Data-Driven Insights in Formula 1 using Machine Learning

Keane Coutinho
Department of Computer Science
Birla Institute of Technology and
Science Pilani, Dubai Campus
Dubai, United Arab Emirates
f20210080@dubai.bits-pilani.ac.in

Abstract

Formula 1 racing is a fast-paced, highly competitive sport in which the success of drivers and teams is decided by a wide range of criteria. This paper provides a thorough examination of predictive modeling in the context of Formula One racing, with a focus on the following essential aspects: predicting whether a driver will finish on the podium, score points or not score points. We hope to disentangle the subtle links between these parameters and the likelihood of a driver winning a victory by using a large dataset that includes previous race data, circuit data, driver, and constructor information. Explore classification models' capacity to predict whether a driver would finish on the podium, and earn points or not earn points in a race using Logistic Regression, Random Forest, Gaussian Naive Bayes, and Decision Trees. This forecast is critical for assessing driver performance during a race and optimizing team strategies. Our findings not only enhance the science of predictive modeling in Formula One racing but also provide insights that can help teams and fans make data-informed decisions. Our findings provide a better understanding of the elements that influence race outcomes, which can be invaluable in the high-stakes world of Formula 1.

Keywords: Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, K-Nearest Neighbors, Formula 1, Winner Prediction

1. INTRODUCTION

Formula One, usually known as F1, is the highest level of international racing for open-wheel single-seater formula racing cars sanctioned by the Fédération Internationale de l'Automobile (FIA). The FIA Formula 1 World Championship has been considered one of the world's finest forms of racing since its inception in 1950. The formula in the term alludes to the set of rules that all team cars must obey. A Formula One season consists of multiple Grand Prix races. Grand Prix are held on purpose-built circuits or closed public roads in a variety of countries and continents throughout the world. Grand Prix has a point system that has different points depending on where you place during the race, this points system is utilized to determine two annual world championships; one for the driver and one for the constructors. As of today, there are 10 constructors and 20 drivers for each season.

A Formula One race takes place throughout the weekend. It usually starts with two free practice sessions on Friday and one on Saturday. Additional drivers (Third Drivers) are permitted to run on Fridays, however, only two cars are permitted per team, meaning a race driver must give up their seat for that practice session for a third driver to drive. F1 has tyre rules that must be rigorously observed. Each driver can't use any more than thirteen sets of dry-weather tyres, four sets of intermediates, and three sets of wet tyres during a race weekend. The Dry weather tyres have a different range of tyres from hardest to softest, C0-C5, C0 being the hardest tyre being C0, and C5 being the softest. Each of these has its own advantages and disadvantages: soft tyres provide much more grip but lose grip very quickly and hence are beneficial towards the end of the race when positions need to be gained. On the other hand, Hard tyres don't provide that much grip but last for a very long time and hence are used when the team doesn't want to pit too early.

The qualifying session then follows the practice sessions. In Formula One, qualifying decides the starting grid for the main race, which usually takes place on Sunday. Qualifying takes place on Saturday and follows a set format designed to display the drivers' speed and ability over a single lap. The Traditional Qualifying session has a certain format which consists of three parts Q1, Q2, and Q3. In Q1, all drivers have 18 minutes to set their fastest lap times. Only the fastest 15 drivers make it into Q2, the remaining 5 drivers are eliminated. In Q2, the remaining 15 drivers have 15 minutes to set their lap times. The slowest 5 drivers are eliminated and the rest of the grid makes it into Q1. Q3 which is the final segment, here 10 drivers are battling for pole position (1st position on the grid). The Driver with the fastest time gets pole position. The key to a good qualifying session is finding a balance between using the proper tire compounds, track circumstances, and car configuration to get the most out of a single flying lap. In Formula 1, a strong starting position is critical because overtaking on some courses may be difficult. Drivers want to qualify as high up the grid as possible to have a greater chance of winning the race.

Then the third is the main event which is the race. In Formula 1, race day is the pinnacle of a weekend of preparation, strategy, and high-speed spectacle. The race begins with a warm-up lap, followed by the cars lining up on the starting grid in the order in which they qualified. This lap is also known as the formation lap since the cars lap in formation with no passing allowed. The warm-up lap allows drivers to inspect the circuit and their vehicle, allows the tyres to warm up to boost traction and grip, and allows pit crews to clear themselves and their equipment off the grid before the race begins. The race starts with a spectacular standing start. Drivers crank their high-performance engines and accelerate quickly from a stop when the lights go out. As cars compete for position entering the first turn, the tone of the race is typically determined. Pit stops are common in Formula 1 races, as teams change tires and perform essential car tweaks. Pit stops are important in racing strategy since they can affect whether a driver gains or loses positions. Teams strive to optimize performance by timing these stops. Drivers and teams make strategic decisions about when to pit, which tire compounds to use, and how to adapt to changing track conditions throughout the race. Strategy frequently develops as the race progresses, and responding to unforeseen situations, like weather changes, can be critical to victory. The safety car may be used in the event of an accident or poor track conditions. This slows the field down and can have an impact on race strategy, as some drivers may choose to pit during a safety car period to gain an edge.

