Formula One Race Winner Predictor using Machine Learning

Savan Patel

Thesis submitted in the fulfillment of the Distinction Requirement for the degree of Bachelor of Science

in the

Department of Computer Science and Engineering
The School of Computing
University of Nebraska - Lincoln

Student Name
Savan Patel

Student Number
77380421

Committee
Dr. Stephen Cooper
Dr. Justin Firestone

Location
256 Avery Hall
Lincoln, NE 68588-0115

Date
2023-04-15

Acknowledgements
I would like to acknowledge my thesis advisor Dr. Stephen Cooper and Dr. Justin Firestone for guiding throughout my thesis research and providing me with valuable advice.

Formula One Race Winner Predictor using Machine Learning and Artificial Intelligence

Savan Patel

Abstract

In the last few decades, the field of machine learning and artificial intelligence has shown very impressive growth. Although, in the area of sports analytics we still have a long way to go. Introducing machine learning and artificial intelligence to analyze sports data can be beneficial for not only the sports business industry but the analytics can be used for future performance enhancement. This research solely focuses on Formula One or F1 race analytics. Research begins with the summary of a few studies done in the past on the historical data of F1 since 1950. There were many changes in the engine regulations since then which can distort the findings of the data. Due to that reason, this research will be based on the data after 2014 when *the turbo hybrid era* of F1 began. The content of this paper includes details on machine learning frameworks and methods used to analyze the race data and the result of the predictions. This can serve as the foundation for future research improvements and can provide some good insights on the F1 race analytics.

1. Introduction

The history of Formula One is dated back to the 1930s, however it didn't emerge until 1950 when it was first introduced as Formula A. Since then, the committee has been diligent in maintaining all the data of the race which includes weather conditions, racers, cars, engines, race positions, finish position, lap time, and more. All the races have a set of rules and regulations that needs to be followed by participants. These rules also specify the type of car or engine that can be considered part of the race. As we all know, F1 is one of the highest class single seater and most popular races of all times. Thus, it was inevitable that it would be left behind when it comes to applying machine learning and artificial intelligence technology to it.

Formula One racing is followed by millions of people around the world and has been an integral part of many university clubs to provide students with a creative learning edge in the sports industry. This being said, the introduction of machine learning and artificial intelligence in the sports industry has brought some fascinating understanding of the sports business analytics, performance enhancement, and future result prediction.

The motivation behind this research was to investigate further on the past findings regarding F1 race winner predictions. Since, most of the research uses either all the past data available since 1950 or partial data from the late 1900s to early 2000. However, there was a major change introduced in 2014 by Formula 1 with the engine requirement. From 2014 onwards F1 race cars started using new hybrid V6 engines. This change came across due to environmental concerns caused by emission during the race and was introduced as *the hybrid era* in the F1 industry.

Deriving from the findings of the hybrid era, this research emphasized on the race data from 2014 to 2022. This will provide better analysis of the data. Scope of this research is to evaluate the winner prediction based on artificial neural network and machine learning frameworks. This research will encapsulate various data analysis techniques such as data extraction, data cleaning and preprocessing, and applying machine learning models to predict race winners. The data was divided into two different parts to train and test the models. Training dataset was kept from 2014-2021 and then the model was tested on 2022 data which is detailed further in the paper.

This research builds on top of the previous research performed by various professionals. Some of the details have been highlighted in the paper along with the findings of their study. Since the data is not always ideal, most of the study was focused on the data cleaning and preprocessing as it is common with any machine learning prediction steps. Further, there is discussion of the models used to predict the race winners based on the 2022 dataset and the methodologies used for the prediction. The structure of the paper includes graphs and tables of some of the key findings in the nature of the study.

2. Literature Review

Millions of people worldwide enjoy the popular motorsport known as Formula 1. In recent years, there has been an increase in interest in using machine learning techniques to predict the results of Formula 1 races. The purpose of this literature review is to provide an overview of some of the recent studies in this area.

