, aka slicks. These are tyres that the teams run when there is no or little rain. Each driver is required to use at least 2 of the 3 compounds during a race. This regulation enforces the drivers to make at least one pitstop during a race. There are 2 wet weather compounds as well, but they are only reserved for when there is rain and the race is declared 'wet', meaning the 2-compound regulation is overruled.

On a race weekend the 3 available compounds are classifies as Soft-Medium-Hard, based on their characteristics. The Soft compound is the most grippiest and fastest but wear quickly. On the other hand, the Hard compound is the least grippiest and slowest, but does not wear as

quickly. In fact, there are multiple occasions when a driver ran the Hard compound for almost the entire duration of the race, only having to pit to meet the regulations.

The 2023 Azerbaijan Grand Prix can be used an example to establish the context. Held on 30th April 2023, it was the 4th round of the 2023 season hosted at the Baku Street Circuit in Baku, Azerbaijan. Before the race, Pirelli announced the possible race strategies as shown in the Figure 1. (Pirelli, 2023)

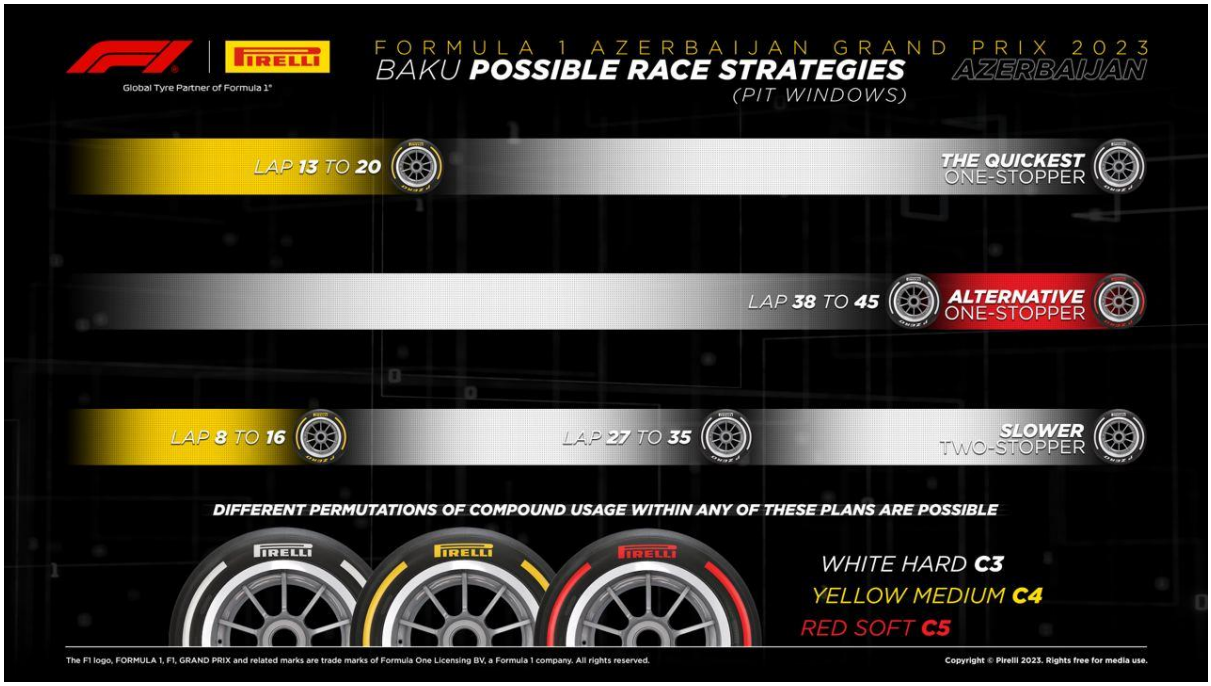


Figure 1 | Pirelli Azerbaijan Grand Prix 2023 Possible Strategies

These strategies are just a suggestion and can be changed as the race progresses and situations arise. Often the drivers start on either one of these strategies and then react to the race. Figure 2 shows the strategies used by the drivers during the race. As is clear most drivers used the 'Quickest' strategy which says start with MEDIUM tyres to about lap 15-20 and then pit for HARD tyres to use them till the end of the race. There was a safety car intervention at lap 12 and it allowed many drivers to change the tyres. The driver GAS (Gasly) had just pitted, for HARD tyres, before the Safety Car period on lap 6 and therefore didn't pit during. He pitted much later around lap 25 for another set of HARD tyres– basically adopting the 3rd strategy.

OCO (Ocon) and (HUL) Hulkenberg adopted the 2nd possible strategy where they stared on HARD compound and went long, till the last few laps of the race. They were able to extend their HARD tyres because of the Safety Car and then pitted on the last lap to meet the regulations. (FastF1, 2023)

This example shows the wide range of possible strategies and why knowing the tyre life is important.

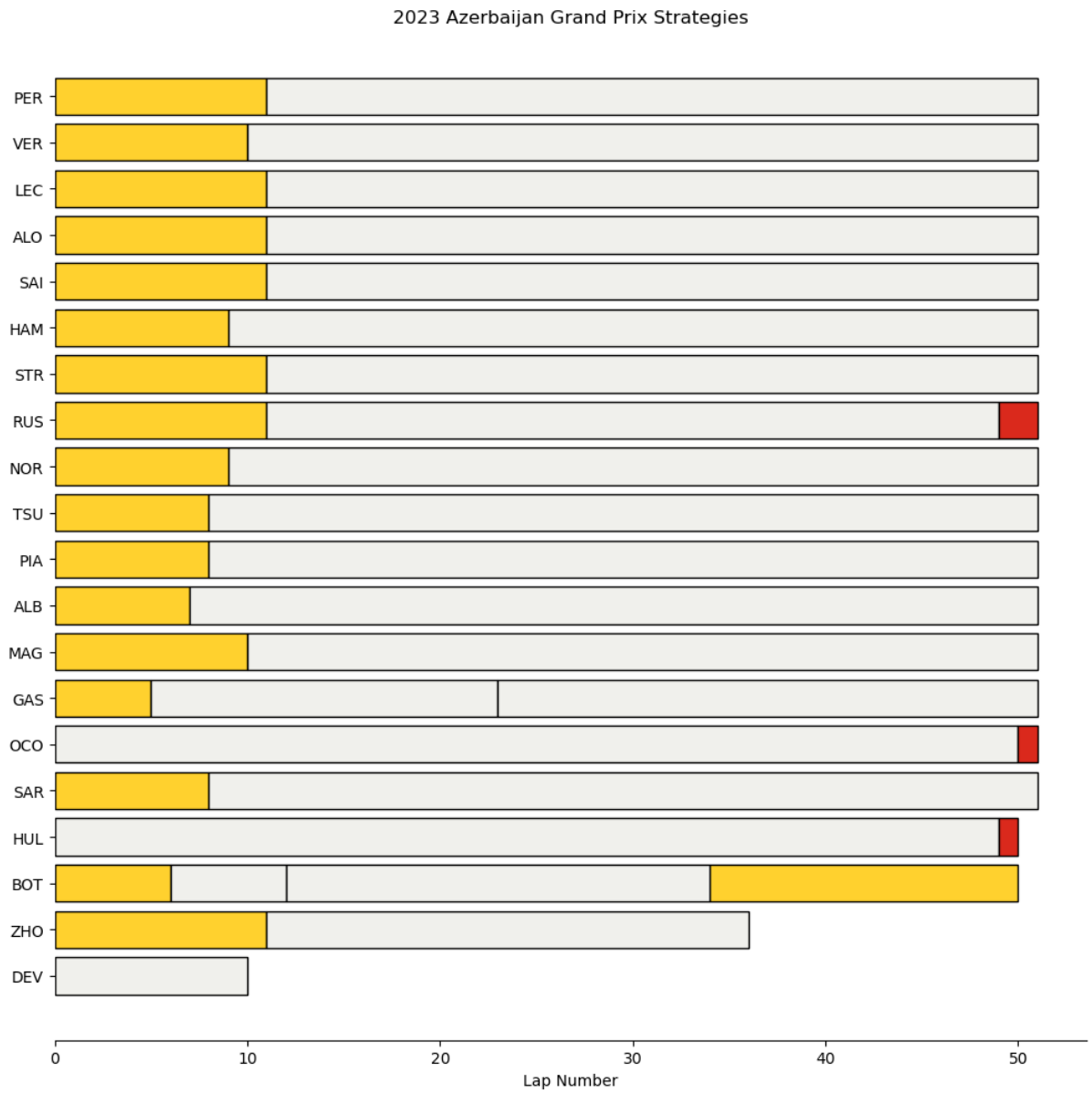


Figure 2 | Azerbaijan Grand Prix 2023 Actual Race Strategies

1.2 Problem Definition

The problem exists because of need for the Formula One teams to choose the tyre to be utilised during a stint in a Grand Prix race. Tyres are one of the most important parts of a F1 car, even when it does not seem like it from. Since they are the only part in contact with the surface, their structural integrity is of utmost importance during a Grand Prix. This means that the useable life of an F1 tyre is one of the prime metrics the teams use in working out a race strategy, that defines the number of and timing of pitstops to be made during the race and which tyres to use.

Tyre life is the number of laps a tyre can run while giving good performance and without increasing the risk factor. As the tyres run longer the degradation increases and performance decreases. It can be considered a function of tyre, track, weather, and car characteristics.

Of course, F1 teams conduct their own simulations and during practice sessions they even test out the tyres to plan a strategy. But the simulations cannot possibly count for each random possibility. Additionally, during the practice sessions at each Grand Prix weekend, the teams are doing a lot more than just modelling a strategy. They are testing new parts and different set-ups to find performance. They also must think about the qualifying and evaluate the next steps in the development cycles. This means that while the practice sessions are representative of the actual performance and conditions on the race day, it is still essential to study the pure racing conditions and pace from previous years. The combination of these 2 aspects means that a tyre life predictor using historical data can be effective. This study tackles these existing issues in a new approach to see if it is possible to predict tyre life using existing historical data.

1.3 Research Questions

1. **RQ1:** Can machine learning techniques accurately predict the life of F1 tyres in a race using open-source data?
2. **RQ2:** What are the most important parameters influencing the running life of F1 tyres?
3. **RQ3:** Is it possible to create a model that enables fans to have a much more knowledgeable discussion about the tyre strategy?

1.4 Aim of the Study

The study aims to identify a Machine Learning technique that is effective and accurate in predicting the life of F1 tyres to provide the teams with an additional source of information. The model should work effectively using open-source data to provide the fans with new information to help healthy discussion about the sport.

1.5 Objectives

- To review the existing studies on F1 tyres and pit-stop strategies in order to define the scope of this work.
- To outline the context under which the results from this study become applicable in the real world.

- To collect data and identify the parameters required for creating an accurate and representative model.
- To create a Machine Learning Model that predicts F1 tyre life from open-source data.
- To verify and validate the results of the predictions, including prediction of different scenarios using the model.
- To identify the weaknesses and suggest further actions that might lead to an improved model.

1.6 Research Methodology

The tyre life prediction model will be created to predict how long each compound of tyre will run in given circumstances. These predictions will be based on existing data of previous races.

The data will be collected from 3 sources and joined together to create the full dataset required for the study. The live data for the weather and race conditions parameters will be collected using an existing Python library called FastF1. This library sources data from open-source information available on F1's official website and a reliable public source called Ergast Developer API. The Ergast Api contains historical data from official sources collected in one location and will be used to retrieve driver, team, and event information. The parameters relating to circuit characteristics will be mined from Pirelli weekend preview documents.

Tyre life will be defined as a target variable. Once we conduct parameter importance and correlation analysis, parameters with high multi-collinearity will be eliminated. Next, outliers will have to be eliminated to ensure accuracy in predictions. Once the data is ready, it will be split into train and test data and a baseline will be established.

A baseline model will be created which acts as a score to beat for any machine learning models. The baseline is important to evaluate the more complex machine learning algorithms. Then the models will be trained and evaluated with several hyperparameter settings. The hyperparameter settings will be adjusted to find the optimal solution. Multiple metrics, such as Root Mean Square Error and R Squared goodness of fit will be used to evaluate the models and compare their performance.

Once the optimal model is identified it will have to be validated. Several situations will be created with real data that has not been fed to the models. The final models will be then used to predict the tyre life in each of those situations. The predictions from those simulations will be helpful in validating the models in an environment away from the training and testing data. This will be a good chance to see the actual performance of the models.

1.7 Research Deliverables

- A comprehensive dataset structure useful for predicting F1 tyre life.
- Models that predict F1 tyre life with the given data.
- Comparison of the ML algorithms and hyperparameter settings.
- Predictions on situations simulated after finalising the models.
- Identify the limitations of the model and the dataset.
- Identify possible further improvements.

1.8 Research Scope

In 2019, Pirelli announced a new way to name the different tyre compounds. There were 5 dry weather compounds in total but only 3 were selected for each race weekend. Rather than having individual names for all 5, the 3 selected for the race were designated SOFT, MEDIUM, HARD depending on their composition relative to each other. This means that the compounds were not named for their own performance, but their performance relative to the others. This rule was made to make it easier to interpret the data and make viewing better. The study will focus on the 4.5 years since this rule was introduced. This is because the data from before 2019 will not be relevant and we will end up with a badly trained model. The SOFT, MEDIUM, and HARD compounds of before 2019 have specific characteristics and do not incorporate the characteristics from 2019 onwards.

2. Literature Review

There is very little research conducted in the field of machine learning in motorsport for race strategy. Historically all the research is towards the mechanics and aerodynamics of the car and its components. And while the teams, suppliers and manufacturers do conduct their own research on strategy, it's all proprietary and nothing is published. But in recent years as machine learning becomes more mainstream and the data becomes openly available, the number is gradually increasing.

2.1 Previous Studies on Machine Learning in Formula 1 tyre life

The previous machine learning research in Formula 1 can be grouped by what they were predicting, into a) predicting the results and championship, and b) predicting the best race strategy. While this project might fall into the second group, that talks about race strategy, it is important to review the literature in the first group as well. Several papers that fall in the first group, predicting race and championships, can help in understanding how various parameters interact with each other while predicting performance in F1.

2.1.1 Predicting F1 Races and Championships

Predicting the championship looks like the more straightforward of the two groups in terms of the technicalities of the models. There is less importance on the lap-by-lap technical aspects, but more on the overall winners and results of each grand prix. The understanding of it is to use the historical results to predict the future races. The approach can be altered with different samples, machine learning models, methods, parameters, and target variables.

Sicoie (2022) developed a model that predicts the race winners and championship standings for the Formula 1 season. The project was developed to use machine learning technologies in Formula 1 predictive models to use modern techniques and leverages the recent push of Formula 1 into analytics. The widely used Ergast API was used to bring in majority of the historic data along with some other sources that help in obtaining more parameters. Supervised regression methods were used to create the model, including Support Vector Regression and Gradient Boosting Regression. The models were developed to predict individual race results and the entire championship as a consolidated result of the race predictions. The author displays a good and insightful use of the data sources and parameter selection, but the project has low accuracy across the different methods. A reoccurring case highlighted was when a driver changes to a different team recently and the training set has very little data for that new combination. This causes the model to be trained on historic information and reduces its accuracy. This also leads to the prediction to primarily be based on a driver – team combination, which might increase correlation. Whereas Rana, et al (2021)

created a model to forecast the results of the Formula One Driver's Championship which uses Lasso Penalised Regression. This comes as an exploration into predicting the championship results using a different tool than earlier studies. Several models including Ridge Penalised Linear Regression and K Nearest Neighbours Regression were considered and tested but, Lasso Penalised Regression gave the best results, followed by Support Vector Machines. Similarly, several algorithms like Neural Networks, Random Forest Regressor and Gradient Boost Regressor were tested, with Random Forest giving the best results.

In Rana et al (2021) the data ranged from right at the beginning of F1 World Championship in 1950. The parameters used in the data included a wide range of race, circuit, weather, driver, team, and even qualifying characteristics. The predictions were accurate in the overall standings at the end of the season but failed to reflect the same accuracy with the individual results and points collected. As the goal of the project was to predict the championship winner, it was a successful result. This also concluded that predicting a driver's championship is more difficult than predicting a constructor's championship and that the car characteristics play a large role in the race results. The study did not utilise weather or circuit characteristics properly. While it is not a major factor in predicting a championship with multiple races, as the law of averages kicks in, it is required to predict individual races and events in each race. One reason for this might be the unavailability of data as far back as 1950. A shorter time span might have provided more parameters as the data recording and storing in the publicly accessible resources has increased manyfold.