The safety car, pitstop strategies, and track characteristics play a huge role in deciding the winner of the race. Safety Car can reshape the whole race's outcome; it leads to bunching up of the field as when a safety car is deployed, it slows down the entire grid allowing them to be in a single line behind it. This allows the cars lagging to catch up and gain positions after the safety car period ends. During the safety car period, drivers can choose to stay out on the track or to pit for a change of tyres to take advantage of reduced speed on the track in turn leading to less time lost due to pitstop. Pitstop strategies mainly include choosing the compound of tyres to use. For example, starting on soft tyres will give the driver more grip on the track so they can move up the grid but doing this will require the drivers to pit earlier than other drivers. Choosing the time of the pitstop is also important as sometimes pitting earlier than the driver ahead to gain an advantage is called an undercut. Similarly staying out until the driver ahead pits and staying out on track for much longer to increase the gap is called the overcut. Track characteristics are an important factor during the race. Some tracks are high-speed favoring cars with strong aerodynamics and engine power, while some of the tracks (street circuits) are more technical benefitting cars with superior handling and braking. Change in weather like it starts raining all of a sudden requires cars to pit stop to change tyres this can change the whole dynamic of the race. Changes in temperature during the race may affect the grip level of different tyre compounds.

Machine learning is reshaping the Formula 1 world, providing teams with a formidable tool for gaining a competitive edge in different parts of the sport. With F1's high stakes and ever-changing nature, any advantage, no matter how minor, might mean the difference between winning and losing. It can also help improve the performance of a Formula 1 car. Algorithms examine massive amounts of telemetry data acquired during testing and racing to uncover patterns and insights that might lead to better setup, suspension tuning, and aerodynamics. Tire wear and degradation management is a critical part of Formula 1. Machine learning models anticipate tire performance under a variety of scenarios, assisting teams in making informed decisions about tire strategy during races, such as when to pit for fresh rubber and which tire compounds to deploy. Optimizing fuel and energy consumption with Formula 1 moving to hybrid power units is critical. Teams rely on machine learning-driven simulations to model various situations, allowing them to predict how changes in car configuration, strategy, and weather conditions may affect race outcomes. This enables teams to make data-driven decisions without the requirement for real-world experimentation. Helps to accelerate R&D by automating the examination of wind tunnel and CFD (Computational Fluid Dynamics) data. This accelerates the design process, allowing teams to develop creative aerodynamic solutions faster. Machine Learning models are taught to predict race outcomes based on historical data and present race conditions. These projections can help teams make real-time decisions, such as whether to try for an overtaking or retain a position.

2. LITERATURE REVIEW

Pudaruth *et al.* [1] In February 2017, the goal is to look into the influence of Champ De Mars in identifying race winners by using odds data at different phases of the race. The study uses a dataset obtained from the Mauritius Turf Club, which includes all races held in 2014 and odds information at various moments during the races. The approach used in this study includes three unique classes: rank, margin, and win/lose. 232 of the total 347 races were utilized to train the model, with data being entered into a neural network enabled by NeuroXL Predictor software. The model's performance was then evaluated using a testing set of 27 races using an Artificial Neural Network (ANN) as the method of choice.

The findings of the algorithm show that the ANN outperforms other regularly used approaches such as Logistic Regression, decision trees, and Support Vector Machines (SVM). Despite its success, this strategy has significant drawbacks. The model's reported accuracy is significantly lower than a random chance prediction accuracy of 11.4%. This indicates a significant drawback of the study, as reaching an accuracy rate of 7.4% implies that the model's predictive capacity is rather limited. This literature analysis emphasizes the need for additional study and model improvement to improve prediction capabilities for predicting race winners based on odds data and Champ De Mars participation. Furthermore, a deeper investigation into the reasons for the ANN's decreased accuracy, as well as a comparison with other viable predictive models, may be beneficial in understanding and predictive capabilities in the field of horse racing.