In the first paper, Sicoie (2022) draws attention to the growing demand for modern software in sport analytics, particularly in the Formula 1 world. The objective of Sicoie's research is to develop, apply, and evaluate supervised learning algorithms to predict championship standings for the 2021 Formula 1 season. Sicoie's research focuses on ensemble methods and regression methods. An ensemble method is a machine learning technique that combines several base models in order to produce one optimal predictive model. The initial results were disappointing, but when the generated rankings were aggregated over the entire season, they were highly correlated with the ground truth. The study found that there is still progress to be made, including algorithmic fine-tuning, the addition of new features, and exploring using different deep learning techniques. This research builds on earlier research and analysis to improve prediction modeling in F1.

In the second paper, Stoppels (2017) compares various prediction techniques and focuses on using ANNs (Artificial Neural Networks) to predict Formula 1 race results. The author created a relatively simple neural network with only 5 layers. The study finds that ANNs perform better than multiclass logistic regression for predicting race, but that performance depends on the sample size and network structure used. The study concludes that the number of samples and the network structure refinement affect how well ANNs predict outcomes.

The study by van Kesteren & Bergkamp uses Bayesian multilevel Beta regression to model the proportion of drivers beaten in Formula 1 races of the hybrid era (2014-2021). Van Kesteren & Bergkamp (2022) focuses on separating the car performance and the driver performance to understand the driver performances. They conclude that the car is more important than the driver when it comes to race results. The authors make inferences about the drivers and constructors competing in the 2021 season. The model accurately represents changes in constructors' seasonal form, for example Ferrari developed a mid level car in 2020 and had to spend entire 2021 season to recover from it before becoming race winner contender again in 2022. The model does not work well for prediction and suggests using other approaches for forecasting. Overall, the model accurately represents driver and constructor performances in the seasons 2014-2020.

Veronica Nigro (2020) discusses the application of machine learning algorithms to Formula 1 race prediction in her article "Formula 1 Race Predictor". The algorithm is trained and tested for accuracy using a variety of data sources, including previous race results and driver statistics. The article also discusses the difficulties in predicting the results of a Formula 1 race, such as the influence of weather and unforeseen events. Overall, the article shows how machine learning might be used to predict Formula 1 race outcomes and points out potential directions for further analysis.

The last article, "Formula One: Extracting and Analyzing Historical Results," by Ciarán Cooney (2020), provides an overview of the process of extracting and studying historical data from Formula One races. The author describes the process for extracting data from the official Formula One website and converting it into a format that can be used for analysis. Cooney's article offers a valuable case study of data visualization technique exploration, as well as insights into best practices for data analysis in this field.

3. Data Collection and Preparation

3.1. Data Collection

The first step in the process of data collection is to determine the data necessary for the analysis and prediction of the race results. All of the following data was scraped and retrieved from the Ergast Developer API and the Formula 1 Results & Statistics. The tables created using the sources include data about races, results, driver standings, constructor standings, and weather.

All of the data tables mostly follow similar steps for the scrapping. The data scraping is done in Python using “requests”, “BeautifulSoup”, “Selenium”, and “Pandas”. The data is retrieved by parsing HTML from the website using BeautifulSoup. After parsing, it extracts all of the relevant information and then stores it in a pandas data frame.

There are different details in the data extraction as we look deeper into each function. For the Races.csv I start with creating a dictionary with each column as the key. The for loop then iterates over each year from 1950 to 2022 and sends an HTTP request to the Ergast API to retrieve data about all the F1 races that occurred in that year. The information is then extracted from the JSON response and appended as a list of the corresponding keys in the “races” dictionary. The possibility of missing data is handled using try-except blocks. Finally, the dictionary is converted into a data frame and stored as a “Races.csv” file. The second function retrieves results for each round of race in the “Races” data frame. It uses the exact same approach of creating a dictionary with each column as a key to store data from the API. The dictionary is then converted into a data frame and stored as “Results.csv”.