Further, Stoppels (2017) created a model to use Artificial Neural Networks to predict the race results of Formula One races. The aim of the research is to display the capabilities of ANN and highlight its uses in the field of Motorsports. The aim is to predict a race result before the start of a race. This means no real-time variations are considered and the model can be developed in a very simple manner. It also means that the number of features is reduced. The research, while insightful in the use of machine learning, lacks in the sampling and feature selection. There are merely 8 features, and the data considers only 4 of the 16 drivers that started all the races and 24 drivers that started at least 1 race. Additionally, as it only considers the 2016 Formula One season, the 21 races in that season mean that the data is only 84 records. They did increase the number of records by doubling each record with an adjustment to increase the variability, but it still lacks the volume required for a good machine learning model.

2.1.2 Predicting F1 Tyre Life and Strategy

The models used for simulating F1 races and predicting race strategies are much more complex and require advanced machine learning techniques, increased domain knowledge, and specialised parameters.

Piccolomini, Evangelista and Rondelli (2022) proposed a model to predict the tyre strategy in Formula 1 as it is a critical decision point in the performance of a Formula 1 car and driver. The FastF1 Python library was used to gather and transform data and parameter tuning. The study focuses on Recurrent Neural Networks. Long Short Term Memory (SLTM) and Gated Recurrent Unit (GRU) being the main focus of the methodology. The project is a model that aims to provide accurate prediction for the choice of tyre in a Formula 1 race, at the start and at every subsequent pitstop. This allows the teams to implement a preferred strategy. The models created presented a great future for the use of deep neural networks in the Formula 1 tyre choice decision making process. The GRU model outperformed the blind classifier by 25% but the other models were only slightly better than the baseline. This presents with an opportunity to tweak the parameters deployed and structure the focus on tyre performance rather than tyre choice. Predicting tyre choice directly leaves a potential weak link as the choices are limited and the conditions are dynamic.

AbdulRazzaq, Fadhel, and Al-Shamma (2021) developed a tyre predictor that uses parallel hardware architecture. This allows for more accurate performance simulations and better predictions. No other study has tried to replicate the tyre wear like they have. It is truly a step in the direction of performance dependent predictions. Rather than predictions based on previous static data. Neural Networks are used to develop a tyre degradation model that allows the motorsport teams to evaluate the wear and tear of the tyres during a stint. Monitoring tyre conditions is likely to help take better strategy calls. Alternatively, Carrasco Heine, and Thraves (2021) developed a model that uses Dynamic Programming and Stochastic Dynamic Programming to develop multiple strategies that show different starting tyres and different yellow flag laps. The study aims to optimise the pit stop strategy before the start of the race to help choose a starting tyre as well. This means that the predictor is limited in its ability to run with changing conditions. This requires the teams to choose the optimal strategy from the predictor and then react to any unknowns. The study develops a model that incorporates random events like safety cars or rain while comparing race strategies. The model does not use real world F1 data, makes it usable in different forms of motorsport by inputting the relevant parameters. But at the same time, it becomes ill-trained to predict Formula 1 strategies as it does not know the conditions and performance of past F1 races and drivers.

Piccinotti (2021) developed a model to identify a race strategy that allows the Formula 1 teams to make the right decision and maximise performance. The problem identified by the study is to create a model that evaluates at each lap weather a pitstop is needed and which tyre to put onto the car if indeed the pitstop is needed. Monte Carlo Tree Search was utilised to make an open loop planning model for identifying race strategies. The study uses multiple other methods including Markov Decision Planning for Stochastic Control and Q-learning. The study was based on a single driver in a specific time frame, reducing the driver – team combinations the model must learn to 1. This allows the training to happen on other parameters like circuit characteristics, weather conditions and the random incidents. The model created has a good accuracy and applicability into the real environment but with a few gaps. As with several other models created by different studies, the variable and dynamic instances of Formula 1 could not be added to the model. Aspects like traffic, overtakes, retirements, penalties are all not considered. Contrastingly, Sulsters (2018) develops a model that allows Formula 1 teams to identify the optimal strategy from a set of options. Other similar projects focus on individual cars, but this model allows the teams to focus on both cars and optimise the result for the team. The model uses discrete event simulation to process lap times and imitates most on track events. Accounting for fuel, tyre wear and overtakes allows for a better simulation and increases the chances for the simulation to be accurate. This model considers the previous retirements, starting grids, positions gained or lost through the course of the season, average pitstop duration, and more parameters most studies overlook. Training a model with these parameters allows the model to handle some variable events. The simulation model predicts the strategy almost accurately as previous works but is worse in predicting lap-times. A pre-race simulation model does not allow for real time contingencies even if some are planned into the model and the author accepts as much. The model is good enough to predict the ideal strategy pre-race with certain input parameters that mimic likely events but can't mimic real world scenarios. The sample utilised is very small as data of only the 2016 season is taken. Dividing a single season into train and test sets does not give enough data to form an accurate model. Even from this limited dataset, rain affected races were removed. This just goes on to provide a slightly untrained model for universal usage.

Heilmeier, Graf, and Lienkamp (2018) developed a race simulator with the purpose of creating a tool that simulates the race and allows the teams to prepare for the strategies in a short time, which is useful during a race weekend. The strategy team needs to run simulations to prepare for and respond to events during a live race. This means that the simulator not only needs to be accurate, but also very fast in giving a result. This project aimed to focus more on speed as a slight deviation in accuracy can be handled with manual inputs and human intuition. Additionally, as the race situation may change several times accuracy becomes a second

priority anyway. The simulation was discretised into laps as each lap has the same characteristics and the race runs for a specified number of laps, but the lap times changes as the characteristics change. The laptimes were calculated in a specific manner with additional complexity with each step. Starting with a base laptime, factors like driver and car ability, fuel mass, tire degradation and starting grid were added to come to a realistic laptime without overtaking and then with overtaking. This allows the laptimes to be broken down into multiple factors and add to the accuracy of the model. The model is very accurate, with the delta to actual race time being just 1 second at an average. Since the race was discretised into laps, it comes down to around 2/100 of a second each lap. The authors used only 1 F1 Grand Prix to evaluate the model and while it was very accurate, the sample size is very small. Additionally, while several factors were considered, circuit and weather characteristics were not. This means for every change in racing conditions, the simulator can either become redundant or requires a lot of changes in the base inputs. Furthermore, Heilmeier, Thomaser, Graf, and Betz (2020) developed a method to automate the strategy aspect of a Formula 1 race team. They call this the Virtual Strategy Engineer (VSE) and it uses two artificial neural networks. Eash solving one decision of the strategy, pit-stop lap and tyre compound choice. The virtual strategy engineer solves the problem of making a strategy decision and allows the teams to focus their resources on other aspects of the race car. As the decision involves multiple variables in a dynamic environment, an automated strategy engineer is much faster in identifying the best strategy any given moment. The supervised neural network architectures solve one part of the decision but must work one after the other. The first neural network determines whether a pit-stop should be made and only then the second neural network kicks in to decide which tyre compound to choose. Therefore, the model is a classification problem. The virtual strategy engineer was trained on real F1 data available, allowing to learn from representative information and be as accurate as possible. Race strategies decisions made by the virtual strategy engineer were accurate with impressive scores in the confusion matrix. The authors point out that the structure of the two neural networks forces the tyre decision to be dependent on the pitstop decision. The tyre choice would not influence the pitstop decision. This influences how the tyres chosen as the full potential performance and tyre life is not considered before making the pitstop decision.

Heilmeier, Graf, Betz, and Lienkamp (2020) worked on creating a race simulator before the Virtual Strategy Engineer. The aim of the study was to develop a method that accounts for random events like accidents and safety cars in a race while planning out a race simulation. The probabilities of these random events vary from race to race, depending on circuit and weather conditions. Basic race simulations can result in accurate lap-times and performance. But any random event can have a big impact on the real performance. Therefore, a model that

considers the probabilistic effects is important in an accurate simulator. Lap-wise discretized race simulation was taken as a base to split the entire simulation into equal individual blocks. Multiple approaches were considered to deal with probabilistic events, but each had a limitation. Hard coding 'what-if' scenarios meant a combination of probabilistic events could not be easily simulated. Another approach was to use full factorial design to discretize the random variables and create a simulation for every combination of the probabilistic events. This approach allowed for better sampling but suffered suffers from the issue of dimensionality and increased computation time. Monte Carlo Simulation turned to be the best solution for tackling combinations of probabilistic events in a race simulation in circuit motorsports. This paper allows for more realistic race simulations with considerations to a wide range of probabilistic effects. Starting position and performance, accidents, mechanical failures, full course yellows and safety cars being some of the major ones. The results can help a race engineer to a more reliable simulation model. The results show that the accuracy of the models increase in accordance with the law of large numbers. There are some inaccuracies acknowledged by the authors regarding the real-world processes that could not be simulated. Similarly, some events cannot be planned for in simulations.

Heilmeier (2022) compiled his earlier research to develop a simulation that aids the strategy decisions during a Formula One race. The project is aimed at finding the research gap that exists between open research and the multi personnel strategy departments of the motorsport teams. Accordingly, the model is desired to take low parameterisation computational effort. Furthermore, there is additional motivation for finding a direction for future research in the field and related areas like autonomous racing. Since the strategy depends on the race simulation which in turn depends on the discretised lap-times, the project had 4 defined steps. Racing Line Generation, which helps in simulating the lap-times, followed by Lap Time and Race Simulation. Last comes the automation of strategy decisions as it requires the complete race simulation to work. The simulation and strategy decision predictor are aimed at finding a model that allows for parameters be determined with little knowledge of the exact values of driver, circuit, vehicle and track characteristics. Additionally, the project aspires to find a way of using the strategies of opposition drivers to benefit the accuracy of objectively evaluating the specified driver's own strategy decisions.

2.1.3 Other Motorsport

Apart from Formula 1, other motorsport series such as Formula E and IndyCar have also attracted some research. This section talks about all of them, without grouping them by the target parameters.

Energy management, rather than tyre management is the name of the game in Formula E. The technology in Formula E cars allows for regeneration and storage of electricity possible as the race progresses. Managing power output and regeneration during the race is the challenge of Energy Strategy Management. Liu and Fotouhi (2020) have shown that the Monte Carlo tree search method is applicable for different motorsports in different capacities. They chose Formula E and its strategy as a topic and have been one of the first to do so. This opens new avenues for research into Formula E strategy and Machine Learning. Artificial Neural Networks were used to predict the car performance. And separating the different performance parameters gave better results than using a single ANN. The model works out one lap at a time to give results based on multiple situations with each lap being different. Most caveats and random events of Formula 1 racing also exist in Formula E. Such as safety car and weather, which are included in the model. A major addition is the implementation of the energy boosts available to the drivers. These also form a part of strategy decision making. The model accounts for it and develops a strategy that includes the laps on which to implement them. Monte Carlo tree search was used to develop a model that predicts a strategy pre-race. It can also be used during the race to factor for any of the random events and changes in conditions, including sub-optimal driving where energy is overconsumed. But even though tyres are not a major decision point in formula E, they still affect the performance of the cars. The study does not include tyre wear or performance as part of the models. Track conditions and race positions are also not considered. Further, Liu, Fotouhi, and Auger (2022) developed an improved model on top of the Monte Carlo tree search method using Bivariate Gaussian distribution. The aim of the project was to improve on the strategy prediction, particularly the multi–lap runs. The earlier MCTS model often generated sub-optimal results with a large variation. That needed to be solved to create a more reliable model. The races were split into multi layered decision making processes like some other motorsport strategy models. BGR improves MCTR's performance with each further expansion step. And Proximal Policy Optimisation was used in interaction with MCTS to improve race time simulation and consistency reducing the variance by 95%.

Peng, et at (2021) created a model to forecast the rank position of race cars in a time series context for the IndyCar. The aim of the project is to create a solution which can conduct probabilistic forecasting and model the pitstops and race positions in car racing. The foundation is to implement model decomposition to use sub-models based on cause-and-effect factors of the position rank, hoping to train the model for the random events efficiently and effectively. As with any good motorsport forecasting model, the biggest aim was to account for the extreme events and pitstops in the most accurate way possible. By using a unique event like the IndyCar Indy500, the model explored some niche factors that are not present in Formula 1, but the basis of a motorsport event and related characteristics are

common across the board. The rank position forecasting was first tested using several models such as CurRank, Random Forest, SVM, DeepAR, DeepState, and XGBoost before implementing their own model. But the best results were reached by suing the new model they created, RankNet. The reason for creating RankNet was based on the inability of the state-of-the-art established models to perform well on a global dependence structure. RankNet is a combination of encoder-decoder network and a separate multilayer perceptron network. It can deliver probabilistic forecasting and model the pitstops and rank position in car racing. Additionally, Tulabandhula and Rudin (2014) set out for designing a prediction and decision system that can be used in real-time during motorsport races. The research was one of the first ones conducted for prediction and decisions of a race strategy. With little previous research, a lack of domain knowledge disrupted the research at first. Therefore, they emphasise that domain knowledge is key in a field like this. Statistical modelling techniques alone cannot provide an accurate model in complex sports analytics. The problem is the need for a fast, efficient and accurate model that helps in strategy decisions. As NASCAR is a virtually domestic series, the parameters included regions like Southern United States and Northern United States along with other circuit characteristics, in addition to the usual racing parameters like circuit and weather characteristics, driver and car details, etc. Along with Random Forest Regression and two baselines, ridge regression, LASSO (least absolute shrinkage and selection operator) and support vector regression (SVR) with a linear kernel were the machine learning techniques tested in the experiment. SVR, Ridge Regression and LASSO were the best performing models in that order, with Random Forest performing better than the baselines as expected.

2.2 Research Gap

From the numerous previous research, Regression, Neural Networks and Reinforcement Learning are the main methods being used for predictive problems in F1. Additionally, it is evident that there is no publicly available study on the life of the tyres. By focussing explicitly on the tyre-life, the research and predictions will be on a completely unique aspect of the sport. The closest matching model is one created by Piccolomini, Evangelista and Rondelli (2022), where their setup and experiments were very straightforward and dependent mostly on training and testing machine learning models, including Deep and Recurrent Neural Networks. The use of the FastF1 Python library allowed the extraction and transformation of accurate data in an easy yet expansive manner. The data available was more than enough, allowing for accurate predictor selection and building an efficient model. With insight from Piccinotti (2021) a better understanding of the advanced technologies utilised for race simulation can be obtained, such as Monte Carlo Method and Markov Decision Process. And finally, the extensive models in Heilmeier (2022) break the strategy decision operation into simpler parts

which can be used as a good basis to create a model that can be built upon in the future and contribute to a more complete strategy identifier.

Only Sicoie (2022) compared ensemble methods in a motorsport context but that was for predicting the results for each grand prix. Therefore, using this approach, creating a tyre life predictor model shall be the specific research gap this dissertation will focus on.