Chung *et al.* [2] in August 2017, wanted to address the difficult challenge of predicting horse racing winners, placers, and showings in order to provide a credible basis for betting decisions. To accomplish this, the researchers used a novel approach, building a committee machine out of Support Vector Machine (SVM) models. The Hong Kong Jockey Club (HKJC) provided the dataset for this study, which ran from January 1, 2012, to June 30, 2015, and included a complete collection of 2691 race records and 33532 horse records. Horse win rate, Trainer win rate, Jockey win rate, Actual weight, Fastest Finish time, and other parameters were methodically extracted from this information. The data from January 1, 2012, to June 30, 2012, was used to train the forecast model, while the remaining dataset was reserved for testing purposes.

The employment of a Committee Machine as the principal prediction method was the study's main novelty. Individual machine learning algorithms such as SVM and Random Forest were dramatically beaten by the Committee Machine, demonstrating the value of merging many models for more accurate predictions. Nonetheless, one significant disadvantage was the Committee Machine's sensitivity to the value of T, which indicated the quantitative difference between the first and second horses. The accuracy varied significantly depending on the T value, ranging from 35.84% to a much more promising 70.86%. This sensitivity to a single parameter could be viewed as a weakness, necessitating additional research into its ideal configuration. Despite this constraint, the committee machine's overall performance was encouraging, with a 70.86% accuracy rate in forecasting horse race outcomes, which has considerable potential for improving betting tactics and decision-making in the domain of horse racing.

In 2021 Ga Yau *et al.* [3] emphasized the application of neural networks in forecasting winning horses based on limited historical data. The issue statement revolves around the challenge of developing predictive models for horse racing outcomes, which is often a data-scarce domain. The raw data used in this study ranges from 2003 to 2020 and includes information on races, horses, courses, jockeys, and trainers. It contains a wealth of features that are essential for training a model to predict horse race winners. To solve this issue, the researchers used a feature engineering process on the dataset. Notably, they split the data into two halves, with races from 2003 to 2018 acting as training data and events from 2018 to 2020 serving as testing data. The neural network design adopted, DH_v1, consists of two hidden layers with 60 units each and a dropout probability of 0.5. While this method represents a determined effort to use machine learning techniques to forecast horse racing outcomes, numerous parts of the methodology deserve greater examination. It is necessary to go into the feature engineering process in order to assess the extent to which hand-engineered features improve model performance, and the specific architecture should be discussed.

The claimed performance metric of 11.1% accuracy suggests the model's potential in the area of horse racing prediction. However, the limitations and implications of this accuracy rate must be investigated further in order to assess its utility in real applications.The capacity of neural networks to exceed random selection and public odds estimates is one of the benefits of using them in this case. This has enormous ramifications for bettors and horse racing fans looking for data-driven insights. However, as previously stated, the model's shortcoming is its inability to learn new past-performance-based features. This constraint makes it difficult to react to changing conditions or incorporate new data sources, which is crucial in a dynamic sector like horse racing. To summarize, while the study represents a commendable effort in utilizing neural networks for horse racing prediction, it requires a thorough examination of its methodology, model architecture, and practical implications in order to assess its viability as a predictive tool in the horse racing domain.

In 2017, Hoffmann, M. [4] addressed the problem of estimating the time required to complete a triathlon while accounting for data normalization owing to variances in individuals' anthropometric and physiological features. The dataset used spans four years, from 2008 to 2012, and includes data from 23 German male triathletes who competed in national or international championships. However, only 11 people were considered for the study. The dataset included both anthropometric and physiological information. The researchers wanted to know how body mass index (BMI), lean body mass, thoracic depth, body fat percentage, body fat in kilograms, pelvic width, body height, seat height, shoulder width, and thorax breadth affected triathlon completion times. The study used a mix of Dominance Paired Comparison (DPC) for data preparation and normalization, as well as two algorithms: Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN) to evaluate the data.

Data standardization was a vital stage in the approach to ensure that race timings could be compared fairly, as it is necessary to account for the intrinsic disparities in the anthropometric and physiological profiles of the triathletes. The process relied heavily on the selection of features, which included numerous anthropometric measurements. Before being fed into the machine learning models, the data was treated to the Dominance Paired Comparison approach. The use of Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN) enabled the development of predictive models, with the ANN outperforming MLR in capturing non-linear correlations and building more complex models. However, the study did have several shortcomings. The sample size was tiny, with only 11 triathletes participating, which may have hampered the findings' generalizability. It was suggested that future research efforts use larger and more diverse datasets to increase the forecasting accuracy of the algorithms.

In terms of performance assessment, the study used R-squared (R2) values to examine the prediction capacities of the MLR and ANN models. These R2 values provide information about the models' goodness of fit for anthropometric and physiological variables. The R2 for anthropometric factors was 0.41 and 0.67 for physiological variables for the MLR model. The ANN model, on the other hand, outperformed the MLR model, with R2 values of 0.43 for anthropometric factors and 0.86 for physiological variables. This result demonstrates the effectiveness of Artificial Neural Networks in capturing the complexities and nuances of the relationships between anthropometric and physiological data and triathlon completion times, while also emphasizing the need for future studies to include larger sample sizes for more robust and accurate results.