The data for the driver standings for each round is retrieved in JSON format from the Ergast API. Later, the relevant data is gathered and added to the respective list in the driver_standings dictionary. After the data is retrieved, a function called lookup is made that takes the driver standings data frame, the driver's name, and the columns that need to be merged and moved (driver_points, driver_wins, driver_standings_pos). The function adds two columns to the data frame, one for the current round and one for the previous round. In order to get the points, position, and number of victories each driver had, coming into the current round, the function compares the data on the lookup keys and merges the data frame with itself. The columns of the previous round's results are then removed, resulting in a final "DriverStandings.csv" file. Constructor standings data is collected similarly to how the driver standings data is gathered. The data is initially stored in JSON, and subsequently, the necessary data for each round is stored in the constructor_standings dictionary. The lookup function adds new columns for the constructor's points, wins, and standings from the previous round. After combining the data from the previous round with the current round, the code removes the columns from the data. The final data frame is then stored as “ConstructorStandings.csv”.

The function runs through the years 1950 to 2022 to scrape data from the website using the following link: "/en/results.html/year/races/" for the qualifying results. The link above appends at the end of the Eargast API to visit the website which contains the specific HTML table data. In the loop for each circuit, the function goes to the starting grid page on the website by changing results.html with starting-grid.html in the website link. The website includes an HTML table with qualifying results for each track. This data is stored in a data frame with the name year_df. A data frame containing all qualifying results up to the current year is created by joining year_df of the year scraped with the results from the previous year once all circuits for a particular year have been scrapped. The "QualifyingResults.csv" file is created after the required data cleaning and appropriate renaming of the columns. For the last dataset, the weather data is scraped from the Wikipedia tables. The URL for wikipedia tables was scrapped with races data from Eargast API. Each race has a seprate URL stored with their respective race in the table which is accessed while scrapping. The weather information is categorized into five weather types: warm, cold, dry, wet, and cloudy. The function loops through each URL to read the first column from the Wikipedia tables. The weather information is added to the list. If weather information is not found, the code uses selenium to navigate to the Italian version of the Wikipedia page and scrape the weather information from there. After the information is collected, a dictionary is created to map the weather information to one of the four categories. A data frame is created with columns corresponding to each weather type. Finally, the weather information is concatenated with the race information to create the "WeatherInfo.csv".

3.2. Data Description

Six different sets of data were scraped to get relevant information about important factors that could affect the result of a Formula 1 race. The following section provides a detailed description of the dataset used in the project. The dataset contains information about the races held in the Formula 1 World Championship from 1950 to 2022.

The "Races.csv" file's data includes information on Formula 1 races that took place between 1950 and 2022. Each row in the dataset represents a different race, and each column contains useful details about that race. The information in the dataset includes the season year, race's round number, name of the circuit, latitude, and longitude of the circuit, the country in which the circuit is located, the race's date, and the URL to the specific race's Wikipedia page from the Eargast API.

The "Results.csv" is the largest of the dataset among all of the datasets. The file contains 25,388 rows with each row containing data about the finishing position of a driver in a particular race from the years 1950 to 2022. The information included in the column is season year, circuit name, driver name, date of birth of the driver, nationality of the driver, the constructor team of the driver, grid position of the driver at the start of the race, time taken to finish the race, status of the result of the driver at the end of the race, points scored during the particular round, finishing position at the end of the race. The results dataset can be uneven because of changes in regulations over the years, especially points scored from the race. The dataset also includes missing values for the finishing time of the race. This data is further cleaned to make it more effective.

The “DriverStandings.csv” dataset is a collection of data on driver standings at the end of each round for a particular season. The dataset includes 6 columns: year of the season, race's round number, name of the driver, points scored during the season so far, number of wins so far in the season, and position in the driver standings. This data provides detailed information about their performance form going into each race and their overall record throughout their career.

The constructor standings are saved in a similar format as driver standings. It is a collection of data on constructor standings at the end of each round for a particular season. The data include the season year, race's round number, constructor name, points scored by the constructor at the end of each race, wins at the end of each race for a particular race, and position in the constructor standings for the season. Similar to driver standings, constructor standings are also important to show current form in the season and overall record in Formula 1. For example, Mercedes and Red Bull are coming from a strong previous season and can be expected to challenge in 2022 but Ferrari is recovering from a poor 2020 season and can be expected to rarely challenge for wins. The beginning era of Formula 1 had different rules and regulations which resulted in missing and inconsistent data in the early 1950s.