Table 1 | Details of parameters and models in existing research

Author and Year	Parameters	Type of Modelling	Focus
Horatiu Sicoie **January 2022**	Weather and Circuit Characteristics, Driver and Constructor Details, Tyre Type, Lap times, Results	Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor	Predicting the race results and championship standings
Elena Loli Piccolomini, Davide Evangelista, and Massimo Rondelli **January 2022**	Driver, Race Location, Lap Time, Tyre Compound, Weather Characteristics, Track Temperature	Deep Neural Networks, Recurrent Neural Networks (LSTM, GRU), Multi-Layer Perceptron	Predict Tyre Strategy during a F1 race
Alexander Heilmier, Andre Thomaser, Michael Graf, Johannes Betz **November 2020**	Tyre Compound, Lap Number, Safety Car, Tyre Age, Racetrack	Artificial Neural Networks	Predicting Race Strategy Decisions
Rahul Rana, Devang Pandey, Satyam Mishra, Neelam Nehra, Deepti Deshwal, Pardeep Sangwan **December 2021**	Weather, Driver, Constructor, Circuit Details Driver Age, Pitstop Timings, Race Length, Results	Linear Regression, Lasso Penalised Regression, Ridge Penalised Linear Regression, K Nearest Neighbours Regressor, Support Vector Regressor	Predicting the Drivers Championship
Bo Peng, Jiayu Li, Selahattin Akkas, Takuya Araki, Ohno Yoshiyuki, Judy Qiu **December 2021**	Lap of Last Pitstop, Lap Time, Track and Lap Status, Historical Stint Distance, Time Behind Leader, Total Pit count, Fuel Level, Tyre type	ARIMA, Random Forest, SVM, Deep Learning Models for Time Series	Predicting a car's future race position based on the event's observed history
Diego Piccinotti **April 2021**	Race, Circuit Details, Driver, Constructor, Lap Times, Pit stops, Lap status, Tyre Compound, Tyre Age, Weather, Safety Car	Reinforcement Learning, Linear Regression, Random Forest	Predicting pit-stop strategy decisions
Xuze Liu, Abbas Fotouhi **March 2020**	Race, Circuit Details, Driver, Constructor, Lap Times, Pit stops, Lap status, Tyre Compound, Tyre Age, Tyre, Weather, Safety Car	Artificial Neural Networks, and Monte Carlo Tree	Predicting Race Strategy

Atheer Akram Abdulrazzaq, Omran Al-Shamma, Mohammed A. Fadhel **October 2020**	Lap Time, Lap Number, Total Time	Neural Networks	Predicting Motorsport Tyre Degredation
Alexander Heilmier, Michael Graf, Markus Lienkamp **November 2018**	Track Characteristics, Driver Characteristics, Tire Characteristics, Car Characteristics, Race Characteristics	Linear Regression	Predicting Race Strategy Decisions
Xuze Liu, Abbas Fotouhi, Daniel Auger **March 2022**	Race, Circuit Details, Driver, Constructor, Lap Times, Pit stops, Lap status, Tyre Compound, Tyre Age, Tyre, Weather, Safety Car	Artificial Neural Networks, and Monte Carlo Tree	Predicting Race Strategy
Oscar F. Carrasco Heine and Charles Thraves **April 2021**	Lap Times, Tyre Compound, tire wear, safety car, weather	Dynamic Programming	Predicting Pit-stop strategy
Claudia Sulsters **February 2021**	Pit Stops, Safety Car, Lap times, Overtakes, Fuel consumption, Tyre Degradation	Linear Regression	Predicting Race Strategies
Alexander Heilmier, Markus Lienkamp, Michael Graf, Johannes Betz **June 2020**	Pit Stops, Safety Car, Tyre Degradation, Driver, Circuit, Weather	Reinforcement Learning, Random Forest, Decision Trees	Predicting Race Strategies
Theja Tulabandhula and Cynthia Rudin **June 2014**	Age of tyre, position in race, starting position, lap time, pit stops, safety cars	Ridge Regression, SVR, Random Forest	Predicting Race Results and Strategies and identifying major challenges
Alexander Heilmier **March 2022**	Pit Stops, Safety Car, Tyre Degradation, Driver, Circuit, Weather, Car, and Race Characteristics	Reinforcement Learning, Random Forest, Decision Trees, Linear Regression	Simulating a Grand Prix and Predicting Race Strategy Decisions
Eloy Stoppels **December 2017**	Circuit Length, Race Laps, Weather, Driver Form, Starting Grid, Results, Qualifying Lap	Artificial Neural Network	Predicting Race Results

3. Methodology

In this section, an explanation has been provided for the steps taken in creating the machine learning models. Each part of this section will discuss the theory and base logic behind the actions taken in the next stage of data analysis and model creation.

3.1 Proposed Methodology

This study follows the CRISP-DM methodology, keeping everything structured and focused on the task at hand. CRISP-DM stands for Cross Industry Standard Process for Data Mining and is the de-facto industry standard data mining workflow. (Schroer, Kruse and Gomez, 2021). It makes the entire data mining process in an easy to follow 6-part process. Figure 3 adapts the shows how this study adapts the CRISP-DM Process.

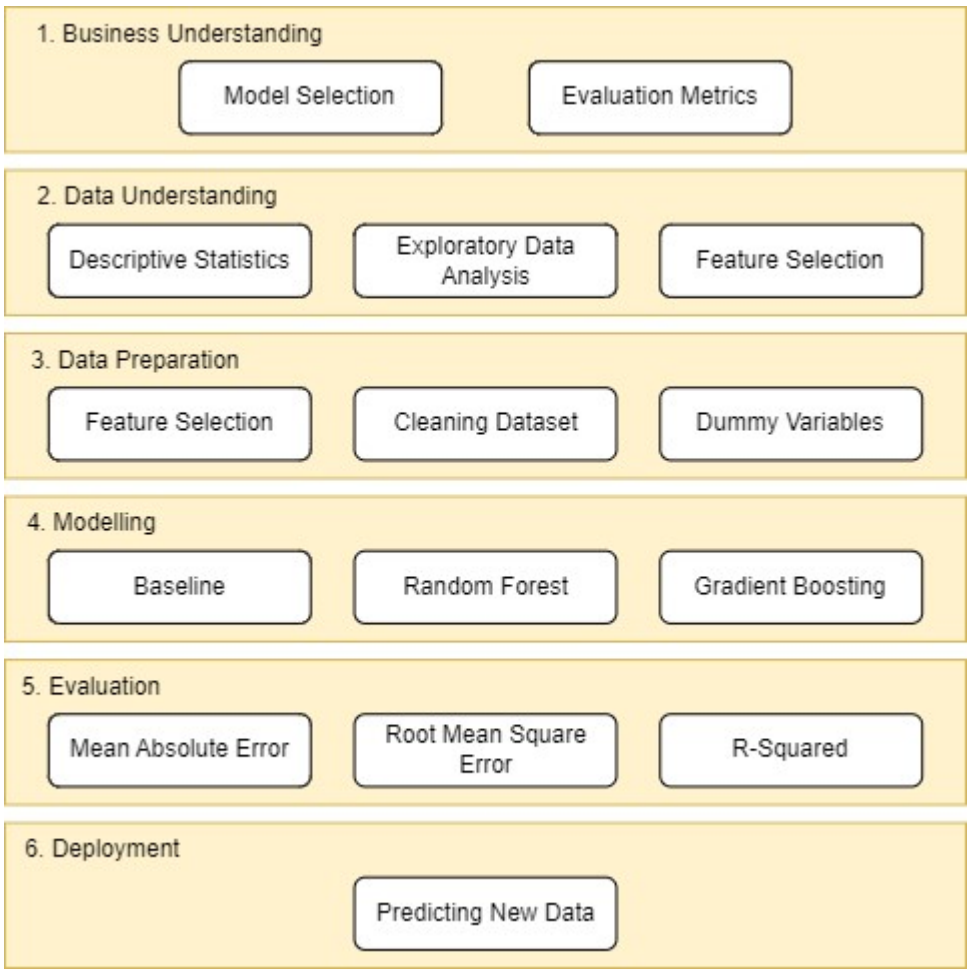


Figure 3 | Proposed Methodology

3.1.1 Business Understanding

Stage 1 of the methodology is Business Understanding. This stage focusses on determining the analysis goals, understanding the context of the problem, and implementing domain knowledge to identify key parameters and performance indicators. We also identify select the modelling technologies and evaluation metrics for the project at this stage to produce a plan for the project. This is the most important stage, and no project can start without completing the requirements of stage 1. The first part of this stage was discussed in chapter 1 when the problem, objectives and deliverables were defined. The next stage will be discussed later in this chapter when we identify the machine learning methods and evaluation metrics to use.

3.1.2 Data Understanding

Stage 2 is Data Understanding, and it mainly involves handling the data before it is processed for the analysis or modelling. At this stage we collect the data from the original source and verify its validity by doing sanity and validity checks. Basic descriptive statistics and exploratory data analysis are done at this stage to fully understand the data. The data is analysed and visualised as we look for relationships between the parameters. It is important to know what we are dealing with and how the data interacts. This will help us understand the results and aid the deployment.

3.1.3 Data Preparation

The next step is Data Preparation where the data is transformed for the analysis and modelling. During data preparation we create the final input data that will be used in the models. This involves cleaning the data of any outliers or empty values that might exist. At this stage the Data Preprocessing process of machine learning is implemented where the parameters are selected and dummy variables for the categorical data created. This is the step which transforms the basic dataset into something that meets the requirements of the models and can provide more knowledge.

3.1.4 Modelling

Once the data is ready to be processed, we build the machine learning models from the algorithms selected. In our project we use Random Forest and Gradient Boosting models. Both are tree-based ensemble algorithms and are comparable as they have common building blocks. During the modelling process, the models require hyperparameter tuning which helps in us to build the best model. Once the models are built, we can evaluate them in the next

stage. Modelling stage is the core of the entire process because the entire project leads up to this stage.

3.1.5 Evaluation

The models can be evaluated various ways to identify the absolute best one. Firstly, the different models created by hyperparameter tuning an algorithm are compared amongst themselves. Once we have the best models from both the algorithms, we compare them with the evaluation metrics identified. Since this is a regression problem, we will primarily use Root Mean Square Error and Mean Absolute Error. Evaluating and comparing model performances is important to identify the best model and understand why it works. The best model will likely be implemented to the business problem.

3.1.6 Deployment

The first 5 steps are often cycled through until an optimised solution is found and only then step 6 is implemented. Deployment can take many forms depending on the business problem being solved. In this project, deployment is predicting a small sample using the model.

3.2 Supervised Machine Learning Models

Prediction of laps is a supervised regression problem and the algorithm selected must be well-recognised regression algorithms. 3 Algorithms were selected to evaluate and identify the ideal one. Following the lead of Sicoie (2022), whose study emphasised on supervised ensemble learning methods, Random Forest Regressor and Gradient Boosting Regressor were identified as candidates. Random Forest and Gradient Boosting Regressor utilise the ensemble decision tree framework and can be good models for generalisation and non-linear relationships, also allowing to understand feature importance. These tree-based algorithms are great for interpreting the predictions and creating non-linear models.

None of the previous literature have used random forest and gradient boosting to predict tyre life in F1. Moreover, Gu, Foster, Shang, and Wei (2019) highlighted the advantages of ensemble learning methods in sports and the robustness of the models created. A comparison of bagging and boosting algorithms with non-ensemble methods was also provided for a better understanding of their performance. Their models were for classification but the accuracy of bagging and boosting models was on average better than non-ensemble models. Providing for strong evidence of their applicability in a sporting context.

3.3.1 Baseline Model

Before creating the main model, a baseline model is created. This is essentially a reference point in the machine learning project. The main purpose is to put all the other models in context of a simple predictor. Baseline models are simple and lack any complexity. Their low predictive power allows them to act as a point of control.

The most widely used baseline in regression models is to use the median of the target variable. The median of the target variable in the training data is used as a prediction to evaluate the testing set. That performance is evaluated, and those results will become the benchmark for the machine learning model. If the predictions in the machine learning models are outperforming this baseline, it can be concluded that the model has learnt something and is using the parameters to make its predictions.

3.3.2 Random Forest Regressor

To obtain a more accurate predictor, Random Forest was chosen as the firs Machine Learning algorithm for this project. Random Forest is an ensemble method constructed from a collection of decision trees. It uses the Bootstrap Aggregation technique, a.k.a. bagging, to help reduce the noise of many unbiased models. Therefore, reducing variance and building a robust model. Decision Trees are ideal candidates for the bootstrapping as the capture complex interactions in the data structure. Although they have low bias, they suffer from being noisy. That is where bootstrap aggregation comes into play and averages the results from multiple decision trees. (Hastie et al. 2008). Figure 4 shows how a Random Forest works, with x being the input variables and y being the aggregated prediction.

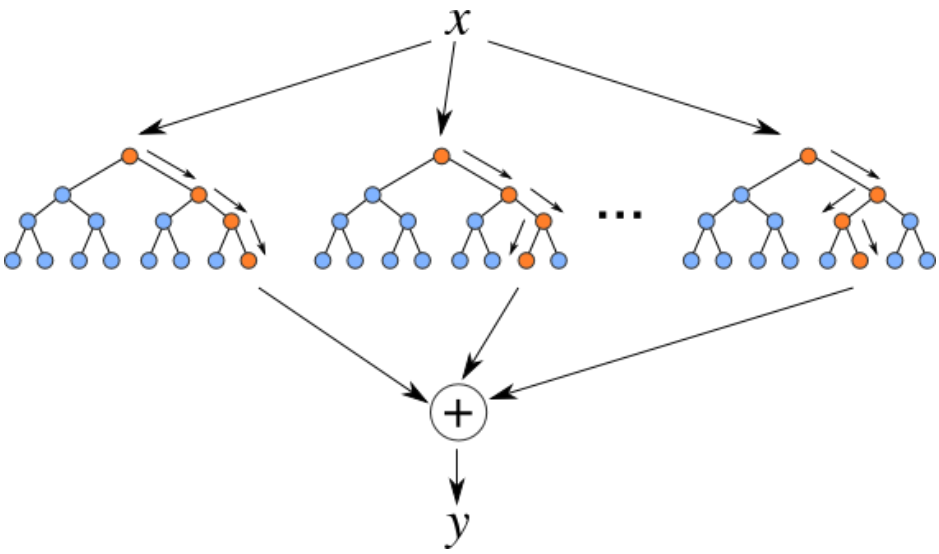


Figure 4 | Random Forest Diagram

In addition to reducing noise, random forests also reduce overfitting, one of the biggest problems with decision trees. Breiman (2001), with the use of the strong law of large numbers, displays that random forests always converge and do not overfit. The inherit randomness of the forest allows it to be trained in a more generalised manner.

Random Forests are constructed by using a subset of the data in each tree, where the subset could include parameters with varying importance. This could still influence some overfitting, but the collective intelligence and aggregated scores of random forests ensures that they have less overfitting and more accuracy than decision trees.

3.3.3 Gradient Boosting Regressor

Gradient Boosting is also an ensemble method that uses multiple decision trees to create a collective intelligence model. It is different from Random Forest in how the trees are trained. While trees in random forests are trained independently and aggregated at the end, trees in a boosting algorithm are trained in an iterative process. This sequential method develops a series of base regression trees to enlarge the model capacity. Each tree learns from its predecessor and the final model is a combination of several weaker models. (Wang et al, 2020) Figure 5 shows how a Gradient Boosting algorithm works.

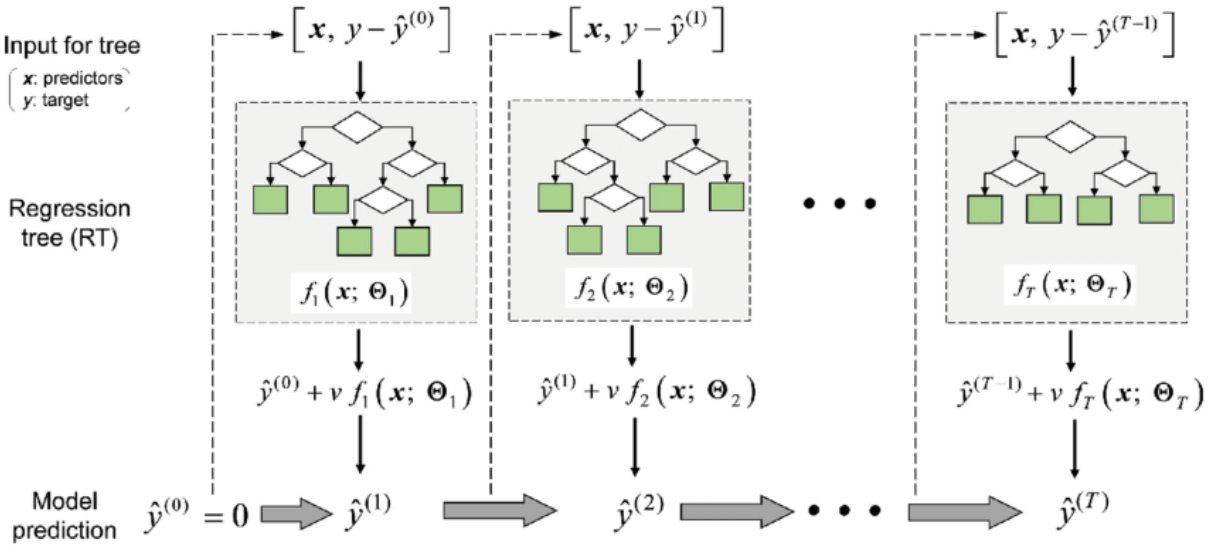


Figure 5 | Gradient Boosting Diagram

One major disadvantage of gradient boosting is the higher chance of overfitting compared to random forest, as the model records the contribution from the new tree. The transfer of knowledge between trees makes the model less generalised. The algorithm tracks if the subset in a tree had a high or low influence in the prediction. But while all subsets of data are considered, the emphasis on training the model as closely as possible makes it vulnerable to

outliers and noise, something that random forests avoid by aggregating and averaging all trees.