In July 2021 Wu, P.P.-Y. *et al* [5] aim to solve a major issue in competitive swimming: the prediction and comprehension of changes in swimmer performance between individual and relay races, with a particular emphasis on the 4x200m freestyle relay. The study makes use of a large dataset that includes race data from 14 international long-course swimming competitions held between 2010 and 2018, including prestigious events like the Olympic Games, Pan Pacific Championships, World Championships, Commonwealth Games, and European Championships. This huge dataset serves as the research's foundation, providing a rich supply of information for investigating the performance dynamics in relay events.

The methodology used to fulfill the study goals entails rigorous feature engineering and the incorporation of extra data, most notably the fastest 200m freestyle time for the season attained by each of the participating swimmers. In addition to individual swimmer data, the study computes team ranking, which is important in relay races. The prediction work categorizes relay teams into Gold, Silver, Bronze, or non-medal positions, and the predictive analysis is performed using Random Forest, a prominent machine-learning technique. The achievement of 100% accuracy in predicting Gold is an obvious advantage of our approach, which could prove beneficial to coaches in their decision-making processes. However, a significant drawback is the limitation to teams with data for all four swimmers, which leads to partial data for certain races and consequently leads to misclassification.

The study's findings are especially intriguing, with the Random Forest model displaying outstanding predictive performance. The model forecasts Gold with 100% accuracy, which has enormous consequences for coaches and team management. Furthermore, sensitivities of up to 41% for Silver, 63% for Bronze, and 93% for non-medal places are achieved by the approach. This sensitivity analysis demonstrates the model's ability to discern between different degrees of performance within relay events. The study revealed a surprising finding: slower individual swimmers can actually perform extraordinarily well in the relay format, implying a novel and potentially game-changing technique for relay event preparation and selection. Overall, this study adds vital knowledge and prediction capacities to the world of competitive swimming by putting light on nuances of relay performance and offering a practical tool for coaches and teams to enhance their decision-making processes.

In 2020 Liu X *et al.* [6] Investigated the usage of Artificial Neural Networks (ANN) and Monte Carlo tree search as predictive techniques for optimizing Formula E racing strategy. The fundamental issue statement revolves around utilizing these technologies to forecast Formula E car performance, allowing for the development of viable racing strategies. The dataset used in this study was painstakingly generated by combining IPG/Carmaker with a Matlab model. It included characteristics such as drive power, regenerate power, torque, and environmental temperature changes, providing a full picture of their impact on crucial performance metrics such as lap duration, battery state of charge, and battery temperature.

ANN was used in two different ways in terms of methodology. The first method involved developing a three-output neural network, but the second involved developing three independent neural networks, each dedicated to one of the important performance indicators. These ANN models were capable of predicting single-lap performance as well as assessing energy consumption. The Monte Carlo tree search technique, on the other hand, was used to evaluate a wide range of racing scenarios, including pre-race planning, aggressive driving strategies, scenarios including the usage of 'Attack mode,' situations involving the safety car, and variations in environmental conditions. This strategy was a great tool for strategic planning as well as reacting quickly and efficiently to unexpected circumstances during a race.

In terms of benefits and drawbacks, ANN models proved to be a significant advance over classic lap simulation models, offering high predicted accuracy. The 3-output MSE was 0.0215, and the 1-output MSE was 0.0108. Nonetheless, the Monte Carlo tree search technique has an accuracy constraint. Regardless, it demonstrated its importance in influencing pre-race plans and providing the ability to quickly respond to unforeseen events during races. In the competitive realm of Formula E, this study illustrates the potential of employing advanced AI techniques such as ANN and Monte Carlo tree search to improve race strategy and performance predictions.

In 2022, Grover, R *et al.*[7] focused on the key subject of whether a driver's car setup and driving style should be primarily tuned for optimal race performance or built to excel in both qualifying and race sessions. The study makes use of a large dataset derived from the Fast-F1 library, an open-source Python tool that allows access to Formula 1 timing data from 1950 to the present. This dataset is organized as Pandas DataFrames and series, allowing for extensive exploration and analysis.

The study's approach focuses on the qualifying and final positions of five drivers. A special emphasis is given to drivers who secured the first Pole position in all races over the 2019-2021 seasons. Aside from these fundamental indications, the research takes into consideration additional variables including weather, safety vehicle interventions, and other influencing factors. To determine the significance of qualifying in relation to race performance, the dataset is exposed to sophisticated data analysis techniques and numerous graphical representations.