The qualifying results are stored in a similar way as the data of race results, each row consists of the result of the driver for each round of each race. The qualifying data is only available from 1986 instead of 1950. The dataset contains information about the grid position of the driver at the end of the qualifying, the name of the driver, the final qualifying lap time, the year of the season, and the race’s round number. The qualifying time stored in the column is the final qualifying time at the end of the session instead of the time recorded at the end of each of the 3 qualifying sub-sessions. This is done to keep data consistent as this is a modern format of qualifying and the best and last qualifying time only matters. To explain it better, a driver is not allowed to run a qualifying lap if he/she gets knocked out after Q1, and the time recorded at the end of Q1 is the final one along with the grid position. The data requires some cleaning as the name of the driver also includes aliases which are not consistent with other datasets.

Lastly, the weather dataset includes information about the weather recorded on the day of the race. Weather plays an important role in a race and can affect the results a lot. The information in the dataset includes season year, race’s round number, name of the circuit, overall recorded weather, and four columns of different categories of the weather namely warm, cold, wet, and cloudy. The weather column includes different keywords which were possible to be summed up. For the same reason, the four categories were created each one with a column. The data in these four columns is of binary type (0 or 1) with the column of the weather during the particular race with a value of 1 while all others have a value of 0.

3.3 Data Cleaning and Preprocessing

One of the major steps performed in a data-intensive project is the process of data cleaning and preparation. Data cleaning includes reducing the data to the most relevant and important features. For this project, I decided to work on race data after the engine regulation changes in 2014 to V6 hybrid power units (2014 F1 Sporting Regulations, n.d.). All the data frames were cropped down to data from 2014 - 2022. Sicoie (2022) and van Kesteren & Bergkamp's (2022) studies also use the same approach of using race data from 2014 and later. Formula 1 regulations are what define the sport and distinguish it from other racing series. Every few years Fédération Internationale de l’Automobile (FIA) makes major regulation changes to improve racing and competitiveness while also giving constructors enough room to show their creativity and engineering skills to build the fastest car. The regulation changes of 2014 are one of the major changes that revolutionize the concept of the car (A racing revolution? understanding 2014's technical regulations 2014). The next major regulation change came in 2022 which included major changes like the re-introduction of ground effect cars, budget cap, and freezing engine changes until 2026 (7 key rule changes for the 2022 season: Formula 1® 2022). This gives a better idea of what factors affect the race results in modern-era cars.