3.3 Evaluation Metrics

Evaluation Metrics are used to decide whether a model is good to work with and which methods perform better than others. Since this is a regression analysis, error metrics were used. The underlying logic behind error metrics is that they aim to summarise the characteristics of all errors in a single number. Mean Absolute Error and Root Mean Square Error are 2 of the most common regression metrics and are used in a wide range of fields and disciplines. (Santhusitha and Karunasingha, 2022)

3.4.1 Mean Absolute Error (MAE)

Mean Absolute Error is a basic error metric that finds the mean error of the predictions. The values are in absolute terms so as not to cancel each other out. MAE helps in understanding the variation in the predictions. This gives equal weight to all errors and is a very straightforward metric to interpret the predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

The MAE in this study can be used as a measure to evaluate the actual tyre life predictions of the model. In a real-world scenario, the F1 teams operate with a buffer to the tyre life predictions. The MAE as a measure be realistic to interpret with the buffer.

3.4.2 Root Mean Square Error (RMSE)

Root Mean Square Error is the most common metric in regression problems. It measures the square root of the mean square error. Mean Square Error (MSE) measures the mean of the squared value of all errors. RMSE is sensitive to the size of the error. As the errors are squared, the magnitude of each error is exaggerated. RMSE presents MSE in a lower value making it more comparable as it is depicted in the same units as the target variable.

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

Unlike MAE, RMSE is cannot be interpreted directly with the real-world buffer for the F1 teams. This is because RMSE is not an absolute value for tyre life. However, it is the most important measure to compare regression models.

3.4.3 Coefficient of Determination (R^2)

The Coefficient of determination, also called R-Squared is a goodness of fit measurement that displays how well the predicted values fit the actual data. It determines how well the parameters explain the predictions. The R^2 value lies between 0.0 and 1.0 and can be interpreted as a percentage. For example, a 0.643 value indicates that in the model 64.3% of the predicted value can be explained by the parameters, the other 35.7% is the random component. A higher coefficient of determination means a more accurate prediction, but it could also indicate overfitting.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

R^2 is calculated by subtracting 1 from the ratio of SSR (sum of square of residuals) and SST (total sum of squares) where residuals are the errors observed upon predicting the target variable. SSR is a sum of the squared value of the difference between predicted and actual values of the target variable. SST is the sum of the squared value of the difference between the mean and the actual values of the target variable. (Acharya, Armaan, and Antony, 2019)

3.4 The Dataset

As a first step for data extraction, some research was conducted into how authors of similar papers have found the data and if those sources are reliable. Ergast API and FastF1 Python library are comprehensively the best open-source data sources available now. Older papers have displayed a struggle in finding reliable data at a large scale, but the newer papers have all used FastF1 as the primary source.

Most of the papers since 2018, including Piccolomini, Evangelista and Rondelli (2022) and Sulsters (2018) have used Ergast API as a reliable source for historic data. This relates to data about races, drivers, and teams that is not collected live and can be verified easily using multiple sources like F1's own websites and other respected media outlets.
There is some data that is recorded by F1 and provided to all teams. All the publicly available data falls under this category. There is more data that is recorded by F1 but not released publicly, only the teams F1, and the governing bodies have access to this. Then there is the data that the teams collect for themselves. It is regulated properly, and the teams have to supply information to F1 and the FIA. The Ergast API and FastF1 library only contain the data that F1 themselves record and make available to the public.

Sicoie (2022) used live data such as track temperature and weather conditions, which impact the decision-making process and variables for F1 teams while choosing the tyre compounds to put on the car. This data consists of all the things that cannot be verified easily after the event ends. Things like track temperature, humidity, air temperature, and air pressure. Additionally, parameters like tyre compound, track status and safety car are also collected. The FastF1 API extracts this data from the live timing data supplied by F1.com during the races and other sessions. Although there is no way to validate this data, it is believed to be accurate for research and testing purposes. It is the best source for open license data and has become the standard in the F1 data analytics community.

A third, additional source has been defined as the 'Weekend Preview' infographics released by Pirelli before each race weekend. These include the fixed facts and characteristics of each circuit such as Asphalt Grip, Asphalt Abrasion, Tyre Grip, Track Evolution, etc. This data is not available as a dataset and had to be scraped for each race weekend. Additional information about the tyres such as minimum tyre pressure is also mentioned. But since this is a minimum recommendation from Pirelli and the actual values are not available as public data, these have not been included.

Heilmeier et al. (2020) advocated the use of lap-wise discretised approach, that uses lap-by-lap data for simulation of a race strategy. This allows the predictor to be accurate in dealing

with decision points placed at the end of each lap. However, this is not practical in the scope of the study as it involves computing the lap-by-lap degradation and the tyre performance on each lap. This requires data for more parameters and specialised preprocessing. It is not applicable to the current data source and methodology. The potential for this process will be discussed further later in the study as part of future research. Therefore, for now, although the data is collected for each lap, it is grouped by each driver's stints. This allows the factors to be aggregated and influence the laps run in a stint.

3.5.1 Feature Selection

The parameters considered for this study range from track characteristics to weather characteristics and tyre details.

Following is a list of the parameters (independent variables) and their details.

- **Driver:** The driver, each driver has their own driving style which may affect the tyre life
- **Team:** The team - car, each car has its own characteristics which may affect the tyre life
- **Compound:** The dry tyre compound, HARD, MEDIUM, and SOFT They have a performance differential in relation to each other.
- **Average Air Temperature:** Air Temperature the tyre is operating in.
- **Average Track Temperature:** Track Temperature the tyre is operating in.
- **Average Humidity:** Air Humidity the tyre is operating in.
- **Average Air Pressure:** Air Pressure the tyre is operating in.
- **Traction:** Traction value of the circuit (Pirelli)
- **Lateral:** Lateral value of the circuit (Pirelli)
- **Braking:** Braking value of the circuit (Pirelli)
- **Tyre Stress:** Tyre Stress value of the circuit (Pirelli)
- **Asphalt Grip:** Asphalt Grip value of the circuit (Pirelli)
- **Asphalt Abrasion:** Asphalt Abrasion value of the circuit (Pirelli)
- **Track Evolution:** Track Evolution value of the circuit (Pirelli)
- **Downforce:** Downforce value of the circuit (Pirelli)

Target (Dependent) Variable –

- **Tyre Life:** The number of laps a tyre ran (race stint + practice or qualifying laps)

Average Lap Time is a major parameter that has been omitted from the selected dataset. This is because in the context of the problem, laptime is a result of a function of strategy. The choice of tyre compounds, tyre wear, and circuit characteristics influence the laptimes drivers can achieve. (Delzell, McCabe, and Mourad, 2019) Additionally, when deciding the strategy, laptime is not used as an input. Target laptime is considered only after implementing the chosen strategy. Laptime is an important variable in F1 performance and predictions, but it fails to fit in the context for this study, therefore it will not be included.

3.5.2 Cleaning the dataset (Random Events)

To build a representative model, it is important that the data is representative to the actual conditions as well. While F1 races rarely run smoothly and without disruption, the teams make the most important strategy decisions assuming ideal race conditions. Filtering out the noise that is not relevant to the model and does not play any role in strategy planning is beneficial because then the model would be efficient. To train a model that is unbiased to random events, it is important to reduce the noise and use a dataset that handles these disruptions.

Oftentimes when there are random events which disrupt the strategies of drivers. 'Random' events in F1 races are unpredictable chance events such as mechanical failures and crashes (Peng et al. 2021). Each event influences the race for every driver. A mechanical failure or a crash could mean that a car is sitting unmoving on the racetrack. Or that it must retire and not complete all the laps. A stopped car or a crash could lead to safety related intervention - Safety Car, Virtual Safety Car or even a Red Flag event where the session is paused completely.

In situations like safety cars or red flags, drivers pit and change their tyres. This means that they are not utilising their tyres to the maximum. During safety car, pitting is not mandatory, but the normal racing is affected, and the tyre life can be extended as they drive slow while maintaining the tyres. These incidents could lead to a very skewed sample and add bias to the training data. Using it in the training model would mean that the model will be influenced by records relating to extraordinary circumstances.

Random Events have been identified as follows: Safety Car, Virtual Safety Car, Red Flag, or any combination of these 3 – indicating interruptions to normal proceedings; Intermediate and Wet tyre compounds – indicating rainfall; Did Not Finish – indicating a mechanical failure or a crash; and less than 10 laps in a stint – indicating a premature pitstop for any of the above or other miscellaneous reasons. A dataset that is free of the above random events can be assumed to be representative of interruption-less dry weather running.

3.5 Data Partitioning

Once the data is loaded and basic sanity checks are conducted the data is split into training set and testing set. It creates a sample dataset for training the model and another dataset for evaluating it. For a machine learning model to be effective, the train – test split must be done properly. The data is divided into an 80/20 split as it is the basic rule of thumb. It allows enough records in the test data to have a good evaluation. There are two types of sampling -random and stratified. Since this is a regression problem, we will be using random sampling.

3.6 Descriptive Statistics and Exploratory Data Analysis

Once the data is split into train and test sets, we set aside the test set and only focus on the train set for now. We do not want to look at the test set. Exploratory Data Analysis is conducted on the train set to better understand the data.

3.6.1 Target Variable

The Descriptive Statistics of the target variable show that the training set has 1788 records where the mean tyre life is 24.75 laps and 23 laps being the median, with a range of 56 laps. It seems to be left skewed in its distribution. This skewness can be explained by the nature of the problem.

Table 2 | Descriptive Statistics of Target Variable

Descriptive Statistics	Tyre Life (Target Variable)
Count	1788
Mean	24.754474
Standard Deviation	9.660181
Minimum	10
25%	17
50%	23
75%	31
Maximum	66

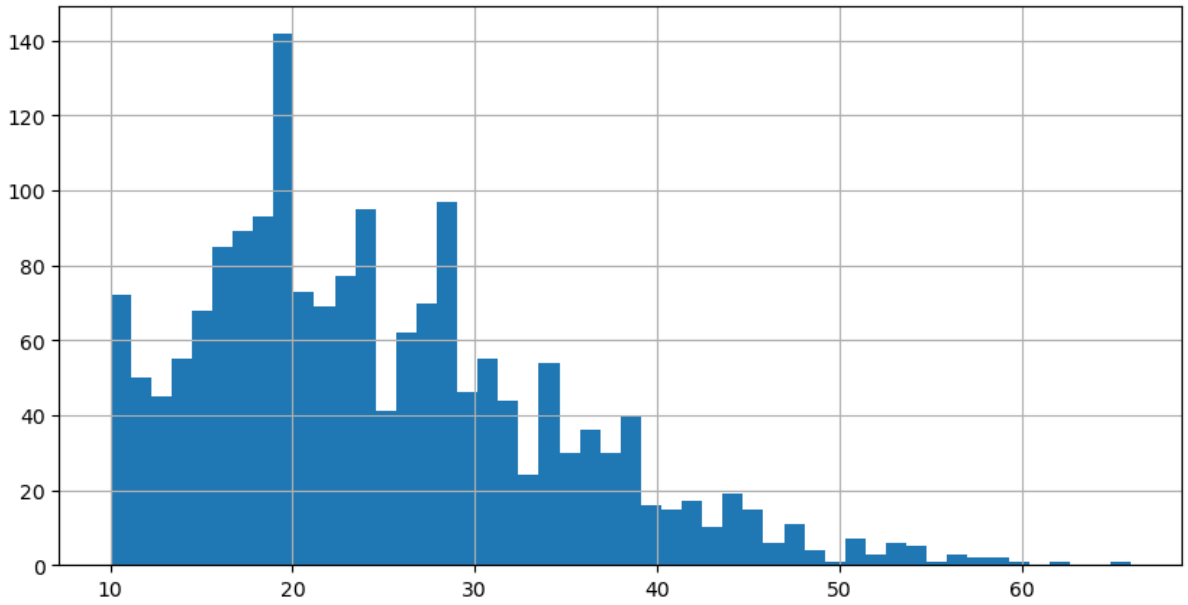


Figure 6 | Distribution of the Target Variable (Tyre Life)

3.6.2 Independent Variables

Mean avg_airtemp and avg_tracktemp is 24.23°C and 36.16°C respectively. Avg_tracktemp has a range of 40°C while avg_airtemp has a range of 28°C. avg_humidity has a mean of 51.63 but this is in a different scale than the temperature variables. Its range is 81.3. The avg_airpressure, in a different scale altogether has a mean of 986.19 and a range of 243.3. The weather data is largely normally distributed, apart from avg_airpressure which is left skewed.

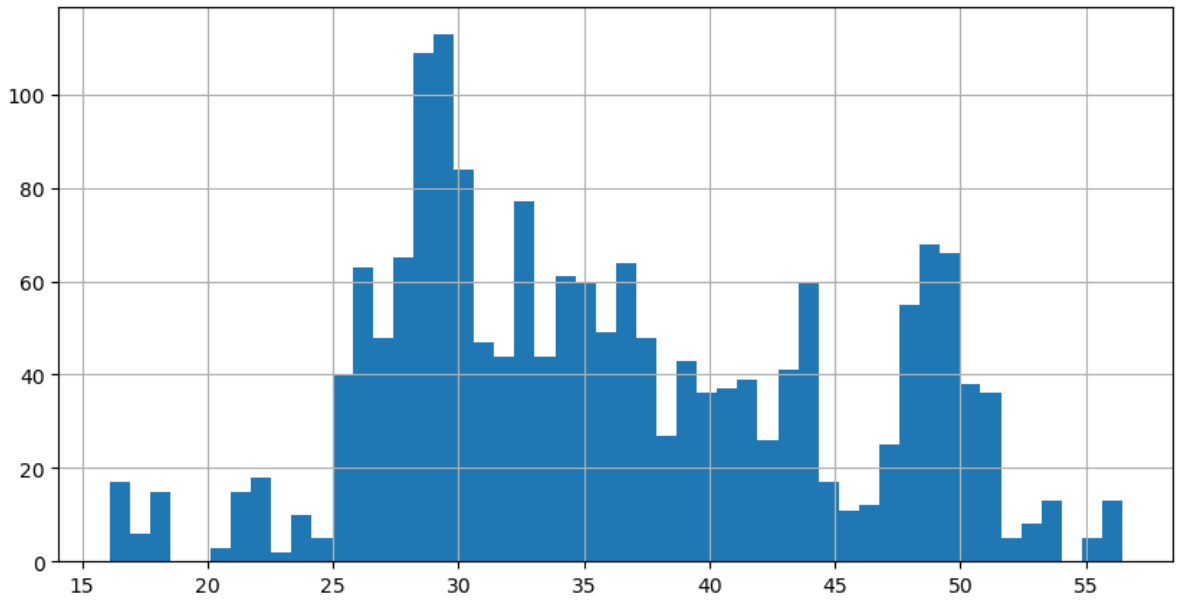


Figure 7 | Distribution of the Average Track Temperature

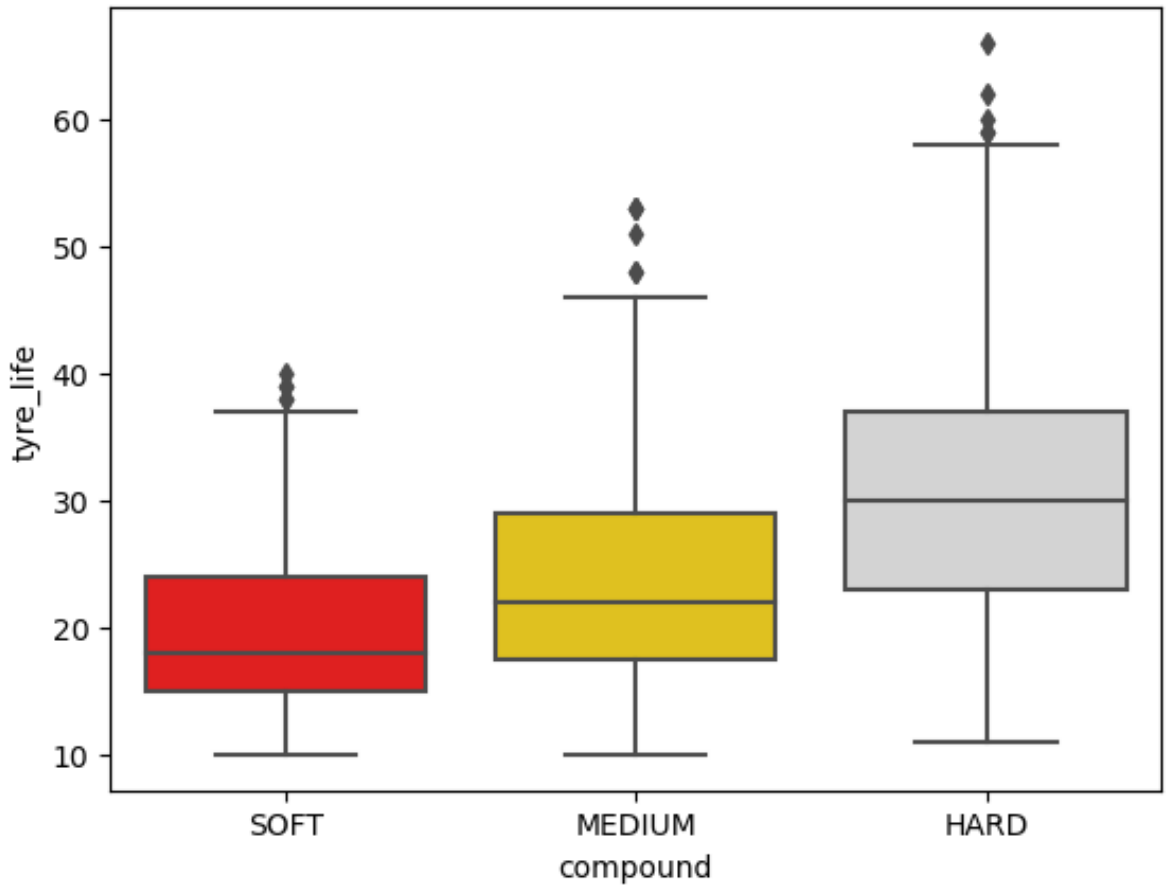


Figure 8 | Tyre life distribution per compound

As for the tyre compounds, there are more records with MEDIUM, than there are for SOFT or HARD. In the context of grand prix racing, MEDIUM is the most likely compound to be used as it is a compromise between the speed of SOFT and durability of HARD. Figure 8 shows the distribution of tyre life per compound, and it is evident that on average SOFT tyres have the least laps and HARD tyres have the most laps run.