One major quality of this study is its thorough evaluation of multiple factors, including as weather conditions and pit stop entrance timings. However, there is a limitation in that only a restricted number of drivers—those rated as the greatest on the grid—are evaluated. The findings of the study emphasize the significance of qualifying in the context of race performance. Nonetheless, it is clear that various other factors, like as safety car deployments, constantly changing weather conditions, and strategic pit stop timings, play critical roles in determining the race's conclusion. As a result, the study reveals that a well-rounded car configuration is critical for ensuring success in both qualifying sessions and subsequent races, which is consistent with the assumption that race-winning cars should have a balanced setup.

In 2023, Patil. A *et al.* [8] focused on a data-driven approach to determining the primary elements influencing each driver's total points scored in Formula 1 racing. The major goal of this research is to use Principal Component Analysis (PCA) to identify the most important components to a driver's performance. The dataset for this analysis was obtained by web scraping from racefans.net and spans five years, from 2015 to 2019. The dataset includes essential statistics such as the average number of pit stops, tire-related data, the number of laps led by the driver, penalties incurred, and overall points won.

The research methods used can be broken down into various steps. Data pre-processing was initially performed to remove redundant and extraneous information. Missing data was then addressed, ensuring that the dataset was comprehensive and dependable. Principal Component Analysis was used to further evaluate the interrelationships between the attributes. The study not only assisted in finding the most important main components, but it also indicated the amount to which they captured variance. Following PCA, a linear regression model was built using web-scraped data from the aforementioned five-year period. The total points scored by each driver were the model's dependent variable. The advantage of this technique is its ability to differentiate the relative importance of numerous parameters contributing to the overall performance of the drivers. However, one weakness of this study is that it only looks at data over a five-year period, which may not account for long-term trends or changes in Formula 1 racing dynamics.

The PCA and linear regression models performed admirably. The total variance explained by the first four main components indicates a substantial association between most race-related characteristics. PC1 emphasizes the value of statistics like as Average Pole Position and laps led, highlighting their significant impact on driver performance. PC2 explains the significance of factors such as Average Pitstop Time and tire-related data. PC3 dives deeper into tire-related issues, giving light on their importance in driver success. The importance of starting positions, accidents, and penalties received by drivers during races is emphasized in PC4. This study also includes comparable estimates and p-values for each feature via linear regression analysis, providing valuable insights into the magnitude of their influence on the total points.

In 2023, Keertish Kumar *et al.* [9] investigated the use of classification models to forecast the performance of both drivers and constructors in Formula One racing, with a specific emphasis on evaluating the efficacy of various machine learning methods. This study's dataset consists of numerous CSV files supplied from a GitHub repository, including "Races," "Results," "Driver," and "Constructors," with features such as race information, race results, driver details, and constructor information. The basic goal of this research is to evaluate classification model performance and decide which algorithm produces the best results.

The suggested methodology is divided into three stages: data pre-processing, machine learning model training, and performance evaluation. To clean, transform, and prepare the dataset for analysis, it is subjected to extensive pre-processing. To develop prediction models, the authors use a variety of classification techniques, including Logistic Regression, Decision Tree, Random Forests, Gaussian Naive Bayes, K Nearest Neighbor, and Support Vector Machine. The accuracy of these models is evaluated to shed light on their effectiveness in predicting driver and constructor performance.

One major aspect of this study is its thorough examination of numerous parameters important in predicting the success of Formula One drivers and constructors. However, the study uncovers a fundamental limitation: when separate models for drivers and constructors are constructed, the constructors' model consistently beats the drivers' model. This disparity indicates a probable lack of data or relevant features for effectively predicting driver performance, implying the necessity for additional research or the inclusion of new driver-specific factors in future investigations. The classification models performed differently in terms of accuracy, with Random Forests and Support Vector Machine achieving the greatest accuracy rates of 92.5% and 92.1%, respectively, and Gaussian Naive Bayes achieving the lowest accuracy of 86.5%.These findings shed light on the potential of machine learning algorithms to anticipate Formula One race outcomes, with effects for both constructors and drivers.

2.2 Problem Statements

- Taking into consideration driver confidence
- Age of the driver is also taken into account.
- Constructor Reliability considered.