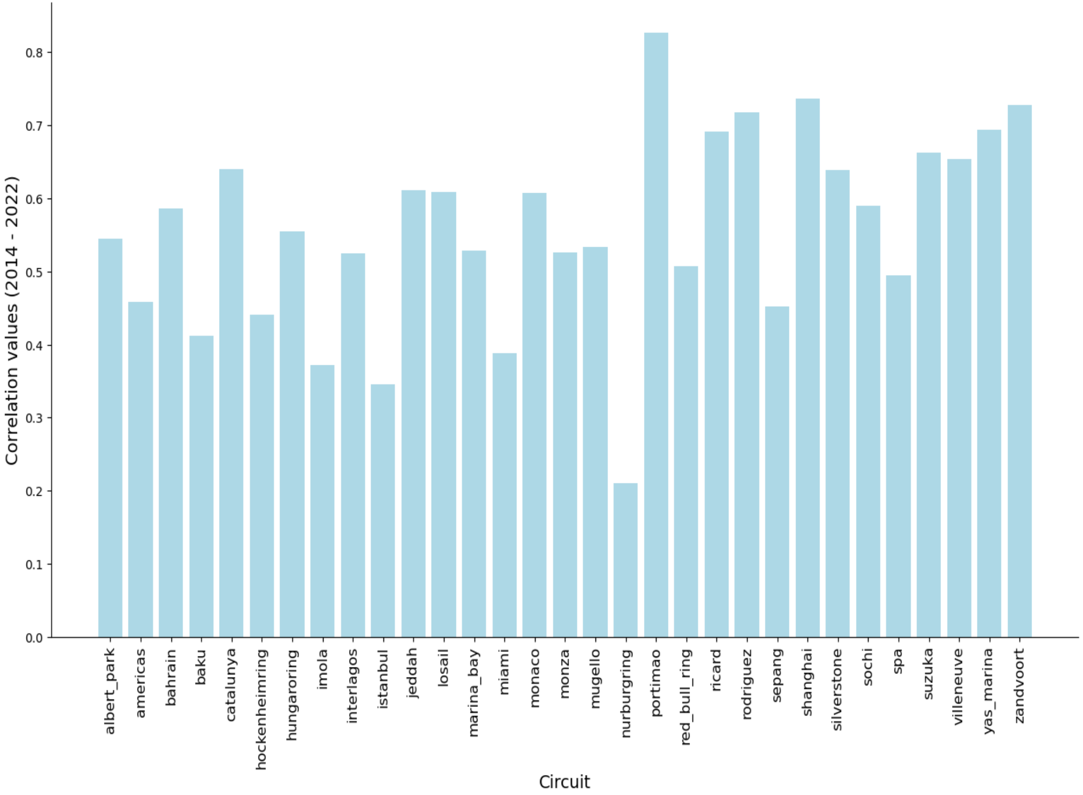
After cropping the data, I move on to cleaning each data set individually. This mostly includes looking for missing values or removing unnecessary columns of data. Mostly all of the datasets are pretty clean with no missing data. Starting with the “races” data, I removed the latitude, longitude, country of the race, and URL columns from the data frame. Latitude, longitude, and country columns were mostly used for the EDA and the data already includes the circuit name which associates the track with its location. And the URL was used for extracting other data during scraping. For “Results” data, the data included 1892 missing race time data and 3 points record. I am currently focusing on predicting the winner of an F1 race and race time can be avoided for now. The 3 missing points data was manually inserted which included decimal scoring records. The results data includes the status of each driver at the end of each race. The status of the drivers was categorized into major categories: “Finished”, “Mechanical Issue”, “Incident”, and “Illness” to simplify the problems of excessive categorical variables.

The scraped Constructor standings, driver standings, and Weather data were clean with only the necessary data present. Similar to the missing race time data, qualifying also included a few missing qualifying lap time records. These records were also removed as I had the starting grid position in the dataset. After cleaning the data, all of the data frames were merged to create a final data frame. The races, results, and weather data were combined on season, round, and name of the circuit. The driver standings and constructor standings were combined on the driver name and constructor name respectively. The final data frame created included data of each driver of each race from 2014 to 2022. As inspired by Sicoie (2022) the driver's age was calculated during each race after subtracting the date of the race and the date of birth of the driver. Sicoie (2022) explains the importance of this feature as age correlates with the experience of the driver and therefore higher rankings. There is a peak in the performance of driver as he/she gains experience with age, this also explains the decline in performance after a certain age. Similar to simplifying the categories of the status of the driver at the end of the race. Over the years, many constructor teams go over various changes. For example, over the years, Force India was rebranded to the Racing Point and again rebranded to the present-running Aston Martin team. It takes time in F1 to become a front-running team from a mid-table team even with new owners and new sponsorships. If the team is rebranded from another team, most of the time engineers and staff remain relatively the same. This correlates with the performance of the team. To keep a record of the new teams, I simplified the constructor's data by mapping all the teams with their predecessors. This was done after creating the new final data frame to keep the data unified. The pre-processing of the data was completed using Onehotencoder of Scikit − learn to tackle the problem of the categorical data (Circuit Name, Nationality of the driver, Constructor Name, status of the driver).