The boxplots of tyre life distribution on circuit characteristics show that there is a relationship between tyre life and Lateral, Asphalt Grip, Asphalt Abrasion, and Tyre Stress. The relationship being that the mean and of tyre life in these categories reduces as the value of each variable goes up. The other 4 circuit characteristics – downforce, track evolution, braking and traction do not have a direct relationship with tyre life. In downforce the means of value 2 through 5 are progressively higher and only the category 1 is out of the trend. On the other hand, in Braking 2 through 4 have a falling mean and only 5 is out of the trend with a higher mean than the rest.

This presents a point of difference among the circuit characteristics. This will have an impact on the model while predicting the tyre life. We shall verify this when the parameter importance is being evaluated.

For example, in figure 9, the distribution of tyre life per tyre stress category displays a negative correlation between them. This holds true in the context of the problem as a higher tyre stress value reflects that the circuit has a high level of degradation on the tyre.

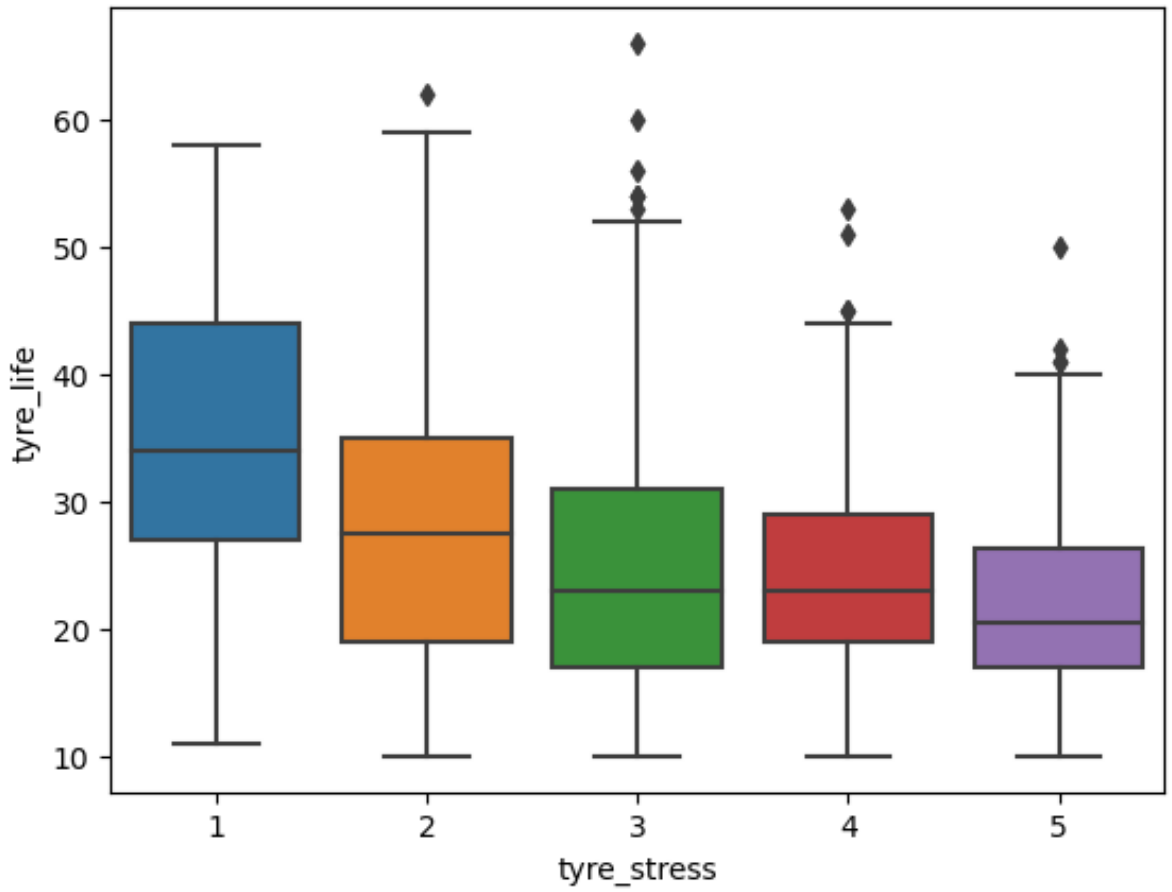


Figure 9 | Distribution of tyre life on tyre stress categories

3.6.4 Correlation Analysis

The correlation analysis shows a high positive correlation of 0.7 between avg_airtemp and avg_tracktemp, along with a negative correlation of -0.5 between those parameters and avg_humidity. This is clearly a reflection of the weather conditions and further validates the data itself. There is no significant correlation between tyre life and any of the weather parameters. Avg_airpressure has the highest correlation with tyre life at -0.2.

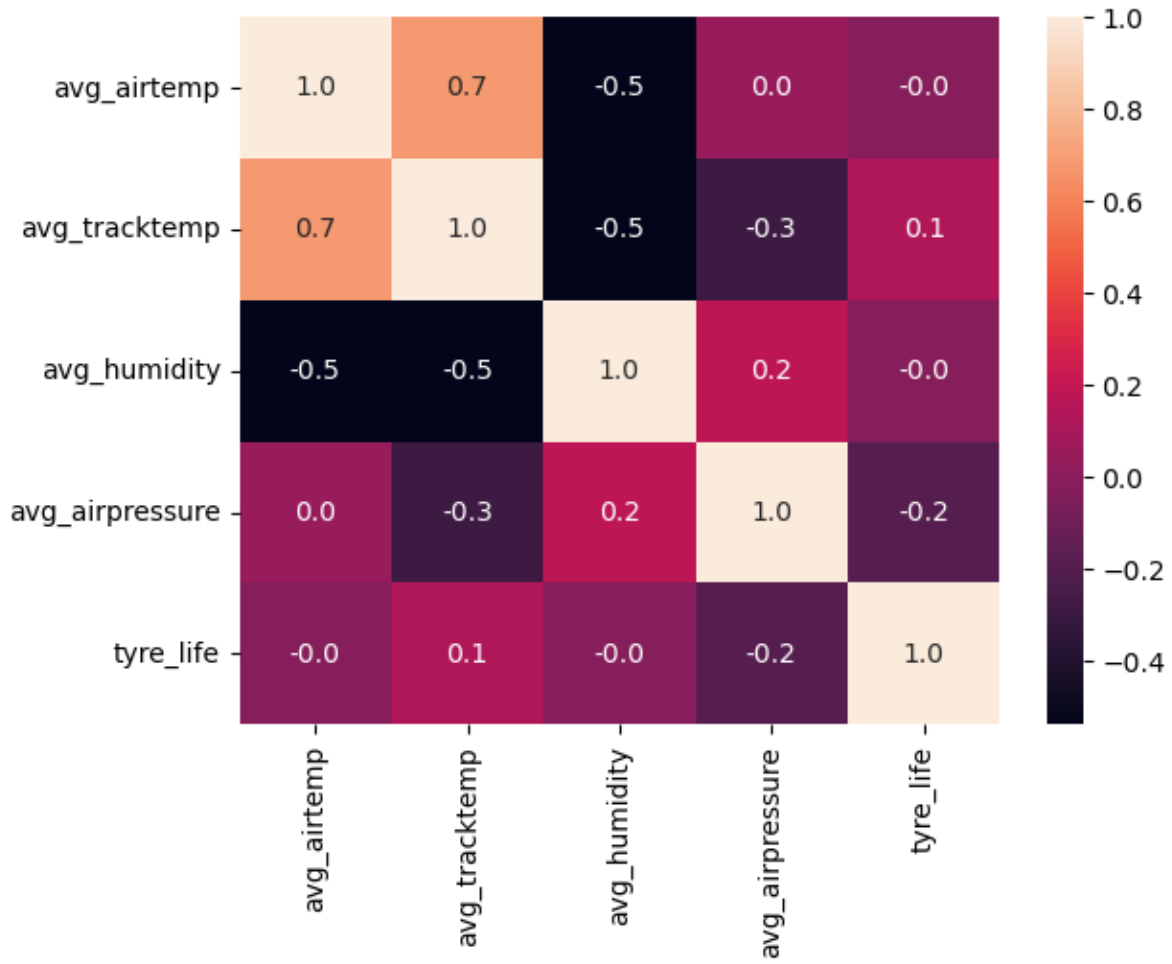


Figure 10 | Correlation Analysis of the numerical parameters

3.7 Data Preprocessing

Data Preprocessing consists of identifying errors in the data and making the data more eligible for the machine learning. This included checking for missing values, duplicate records, and outliers. Thereafter, the independent variables were scaled so that all the values are in the same scale. This allows the machine learning models to perform better. Since Random Forest and Gradient Tree Boosting are tree-based models, scaling is not required. Lastly the categorical variables are transformed into dummy variables as required for machine learning problems.

4. Results

In this section, we the results and insights from the modelling will be presented. The two machine learning algorithms were trained and tested on the data to see if they can be used to predict tyre life from a set of variables that represent driver, circuit, and weather characteristics. The results will be discussed along with the different hyperparameter settings. Feature importance and overfitting will also be presented to understand the predictions and results better. We will then conduct the simulations in the next section to discuss the findings properly.

4.1 Hyper-parameter settings

Several hyperparameter settings were investigated for both Random Forest and Gradient Boosting algorithms using the Grid Search CV function in Scikit-learn library. Grid Search CV goes through all the hyperparameter settings and makes predictions on the training set, multiple times with each setting. The mean performance of those predictions are used to evaluate the different hyperparameter settings. This process assesses and cross validates all the hyperparameter settings assigned and returns the performance scores for each. Then those scores can be used to identify the best hyperparameter settings in the algorithm to make the final model and run it on the test set. Tables 3 and 4 show the 5 best hyperparameter settings considered for Random Forest and Gradient Boosting respectively.

For Random Forest, the hyperparameters changed were number of trees and max depth of each tree. Each setting underwent a 10-fold cross validation. The number of trees considered were 10, 30, 50 ,100, 150, and 200. Max depth considered was 3, 4, and 5. Limiting the max depth is important as very deep trees perform excellent on training data and are prone to overfitting. For Gradient Boosting, the same settings were considered for the common hyperparameters. But there is an additional learning rate hyperparameter that was considered for 0.01, 0.05, and 0.1. This is the gradient learning specific part of gradient boosting algorithms.

With all hyperparameter considered, there were 18 settings for random forest and 54 for Gradient Boosting. Each setting was fitted 10 times for cross-validation. The resulting models were considered while searching for the best model in each algorithm.

Table 3 | Top 5 Hyperparameter Settings for Random Forest

hyperparameters	train score	test score	difference (%)
{'max_depth': 5, 'n_estimators': 150}	6.502672	6.953081	- 6.926522
{'max_depth': 5, 'n_estimators': 200}	6.501340	6.953862	- 6.960445
{'max_depth': 5, 'n_estimators': 100}	6.501340	6.956871	- 6.949049
{'max_depth': 5, 'n_estimators': 50}	6.509527	6.960489	- 6.927720
{'max_depth': 5, 'n_estimators': 30}	6.522578	6.965202	- 6.786035

As Table 3 shows, the best Random Forest model has a max depth of 5 and 150 trees. This setting is very close to the next best ones, all with a max depth of 5 and progressively lesser trees. The train and test scores are very similar to each other with less than a 7% difference between them. This indicates that the Random Forest model has a good generalised fit and does not overfit.

All the top 5 models have a max depth of 5 and the complete set, available in the appendix, shows that the top 6 have a max depth of 5, followed by 6 with a max depth of 4 and the final 6 with a max depth of 3. The number of trees fluctuate in each group, but the max depth seems like an important hyperparameter.

Table 4 shows the hyperparameter settings with the best 5 scores in Gradient Boosting. The selected model, the one with the best score, has a max depth of 5 and 100 trees, with a learning rate of 0.1. The next best model is similar, but with 150 trees. 4 of the top 5 scores have a max depth of 5. 4 of the top 5 scores also have a learning rate of 0.1, making it an important hyperparameter. This is relevant because it is a unique hyperparameter to Gradient Boosting.

The issue with Gradient Boosting seems to be overfitting. The train score is considerably lower than the test score for all top 5 settings and the next 17 best settings. This shows that the model performs too well on the training data. It fails at generalisation and performs poorly at the testing data. It reflects that there might be some overfitting in the Gradient Boosting Model. Overfitting is discussed further, later in the chapter.

The complete set of settings, available in the appendix, shows that the 26 of the settings have a difference between test and train scores that is lower than their difference in the best Random Forest scores. But these scores are very low in performance and are mainly made up of settings with a lower number of trees and a lower learning rate.

Table 4 | Top 5 Hyperparameter Settings for Gradient Boosting

Hyperparameters	train score	test score	difference (%)
{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}	2.647685	5.567573	- 110.280768
{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 150}	3.184264	5.665127	- 77.910104
{'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 150}	3.605842	5.799317	- 60.831140
{'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 150}	3.935326	5.802376	- 47.443345
{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50}	3.878488	5.814415	- 49.914456

4.2 Model Improvement

Figure 11 visualises how the hyperparameter settings changed and improved the performance of the Random Forest models. The hyperparameter settings in Random Forest progressed as follows:

- First level is Max Depth – 3, 4, 5.
- Second level is Number of trees – 10, 30, 50, 100, 150, 200.

Each Max Depth iterates through the number of trees and moves on to the next. First all the tree settings of max depth 3 are considered, then for max depth 4 and lastly for max depth 5. This can be seen in the graph above as the score improvement can be classified into 3 progressive groups of settings. The drastic improvement in the score represents a change in max depth. This clearly shows that deeper trees always perform better in Random Forest.