3. METHODOLOGY

3.1 Dataset Description

The Formula 1 World Championship (1950-2023) [10] from Kaggle consists of various CSV files which include circuits.csv, constructor_results.csv, constructor_standings.csv, constructors.csv, driver.csv, driver_standings.csv, lap_times.csv, pit_stops.csv, qualifying.csv, races.csv, seasons.csv, results.csv and status.csv. From these 8 CSV files were used, including circuit.csv which contains information about the circuit like country, latitude, and longitude, race.csv gives the details of all the races ever done like the date, circuit, FP1, FP2, and qualifying times. Drivers.csv gave information about the drivers like their name, car number, date of birth, and nationality. Constructor.csv tells us which constructors have taken part in F1 and their Nationalities. Status.csv contained possibilities for races like Finished, Disqualified, Engine Failure, Puncture, and Collision. Results.csv details the race results like the start and end position and number of points earned.

3.2 Experimentation

The Architecture of Implementation is described in Fig 1.

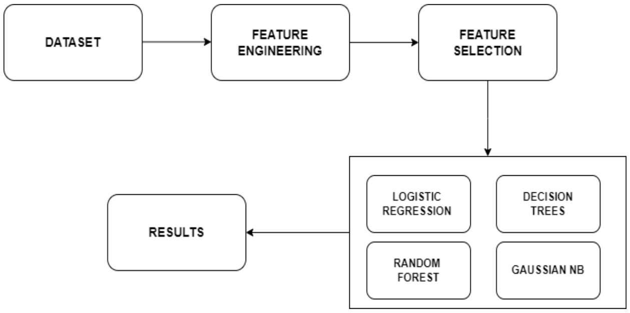


Fig 1. Architecture Diagram

(i) Data Extraction and Pre-processing

Since the dataset consists of multiple CSV files we combined all the following CSV's which are in the form of data frames races.csv, results.csv, qualifying.csv, drivers.csv, constructors.csv, and circuits.csv. Following this, all the columns whose data wasn't required have been dropped. Then only races after 2015 have been considered as the dataset includes data from 1950 and many teams that aren't as good right now will be shown as good. Since the team names have changed over time all the old team names have been updated to the current team names so there is a mismatch between the data. All the columns country-related columns have been converted to their short form.

(ii) Feature Extraction

The first feature extracted from the dataset was the age of the drivers in days this feature plays an important role as many younger drivers still have a long way ahead for development and improvement meanwhile the older drivers have much more experience and maturity, on the other hand, physical fitness plays a key role since F1 is physically demanding and younger drivers have an advantage in terms of endurance and recovery. The dataset then had the driver's forename and surname separately mentioned as two features it was combined to form one feature. For the combined data frame every time a driver Did Not Finish (DNF) there was a status ID present which gave information as to why the driver DNF so for every DNF two binary columns were created Driver_DNF where 1 meant it was the driver's fault and a similar thing for the Constructors_DNF.

By grouping the drivers and seeing their DNFs a DNF Ratio was calculated and then used to calculate the driver's confidence as shown in Fig 2. Similarly, for the constructors a feature called Constructors_Reliability was calculated. The position column was then converted into 3 classes instead of 20 the three classes include Podium, Points, and no points.After this the dataset was reduced only to the current drivers and constructors and some other unwanted features were also dropped. Finally, the remaining features were GP_name, qualifying position, driver, constructor, driver_confidence, constructor_reliability, and age at GP in days.

$$Driver\ confidence = 1 - \frac{no: of\ DNFs}{no: of\ races}$$

Fig 2. Driver Confidence

(iii) Post-processing

After achieving all 7 columns like GP_name, driver, and constructor were transformed using label encoding. The dataset was tested among different models which include Logistic Regression, Decision Trees, Random Forrest, and Gaussian NB. These methods were tested using cross-validation and 10 splits and using the scoring metric as accuracy. The front end was added to the model to make it interactive and predict if the driver ended up on the podium or in point or no points using Streamlit as shown in Fig 4 and the working of the application is shown in Fig 3.

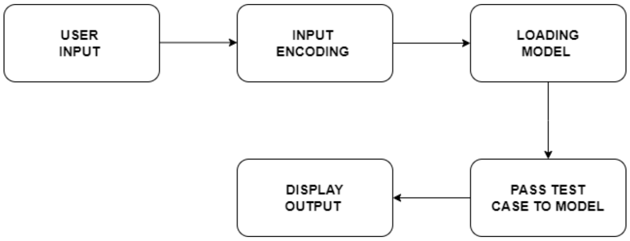
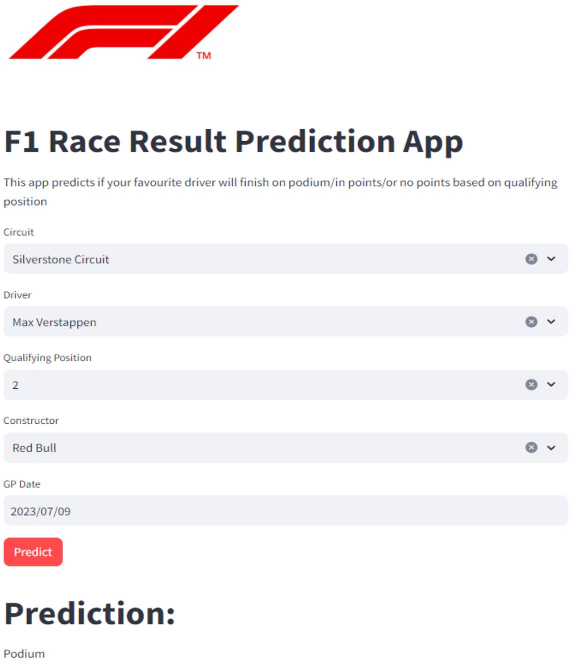


Fig 3. Working of Application



F1 Race Result Prediction App

This app predicts if your favourite driver will finish on podium/in points/or no points based on qualifying position

Circuit: Silverstone Circuit

Driver: Max Verstappen

Qualifying Position: 2

Constructor: Red Bull

GP Date: 2023/07/09

Prediction:

Podium

Fig 4. Application GUI

3.3 Models

3.3.1 Logistic Regression

Process of modeling the probability of a discrete outcome for the given input variables. Most commonly used to predict a binary outcome like true/false or yes/no. Multinomial logistic regression can model scenarios where there are more than 2 classes.

3.3.2 Decision Trees

Decision Trees, are a non-parametric supervised learning method used for classification but can also be used for regression. The model is created by learning simple decision rules inferred from features of the dataset.

3.3.3 Random Forest

A Machine Learning algorithm that combines the output of multiple decision trees to reach a single result. It can be used for both classification as well as regression. It acknowledges feature bagging which usually causes low correlation between decision trees.

3.3.4 Gaussian NB

Is a classification technique used in machine learning based on the probabilistic approach and Gaussian distribution. It also assumes that each feature has an independent capacity for predicting the result.

4. IMPLEMENTATION AND EVALUATION

4.1 Implementation

In our experiment, the minimum requirements are Windows 10 or more as an operating system and Google Colab or Jupyter Notebook as an integrated development environment (IDE). Python modules include numpy, pandas, matplotlib, pickle, seaborn, and Joblib.

4.2 Evaluation Metrics

This paper explores logistic regression, random forest, decision tree model, and Gaussian NB to predict where the driver will place at the end of the race. Following are the benchmarks used to measure the performance of the compiled models.

Accuracy score: Accuracy is the proportion of correct predictions made by a machine learning model over total predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision can be viewed as the proportion of samples that were correctly classified as a specific class over the total number of predictions of the class made by the model.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall is the measure of correctly classified samples of a class over the total number of samples present from the class.

$$Recall = \frac{TP}{TP + FN}$$

F1 score:] be the harmonic mean of precision and recall. Because the F1 score integrates recall and precision into a single metric, it is a little less intuitive.

$$F1 \quad 2 \times \text{Precision} \times \text{Recall}$$

5. RESULTS

Using 10-fold cross-validation on 4 models based on Logistic Regression, Decision trees, Random Forrest, and Gaussian NB, based on accuracy the scores were 83.75%, 88.93%, 92.65%, and 82.76% respectively. Since Random Forest performed the best feature importance and the confusion matrix as a heatmap is shown in Fig 5 and Fig 6.