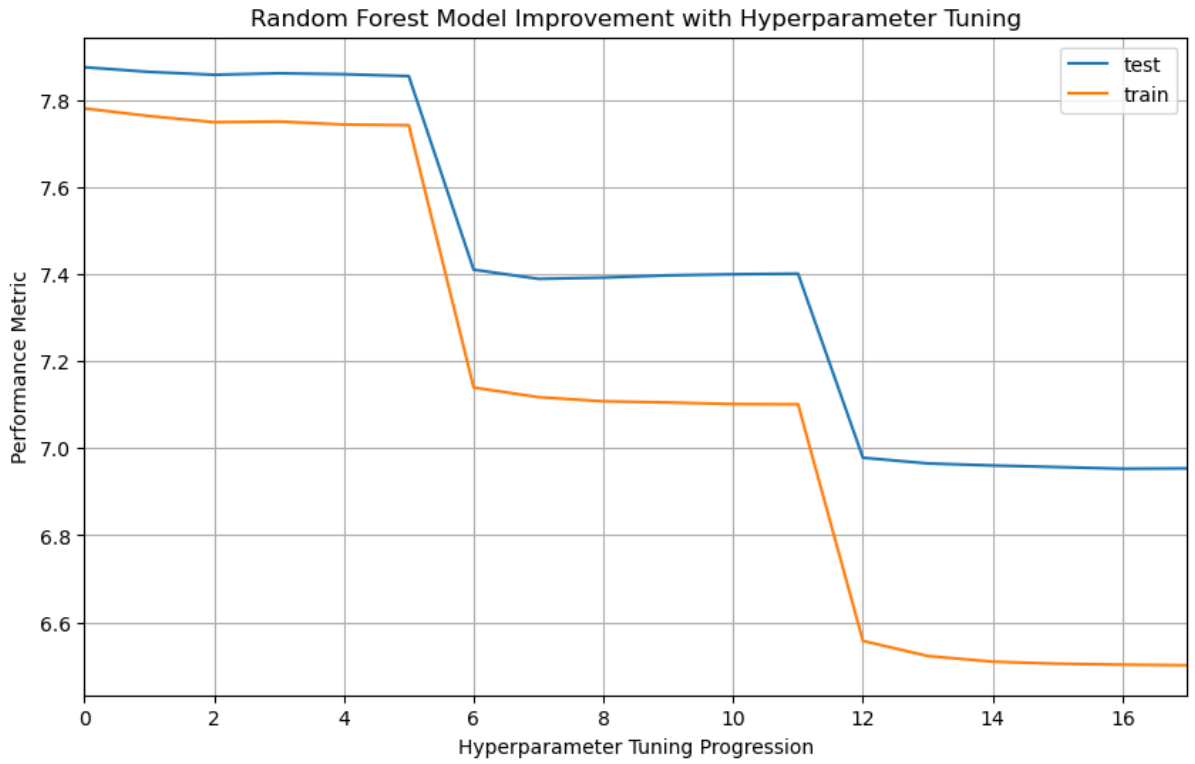


Figure 11 | Model Improvement for Random Forest

Figure 12 visualises hyperparameter settings and performance improvement for Gradient Boosting models. The hyperparameter settings in Gradient Boosting progressed as follows:

- First Level is Learning Rate – 0.01, 0.5, 0.1.
- Second level is Max Depth – 3, 4, 5.
- Third level is Number of trees – 10, 30, 50, 100, 150, 200.

The basic trend that exists in random forest also exists in gradient boosting. Meaning that the depth of the trees severely influences the performance of the models. But the additional hyperparameter of Learning Rate is implemented before the max depth. This breaks the progression into roughly 6 parts as shown in the figure below.

The settings for learning rate 0.01 and max depth 3 are considered with number of trees going in progression. This is the first downward trend. Then, the max depth changes to 4 for the learning rate of 0.01, creating the second downward trend which begins above the end of the last trend. And so, each hyperparameter setting iterates through the 54 combinations to find the setting with the best performance.

As with max depth in random forest, learning rate seems to have a significant influence on the performance. Another major trend visible in gradient boosting is the difference between training and testing data. As learning rate increases, the difference between the performance

on training and testing set also increases. There are no horizontal lines in the chart which depicts that the hyperparameter settings are important, as opposed to random forest.

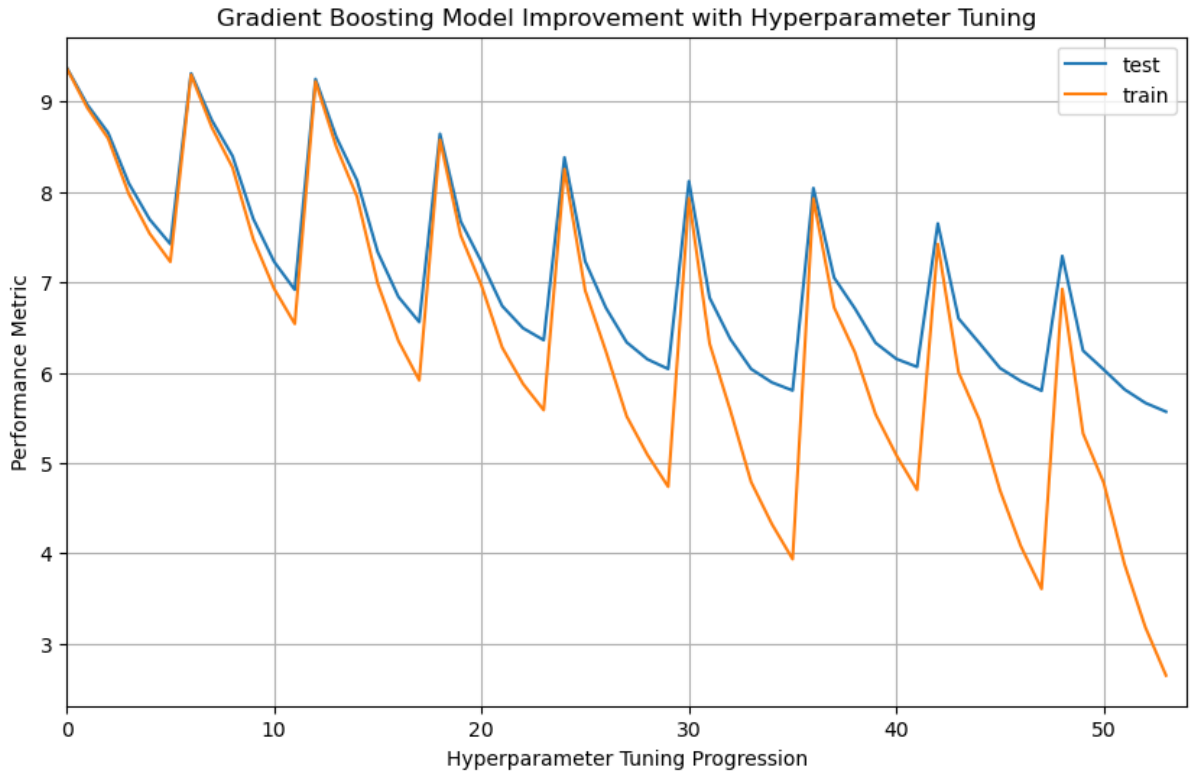


Figure 12 | Model Improvement for Gradient Boosting

The shape of the line is vastly different in both the models. In Random Forest, the number of trees is not very influential in the score as the progression lines are mostly flat but drop drastically when the max depth is changed. The shape of the progression line in gradient boosting is different as the score drops with the number of trees but climbs back up as the learning rate and depth change. This is a proof that the number of trees is more influential in gradient boosting than in random forest. As the number of trees increases, the gradient boosting model improves the score drastically.

The number of trees act differently when combined with learning rate in a boosting model. The bagging and boosting fundamentals of the two models make this aspect unique. In random forest the 30 or 200 trees are created in parallel and are averaged out. But in gradient boosting each tree learns from the previous one, so a the 200th tree will have information from 199 previous trees. This level of training can affect the generalisation capabilities of a machine learning model.

4.3 Model Comparison

The two best models were tested, and their performance measured using two metrics – Mean Absolute Error and Root Mean Squared Error. Table 5 provides the values of the measures making them comparable. The results clearly show that the baseline model performs worse than the machine learning models, indicating that the machine learning is at least better than using the mean as the predicted value. RMSE is used to compare the metrics and MAE is used to interpret the predictions of the models independently. The models were considered for their goodness of fit using the coefficient of determination (R^2).

While considering MAE, the baseline model is off by 7.61 laps on average. The Random Forest model is off by 6.05 laps on average whereas the Gradient Boosting model performs better than the other two as it is off by 4.06 laps on average. The RMSE scores reflect that the Random Forest model is off by 8.01 laps on average and Gradient Boosting is off by 6.11 laps on average. The baseline model is off by 9.81 laps on average.

Random Forest has a higher score than Gradient Boosting in both MAE and RMSE metrics. This means that Gradient Boosting is the better performing algorithm. The difference between Random Forest and Gradient Boosting is about 2, meaning that the Random Forest predictions are about 50% further away than Gradient Boosting. For RMSE, this difference is almost 33%.

Table 5 | Model Comparison

	Mean Absolute Error	Root Mean Squared Error	R Squared
Baseline Model	7.618009	9.815553	
Random Forest	6.052311	8.015541	0.441549
Gradient Boosting	4.068472	6.117151	0.674750

The R^2 values for the two models was calculated to determine the goodness of fit, and check for symptoms of overfitting. Random Forest has a R2 value of 0.441 while Gradient Boosting has a R2 value of 0.674. This indicates that that gradient boosting has a better fit and explains the predictions better. Gradient Boosting's R^2 value is not very high, but it is still considerably higher than Random Forest. The Gradient Boosting model can explain 67.4% of the value of the tyre life, while Random Forest can explain only 44.1%.

As a result of the better scores, Gradient Boosting would be considered as the chosen machine learning model for this problem.

4.4 Overfitting

While the Gradient Boosting model is a better model, it has some level of overfitting. Figures 13 and 14 show the predicted value and actual value from the training set for Random Forest and Gradient Boosting respectively. These two charts show us how both the models perform on the training data. In Gradient Boosting the predicted values are much closer and to the actual values, while in Random Forest they are generally further away. This shows slight overfitting in the gradient boosting model. The charts contain just 50 of the 1788 records in the test set to use as visualised examples.

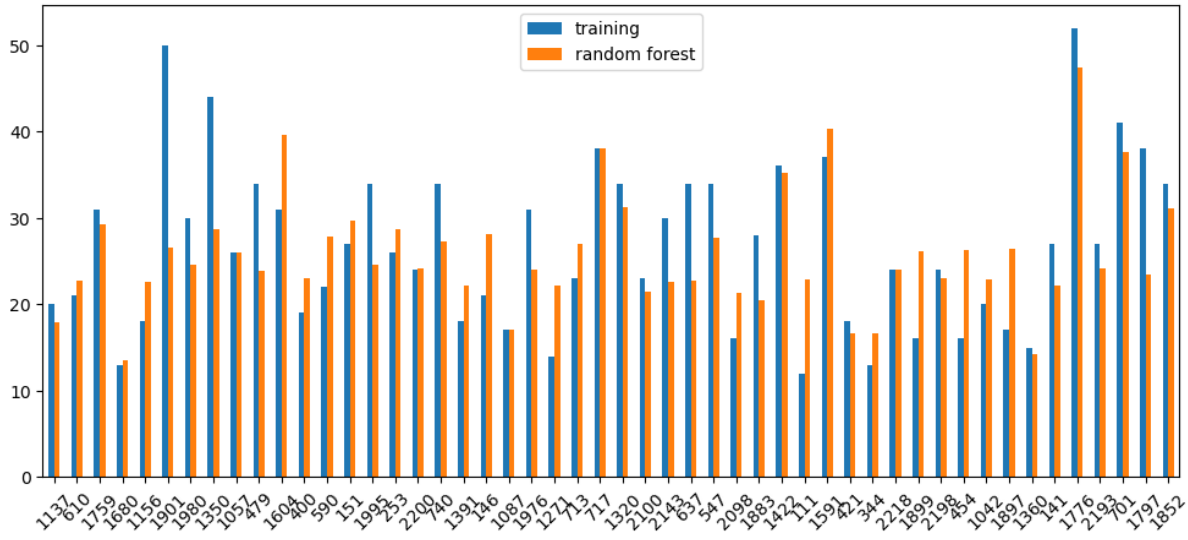


Figure 13 | Random Forest performance on Training Data

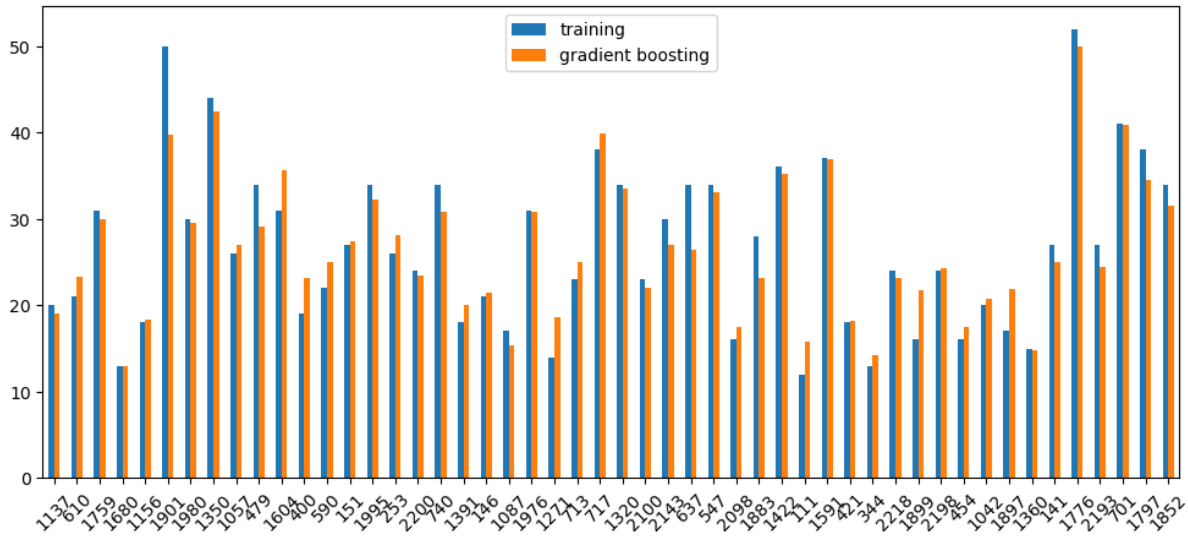


Figure 14 | Gradient Boosting performance on Training Data

4.5 Feature Importance

Since both the algorithms considered were based on decision trees, it is easy to find out and interpret the feature importance. This allows us to understand the dataset better and put the predictions into context of the problem. Figures 15 and 16 show features importance in Random Forest and Gradient Boosting respectively. While looking at this we need to consider the context of the problem as the compounds affects the tyre life and are the most important factor by their nature.

In Random Forest Compound is exceptionally more important than any other feature, with an importance of more than 0.35. The next best feature is Track Evolution with 0.12. This is interesting because during the exploratory data analysis stage we identified that the mean of tyre life distributed across Track Evolution did not follow a trend. Circuit is the third most important feature, and it is in line with the setting of the problem.

3 of the 4 weather features follow next in the top 6, with only Avg Air Temperature having an importance score of 0. The 8 circuit characteristics are spread across the spectrum, the lowest of which are Braking and Tyre Stress.

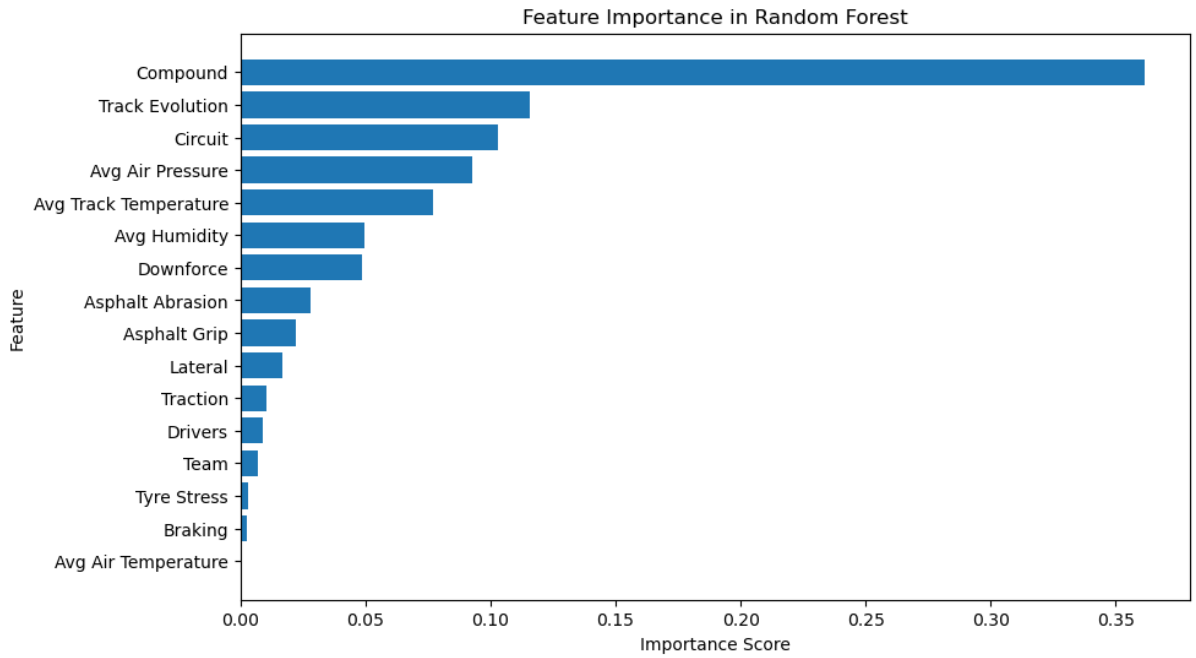


Figure 15 | Feature Importance in Random Forest

The feature importance chart for Gradient Boosting looks much different as Compound has a much lower importance at just under 0.2. And even the gap between Compound and the second most important features is down to just about 0.05 from 0.25. The second most important feature being Avg Air Pressure. Indeed, avg air pressure was the numerical feature

with the highest correction to tyre life. Once again, the Avg Air Temperature is at the bottom, and the other 3 weather features are near the top. Avg Air Pressure is followed by Avg Track Temperature and Humidity. Circuit and Track Evolution have a lower importance score than in Random Forest. 4 of the 8 circuit characteristics are in the bottom 5.

On comparing both the feature importance charts, it can be concluded that the general trend of feature importance is similar in both the models, even if the scores and their ordinal positions change. Perhaps a different model would give a different feature importance chart, but compound, avg air pressure, circuit, and avg track temperature will be near the top.

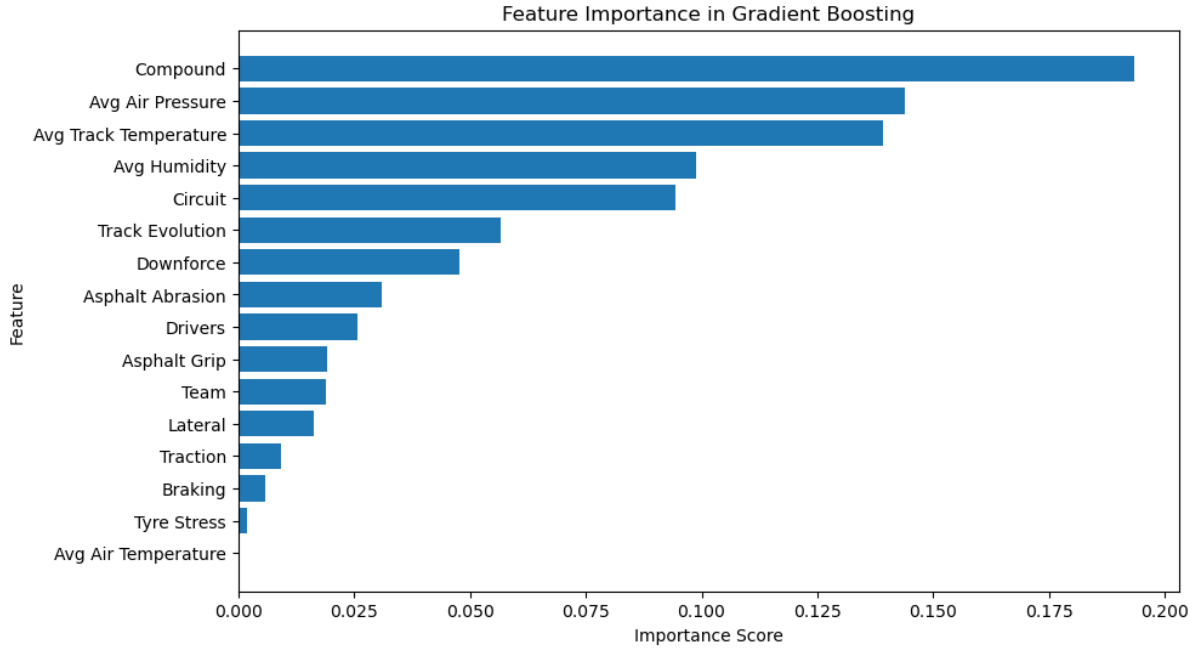


Figure 16 | Feature Importance in Gradient Boosting

5. Simulation

Once the best models were identified, they were used to run some simulations, to interpret the models better. This also acts as a validation exercise for the models. Checking how the perform outside the original train-test dataset The simulations were run on 6 different situations.

5.1 Situations

The scenarios will be different from one another by small changes, to identify the variations and compare the predictions. This will allow for the models to be evaluated with some real-world data. The inputs will be from Grand Prix conducted after the data for this model was extracted. The predictions made can be used to further assess the models and form conclusions.

The 6 situations can be split into 3 pairs depending on multiple factors. This is done because 3 Grands Prix were conducted after the training data had been collected. This gives 6 new records to try the models on and verify them against the actual data.

The records can be paired on circuit, compound, and team. There are 2 records from each of the 3 races mentioned and each compound has been accounted for twice. For teams, there are 2 records from each Ferrari and Haas and the other 2 records are from different teams Those last two records are from the same circuit, providing a common ground.

Table 6 | Situations to be simulated.

Parameters	Situation 1	Situation 2	Situation 3	Situation 4	Situation 5	Situation 6
circuit	Zandvoort	Zandvoort	Monza	Monza	Singapore	Singapore
driver	ZHO	OCO	SAI	MAG	HUL	LEC
team	Alfa Romeo	Alpine	Ferrari	Haas	Haas	Ferrari
compound	MEDIUM	SOFT	HARD	MEDIUM	HARD	SOFT
avg_airtemp	16.95238	17.35556	29.95625	30.02353	29.74054	29.99474
avg_humidity	64.66667	69.33333	40.65625	40.64706	74.43243	72
avg_airpressure	1007.643	1007.928	1000.231	1000.212	1009.695	1008.895
avg_tracktemp	27.02381	26.08889	42.57813	41.99412	35.38108	36.66316
traction	4	4	3	3	4	4
braking	3	3	4	4	5	5
lateral	4	4	2	2	2	2
tyre_stress	5	5	3	3	2	2
asphalt_grip	2	2	2	2	3	3
asphalt_abrasion	3	3	3	3	3	3
track_evolution	4	4	3	3	4	4
downforce	4	4	1	1	5	5

5.2 Simulation Predictions

Table 7 | Simulation Predictions (in Laps)

Model	Situation 1	Situation 2	Situation 3	Situation 4	Situation 5	Situation 6
Random Forest	23	17	27	23	27	21
Gradient Boosting	24	20	25	21	26	24
Actual	**27**	**18**	**32**	**17**	**42**	**22**

It must be noted that the predictions have been rounded to an integer to reflect real world application.

The 6 situations representing 6 unique records with slight changes and some commonalities were simulated using both machine learning models developed. On first inspection, it looks like Gradient Boosting predictions are further away from the actual tyre life more often than Random Forest predictions. But upon calculating the MAE for both, they are almost equal. Random Forest has a MAE of 5.33 and Gradient Boosting has a MAE of 5.66 laps. They have a RMSE of 7.11 laps and 7.5 laps respectively. We are not evaluating the performance here; this is just an observation as the evaluation was completed in the modelling stage.

There is only one situation where the error in greater than 7 laps, in situation 5. Other than that, every prediction is very close to the actual value of the tyre life.

When comparing absolute predictions and not the error, it must be noted that predictions on both models are close to each other. The biggest difference in the predictions on a situation is only 3 laps.

Comparing the predictions based on the pairs created is a good way to check the validity of these predictions in the real world. For instance, let's consider the 3 pairings made on the Grand Prix. The models correctly predicted that the harder compound would run longer than the softer compound. In fact, when the 6 situations are considered together, the 2 highest are both HARD tyres, and the lowest prediction is for SOFT tyres.

6. Discussion

As the 2 models have been trained, tested, and validated with simulations to make new predictions, it is now necessary to evaluate the project. This section will wrap questions raised and objectives set in chapter 1 and discuss the interpretations of the models.

6.1 Problem Solution

The MAE and RMSE scores were essential in evaluating the performance and practicality of the selected models. The Gradient Boosting model had the better scores, but the Random Forest had a model with very low overfitting. Overall, the models seem to have predicted the tyre life well but with a moderate goodness of fit. The models can be used by teams to predict the tyre life well and aid in strategy planning during an F1 Grand Prix.

6.2 Research Questions

In this section, we answer the research questions outlined in chapter 1. These questions were the basis of the study and answering them is one of the main objectives.

6.2.1 RQ1 Can machine learning techniques accurately predict the life of F1 tyres in a race using open-source data?

The models developed made in way that anyone with a basic knowledge of machine learning and access to open-source data can replicate them and create their own models with varying data. In that regard the research has been moderately successful. The models upon evaluation, gave good prediction on the testing data and on the simulations.

The models created and chosen for the simulations were good in terms of MAE scores but did not perform well on the goodness of fit. Overall, it appears that the models are not optimised and need further improvement. At this stage they should not be used to predict tyre life in F1. The Random Forest model performed poorly, and the Gradient Boosting model appears to have overfitting. But the simulations had good accuracy and returned with largely acceptable results. Further improvement on the models can lead to much more accurate predictions and improved viability.

6.2.2 RQ2 What are the most important parameters influencing the running life of F1 tyres?

The feature with the highest importance scores is Compound but this does not give any additional information as it is the nature of the compounds to influence the tyre life. The important take aways from this study lie in the feature importance after compounds.

Firstly, let's get the least important features out of the way. Avg Air Temperature the least important feature with a score of 0 in both models. Tyre Stress and Braking are the next least important features. Traction, Lateral, Asphalt Grip, Asphalt Abrasion, Downforce, Team, and Drivers are the features that cover medium to low importance.

The 5 most important features other than compound are Track Evolution, Circuit, Avg Air Pressure, Avg Track Temperature, and Avg Humidity. Avg Air Pressure is the second most important feature in gradient boosting and when considering the aggregate of both models.

Certainly, the feature importance does not seem very accurate when you consider the basic principles of tyre wear in Formula 1 and motorsport. They also deviate from the boxplot analysis of the categorical variables done in the exploratory data analysis.

6.2.3 RQ3 Is it possible to create a model that enables fans to have a much more knowledgeable discussion about the tyre strategy?

While the machine learning models created were not perfect, they certainly give a bot more information to the fans to have a more knowledgeable discussion about the sport. The simulations helped validate the results from models with their predictions being reflective of the actual situations in the real world. The main purpose of using open-source data was to make it accessible. FastF1 was a very important source of data and helped in making an intelligent model.

Fans of the sport enjoy speculation and discussing strategies amongst themselves before they are played out on the track. In the social media society of today, fans can engage with the community even while sitting in at their homes. For a fan sitting at home, watching the race on TV, this model enables them to predict the tyre life when a driver implements a strategy. That prediction can be used as a discussion point. Additionally, they can use this data to make more insightful speculative bets about tyre life and strategy. Potentially making it a very useful tool in that regards.

7. Conclusion

This brings us to the end of a largely successful project. There have been some learnings and weaknesses of the models have been identified. They will be presented in this section. That will be followed by a conclusion of the study.

7.1 Limitations

While creating the machine learning models and using them to predict a set of scenarios, a few limitations were identified.

The main limitations arise from the open-source data available for use. For example, while the circuit characteristics and weather conditions do play a role in the longevity of an F1 tyre, there are more variables that are not accessible to the public. Parameters like tyre degradation, tyre temperature, tyre pressures, time spent in traffic, and tyre deformations like flat spots or blisters are some of the data not available in the data source chosen. Access to these can significantly improve the predictions.

Secondly, the data taken did not account for the technical changes that happen to F1 cars all the time. In the scope of the study, from 2019 to 2023, there was a major rule change in 2022. In 2022 the design of the cars was completely changed and with it their performance. Using teams as a parameter was a way to account for the basic characteristics of a given team's car but it does not account for the technicalities of the rule changes. Even categorising the team parameter into two, pre and post 2022, could perhaps improve the predictions. In the same capacity, the size of the tyres was changed to work with the new cars in 2022. This change was also not represented in the models. The decision to not implement this was taken to keep the data dimensionality low and not introduce too many categorical variables.

The biggest limitation arises from how the data is interpreted. In this project the original data mining process gave us data where each record depicted a lap. This data was then grouped on stint to determine the stint length and all the aggregated variables were average values of that stint. More robust and complex models work on a time series interpretation of the data. They do not average the weather conditions on the stint. Rather they use the progression of each stint to make predictions. Sulsters (2019) and Heimler (2022) both have used laps wise discretised events to simulate and predict race strategies.

For example, if a tyre runs for 67 laps, each of the 67 laps and the data acquired from it is important to determine how the tyre is performing. This includes laptimes, track temperature,

and all the numerical parameters be it weather or tyre characteristics. Accurate models also consider the amount of fuel in the car because a full tank means a heavy car and that means more work for the tyres to do. A full tank also means slower laptimes. Hence, to incorporate progressive laptimes, fuel is important.

Data from practice sessions was not considered in this project, but perhaps it can be used in a project with which evaluates tyre degradation and uses lap-by-lap records. Cars typically do not go to the full potential of a tyre during the practice session, but the tyres will present their characteristics in the data. this can be used together with the race data to train the model about a tyre's characteristics. Additional circuit characteristics like number of turns, length of lap might be helpful in training the models but may lead to multicollinearity amongst the different circuit characteristics.

As the interpretation of data moves from stints to lap by lap, the handling of random events can be improved. When handling the random variables on a lap-by-lap basis, only the affected laps need to be considered, the unaffected laps can be used as normal. Suppose a random event lasted just 3 laps in a 30-lap stint. The current methodology would require the entire 30 laps to be eliminated. In a lap-by-lap methodology, the other 27 laps will be considered for the model.

Then lastly, part of the least accessible data is the internal settings each team and driver use while operating the car. This is especially tricky because unlike everything else, these settings are not standardised. All the performance parameters, even if they are inaccessible, are measured in the same way. But each team has their own settings and modes that cannot be replicated in the data of all teams.

7.2 Recommendations

This dissertation presented a look at the potential of creating machine learning models to predict tyre life in F1. This was not a very comprehensive model, and the parameters were not exhaustive even for open-source data. There were several aspects that can be improved upon from this project and used to further develop this aspect of machine learning implementation in F1.

Firstly, a more comprehensive dataset which brings in more data that can be used to predict tyre life. A better dataset and a better understanding of which parameters are influential towards tyre life. A set of input variables which where each parameter is very important in contributing to the model is a great starting point for a regression model. Increasing the scope of the study to get data from before 2019 can bring about an improvement to the results with

more training data. But for that, the post 2019 and pre 2019 data will have to be combined in a manner that the tyre compounds are compatible with both the sets. This would mean that the data would include information about the compound and not just the visible label. It can be achieved as the data is available but the data mining for it would be a little more complex. Similarly, as the years roll on, the training data will increase on the post 2019 data. And a few years down the line, we will have a much larger dataset which can potentially train the data better. Another aspect of the data set is the handling of random events and outliers. A different approach on that front can lead to different results.

Using different hyperparameter settings on Random Forest and Gradient Boosting can also be a way to improve the model, but this study did investigate the most common settings so anything else might not give the best results. For example, training a model with deep trees can lead to overfitting. Increasing the number of trees in the models beyond 200 can be helpful but that might increase the computation time of the models. Additionally, early stopping can be used in the gradient boosting model to prevent overfitting and find a generalised model. This might result in less accuracy, but it would certainly be a more robust model.

Another way to find better results is to use a more complex model than gradient boosting. Decision Tree Regression probably will not produce results more accurate than the ensemble algorithms, but something like a Lasso Regression model can. Similarly, Support Vector Regression can be an improvement upon the tree-based algorithms.