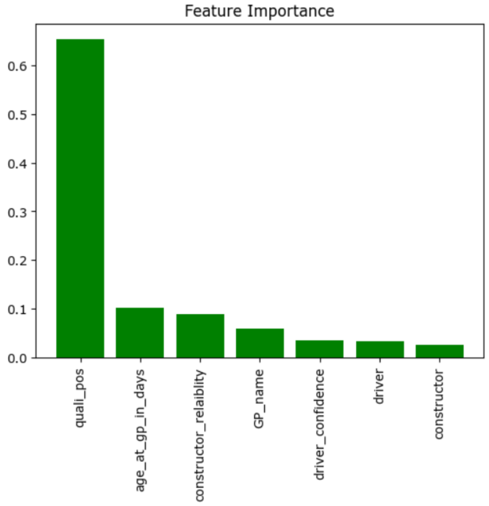


Fig 5. Feature Importance for Random Forest

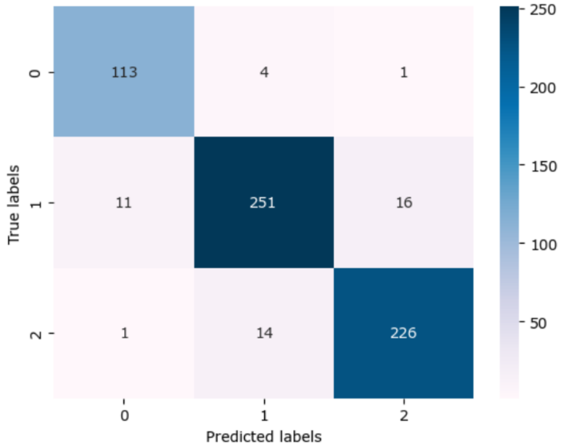


Fig 6. Confusion Matrix for Random Forest

6. LIMITATIONS

The dataset used gave information about the drivers, circuits, constructors, races, and reasons for crashes but whether the driver will get podium, points, or no points depends on various other factors like safety cars, Red flags, weather, number of pitstops, time take for a pitstop during the race. The weather wasn't considered during this experimentation but weather plays a crucial role as according to weather conditions the tyres are chosen and the race pace reduced, especially during rain the drivers are more prone to losing control and crashing leading to more safety cars, yellow, and red flags as well.

The number of pitstops and the time taken for a pitstop matters and in F1 everything comes down to milliseconds sometimes front wing needs to be changed or the pitstop mechanics are slow leading to loss of positions. Here driver confidence and constructor reliability is taken into consideration but each year a new car is built and sometimes one team has a faster car than the other, some cars have faster straight-line speed so certain circuits are favoured. Also at times, some cars may be faster than others but they have higher degradation leading to more pitstops so car details have to be considered as well.

7. FUTURE SCOPE

The current model only takes into account a few features that determine the results of a race. The future iterations can include more features like pitstop times and number of pitstops fastest lap times for a circuit, average lap time for each driver on each circuit, number of safety cars during the race, and weather conditions as well. Also, include information about creating a car_power attribute based on the specifications of the car. This would increase the complexity of the model but would allow it to capture more complex relationships and dependencies.

8. CONCLUSION

The study was conducted in the following series of steps. First, the data was sourced from Kaggle consisting of information on all past races, circuits, drivers, constructors, etc. The dataset was then combined into one Dataframe and then moved on to feature selection and removing unwanted features. The data was then pre-processed and valuable features were extracted from this, like age of drivers in days, driver confidence, and constructor confidence. Then the target class (position) was converted from a 20-class column to a 3-class column. The features were then passed through 4 Machine learning algorithms which include Logistic Regression, Decision Trees, Random Forest, and Gaussian NB. Random Forest had the best accuracy among the 4 models with an accuracy of 92.65 %. The paper also suggests more ways to increase the accuracy of the model by considering different features.

REFERENCES

[1] Pudaruth, S. (2017) ‘Impact of the variation of horse racing odds on the outcome of horse races at the Champ de Mars’, *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*.

[2] Chung, W.-C., Chang, C.-Y. and Ko, C.-C. (2017b) ‘A SVM-based committee machine for prediction of Hong Kong horse racing’, *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)*

[3] Torné, Olaf. “Ga Yau: Machine analysis of Hong Kong horse racing data.” (2021).

[4] Hoffmann, M. *et al.* (2017) ‘Predicting Elite Triathlon Performance: A comparison of multiple regressions and artificial neural networks’, *International Journal of Computer Science in Sport*, 16(2), pp. 101–116. doi:10.1515/ijcss-2017-0009.

[5] Wu, P.P.-Y. *et al.* (2021) ‘Predicting performance in 4 x 200-m freestyle swimming relay events’, *PLOS ONE*, 16(7). doi:10.1371/journal.pone.0254538.

[6] 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(18), pp. 15191–15207. doi:10.1007/s00521-020-04871-1.

[7] Grover, R. (2022) ‘Analysing the importance of qualifying in Formula 1 using the FASTF1 library in Python’, *International Journal of Advanced Research*, 10(08), pp. 1138–1150. doi:10.21474/ijar01/15280.

[8] Patil, A., Jain, N., Agrahari, R., Hossari, M., Orlandi, F., Dev, S. (2023). A Data-Driven Analysis of Formula 1 Car Races Outcome. In: Longo, L., O’Reilly, R. (eds) Artificial Intelligence and Cognitive Science. AICS 2022. Communications in Computer and Information Science, vol 1662. Springer, Cham. https://doi.org/10.1007/978-3-031-26438-2_11

[9] Keertish Kumar, M. and Preethi, N. (2023) ‘Formula One race analysis using machine learning’, *Lecture Notes in Networks and Systems*, pp. 533–540. doi:10.1007/978-981-19-6088-8_47.

10] Vopani (2023) *Formula 1 World Championship (1950 - 2023)*, *Kaggle*. Available at: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020 (Accessed: 13 December 2023).