A new approach would be to classify the laps into pit windows. The teams use pit windows to create strategies, these are groups of laps where the pitstop can happen on any own. For example, there can be a pit window for laps 16-22. That's a 6-lap window. This also reflects the buffer on the predictions mentioned above. In a future study, this approach can be used to transform the entire dataset into brackets based on laps and make this a classification problem, rather than a regression problem. This will be more realistic, and the accuracy will certainly improve, albeit with different measures.

The models created can be adapted to use towards predictions in other motorsport series like Formula E and WEC. All parameters used in models are common across the motorsport industry. The availability of the data will surely be the biggest challenge in other series as most of them do not have large publicly accessible repositories like F1 does.

7.3 Final Remarks

In the end, this project was successful in creating a machine learning model that is publicly available and accessible for everyone. The model is also basic enough for more fans to

develop upon it further. This will encourage more fans to engage in insightful discussions and increase the engagement of the viewer during live races.

The models create a base understanding that can be used by F1 teams to create further models with more data available to them. The teams and the operating bodies will have more access to data that can be used to make better predictions.

Bibliography

Sicoie, H., 2022, *Machine Learning Framework For Formula 1 Race Winner And Championship Standings Predictor* (Doctoral Dissertation, Tilburg University).

Piccolomini, E. L., Evangelista, D. and Rondelli, M., The Future of Formula 1 Racing: Neural Networks to Predict Tire Strategy.

Heilmeier, A., Thomaser, A., Graf, M. and Betz, J., 2020. Virtual strategy engineer: Using artificial neural networks for making race strategy decisions in circuit motorsport. *Applied Sciences*, *10*(21), p.7805.

Rana, R., Pandey, D., Mishra, S., Nehra, N., Deshwal, D. and Sangwan, P., 2021, December. Predicting Standings in F1 Sports Driver's Championship using Lasso Penalised Regression. In *2021 International Conference on Industrial Electronics Research and Applications (ICIERA)* (pp. 1-5). IEEE.

Peng, B., Li, J., Akkas, S., Araki, T., Yoshiyuki, O. and Qiu, J., 2021, May. Rank position forecasting in car racing. In *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (pp. 724-733). IEEE.

Piccinotti, D., 2021. Open loop planning for Formula 1 race strategy identification. Master's degree Thesis, Politecnico de Milano.

Liu, X. and Fotouhi, A., 2020. Formula-E race strategy development using artificial neural networks and Monte Carlo tree search. *Neural Computing and Applications*, *32*, pp.15191-15207.

AbdulRazzaq, A.A., Fadhel, M.A. and Al-Shamma, O., AN INTELLIGENT TIRE WEAR PREDICTION USING PARALLEL HARDWARE ARCHITECTURE.

Heilmeier, A., Graf, M. and Lienkamp, M., 2018, November. A race simulation for strategy decisions in circuit motorsports. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 2986-2993). IEEE.

Dhanvanth, S., Rajesh, R., Samyukth, S.S. and Jeyakumar, G., 2022, September. Machine Learning-Based Analytical and Predictive Study on Formula 1 and Its Safety. In *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2021* (pp. 257-266). Singapore: Springer Nature Singapore.

West, W.J. and Limebeer, D.J., 2020. Optimal tyre management of a formula one car. *IFAC-PapersOnLine*, *53*(2), pp.14456-14461.

Liu, X., Fotouhi, A. and Auger, D., 2022. Application of advanced tree search and proximal policy optimization on formula-E race strategy development. *Expert Systems with Applications*, *197*, p.116718.

Heine, O.F.C. and Thraves, C., 2023. On the optimization of pit stop strategies via dynamic programming. *Central European Journal of Operations Research*, *31*(1), pp.239-268.

Liu, X., Fotouhi, A. and Auger, D.J., 2021. Formula-E race strategy development using distributed policy gradient reinforcement learning. *Knowledge-Based Systems*, *216*, p.106781.

Lasrado, S., "Predicting winners in the Formula 1 car racing season" (2021). Symposium on Undergraduate Research and Creative Expression (SOURCE). 974.

Sulsters, C. and Bekker, R., 2018. Simulating formula one race strategies, Master's Research Paper, *Vrije Universiteit Amsterdam*.

Piccinotti, D., Likmeta, A., Brunello, N. and Restelli, M., 2021, "Online Planning for F1 Race Strategy Identification", In *International Conference on Automated Planning and Scheduling (ICAPS)* (p. 5126).

Heilmeier, A., Graf, M., Betz, J. and Lienkamp, M., 2020. Application of Monte Carlo methods to consider probabilistic effects in a race simulation for circuit motorsport. *Applied Sciences*, *10*(12), p.4229.

Heilmeier, A.M., 2022. *Simulation of Circuit Races for the Objective Evaluation of Race Strategy Decisions* (Doctoral dissertation, Technische Universität München).

Massaro, M. and Limebeer, D.J.N., 2021. Minimum-lap-time optimisation and simulation. *Vehicle System Dynamics*, *59*(7), pp.1069-1113.

Tulabandhula, T. and Rudin, C., 2014. Tire changes, fresh air, and yellow flags: challenges in predictive analytics for professional racing. *Big data*, *2*(2), pp.97-112.

Choo, C.L.W., 2015. *Real-time decision making in motorsports: analytics for improving professional car race strategy* (Doctoral dissertation, Massachusetts Institute of Technology).

Stoppels, E., 2017. *Predicting race results using artificial neural networks* (Master's thesis, University of Twente).

Gu, W., Foster, K., Shang, J. and Wei, L., 2019. A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, *130*, pp.293-305.

Wang, M.X., Huang, D., Wang, G. and Li, D.Q., 2020. SS-XGBoost: a machine learning framework for predicting newmark sliding displacements of slopes. *Journal of Geotechnical and Geoenvironmental Engineering*, *146*(9), p.04020074.

Karunasingha, D.S.K., 2022. Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, *585*, pp.609-629.

Yoon, J., 2021. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), pp.247-265.

Rosso, G. and Rosso, A.F., 2016. Statistical Analysis of F1 Monaco Grand Prix 2016. Relations Between Weather, Tyre Type and Race Stints.

Serapiglia, A., 2018. Formula One–a database project from start to finish. *Information Systems Education Journal*, *16*(2), p.34.

Lampprecht, T., Salb, D., Mauser, M., Van De Wetering, H., Burch, M. and Kloos, U., 2019, July. Visual analysis of Formula One races. In *2019 23rd International Conference Information Visualisation (IV)* (pp. 94-99). IEEE.

Mourad, A.J., Delzell, P.J. and McCabe, P.C., 2019. Automation of data analysis in Formula 1.

Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009) The elements of statistical learning: data mining, inference, and prediction. Second edn. New York: Springer (Springer series in statistics).

Jolly, K 2018, Machine Learning with Scikit-Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python, Packt Publishing, Limited, Birmingham. Available from: ProQuest Ebook Central. [26 September 2023].

Alpaydin, E. (2021) Machine learning. Revised and updated edn. Cambridge, Massachusetts: MIT Press (The MIT Press Essential Knowledge Series).

Acharya, M.S., Armaan, A. and Antony, A.S., 2019, February. A comparison of regression models for prediction of graduate admissions. In 2019 international conference on computational intelligence in data science (ICCIDS) (pp. 1-5). IEEE.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied computing and informatics.*

Haghighat, M., Rastegari, H., Nourafza, N., Branch, N., & Esfahan, I. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal.*

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5), e1380.

Jenkins, M., & Floyd, S. (2001). Trajectories in the evolution of technology: A multi-level study of competition in formula 1 racing. *Organization studies,* 22(6)

Schröer Christoph, Kruse, F. and Gómez Jorge Marx (2021) “A Systematic Literature Review on Applying Crisp-Dm Process Model,” Procedia Computer Science, 181, pp. 526–534.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random Forest. In *International conference on information computing and applications*

Raschka, S. and Mirjalili, V. (2019) Python machine learning: machine learning and deep learning with python, scikit-learn, and tensorflow 2. Third edn. Birmingham: Packt Publishing, Limited.

Pirelli Grand Prix Preview: For example, Pirelli’s preview for the Azerbaijan Grand Prix (2023) Available: https://press.pirelli.com/2023-azerbaijan-grand-prix--preview-0/ (Accessed May 30, 2023)

Pirelli Suggested Race Strategy: For example, Pirelli's preview for the Azerbaijan Grand Prix (2023) Available: https://twitter.com/pirellisport/status/1652569402699571200/photo/1 (Accessed May 31, 2023)

Formula One Database API Available https://ergast.com (Accessed June 05, 2023)

F1 Live Laps, Timing, and Weather Data access through FastF1 python library Available https://docs.fastf1.dev/index.html (Accessed August 02, 2023)

Formula One Corporate Website Available: https://corp.formula1.com/ (Accessed May 12, 2023)

Virtual Strategy Engineer Available: https://github.com/TUMFTM/race-simulation (Accessed June 16, 2023)

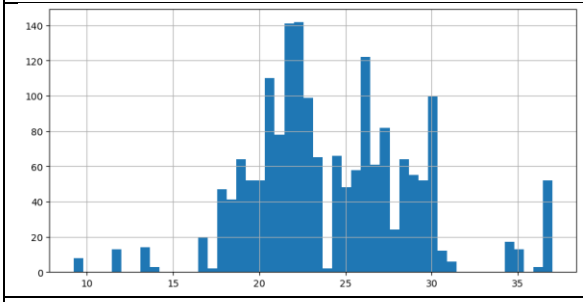
Appendix

1. Descriptive Statistics for Numerical Data

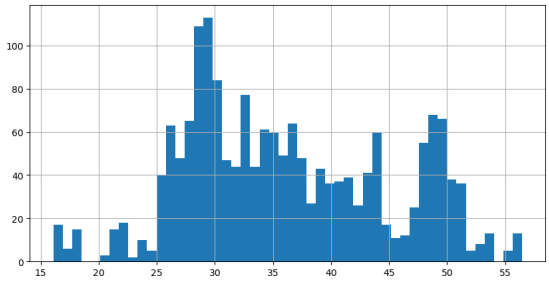
Details	avg_airtemp	avg_humidity	avg_airpressure	avg_tracktemp
count	1788	1788	1788	1788
mean	24.199352	51.631621	986.192979	36.144517
std	4.736915	17.36254	46.015542	8.773392
min	9.228571	5.714286	779.876191	16.078571
25%	21.01625	38.670565	976.916377	29.145865
50%	23.252619	55.988889	1001.19337	34.697917
75%	27.179762	63.118466	1013.49458	43.261607
max	37.009524	87	1023.17692	56.457143

2. Univariate Data Visualisations

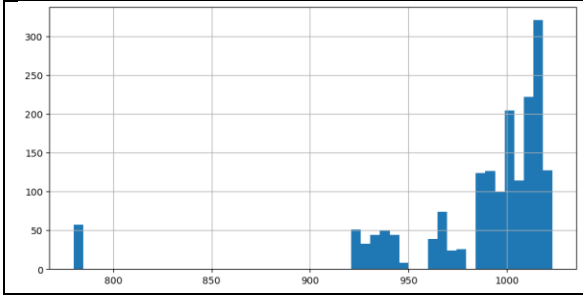
Average Air Temperature



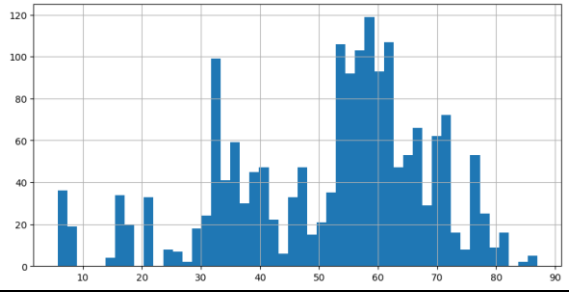
Average Track Temperature



Average Air Pressure

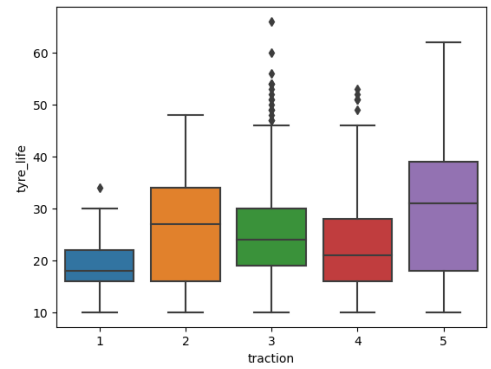


Average Humidity

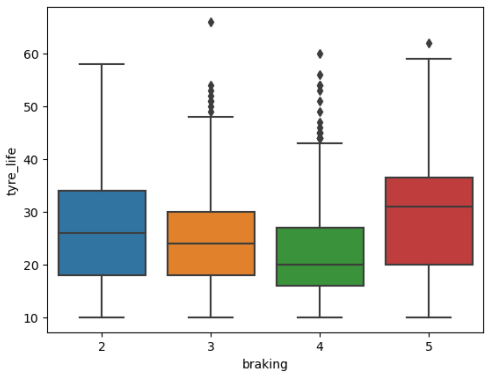


3. Bivariate Data Visualisations

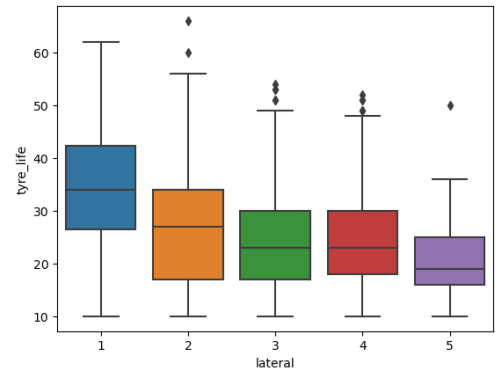
Traction



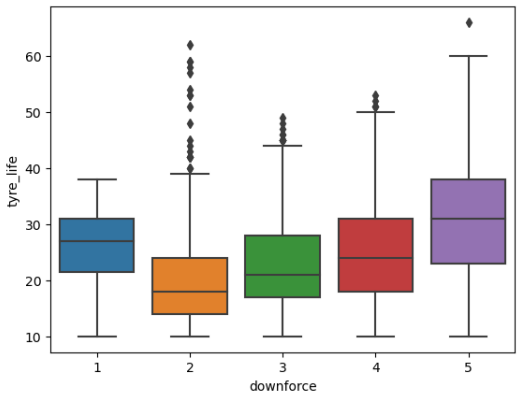
Braking



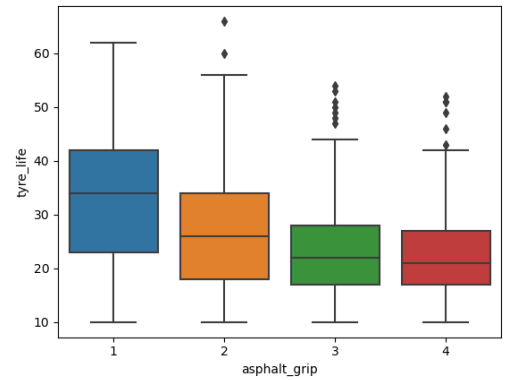
Lateral



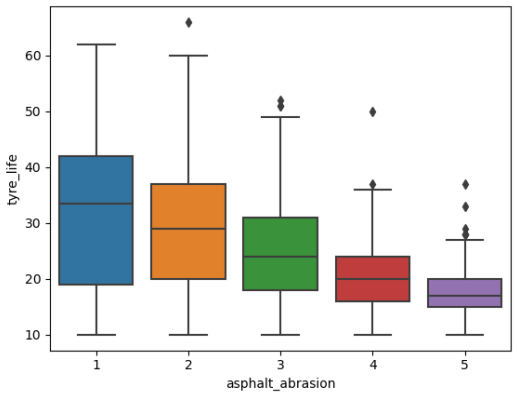
Downforce



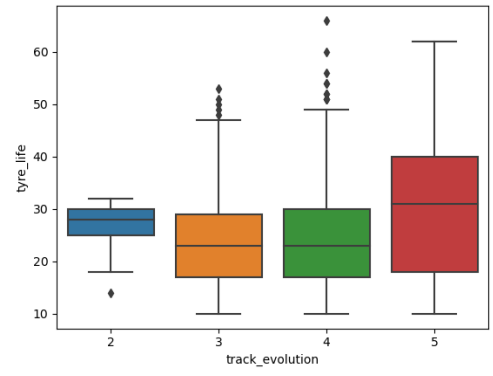
Asphalt Grip



Asphalt Abrasion



Track Evolution



Tyre Stress