3.4 Data Exploration

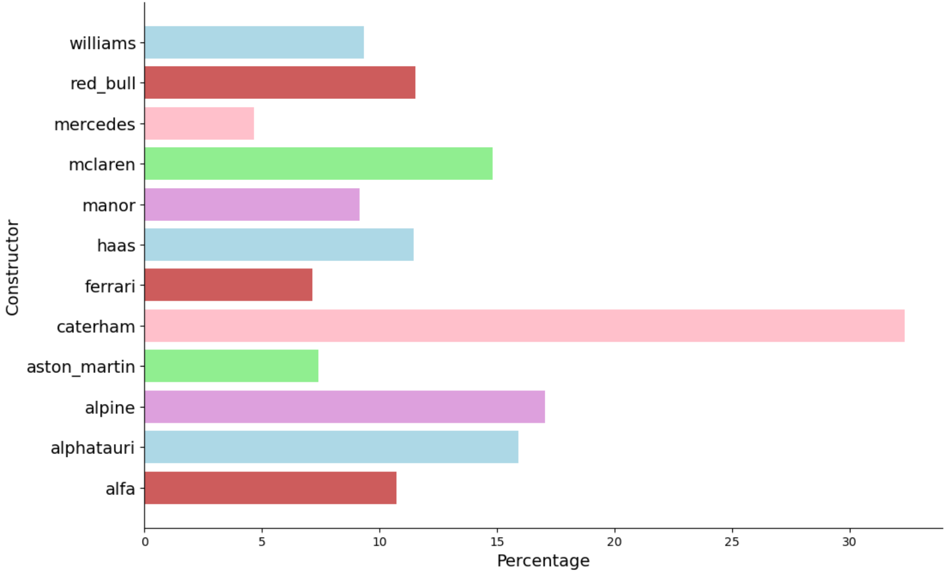
Exploratory Data Analysis (EDA) is an important step in any data analysis project. It involves examining and summarizing the important features and patterns in the data to gain insights and identify potential issues. Results in a Formula 1 race depend on many factors and determining these factors helps to improve the prediction. In this section, we will discuss the main findings from the EDA.

Figure1: Correlation values of starting position and finishing position per circuit

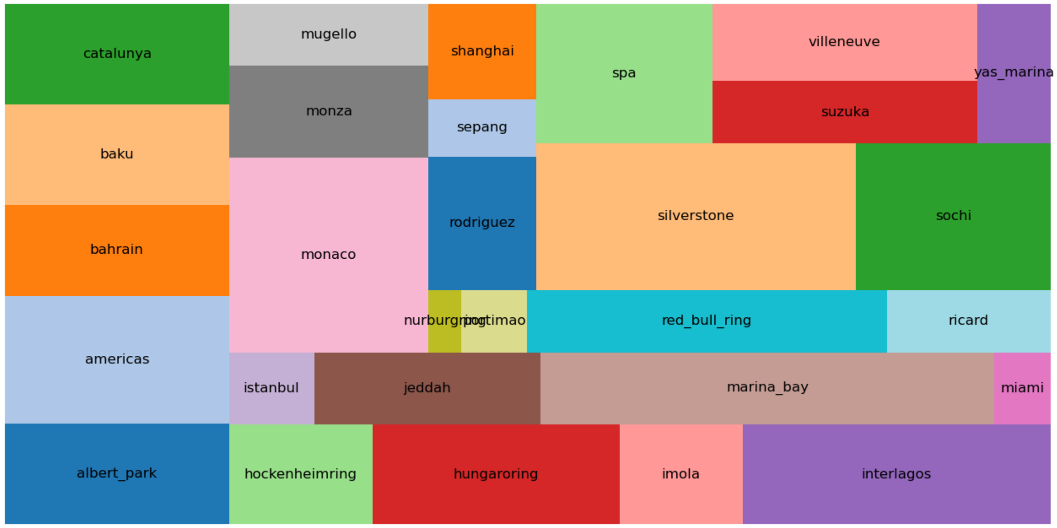


Firstly, the Position Correlation graph (Figure 1) shows the importance of starting the grid of a driver per circuit. The graph illustrates the correlation of the finishing position and the starting position of a driver, it shows the likelihood of finishing in the same position as starting position at each circuit. This trend holds true across most circuits, indicating the importance of a good starting position for achieving a good finishing position.

Figure 2: Percentage of failures by constructor

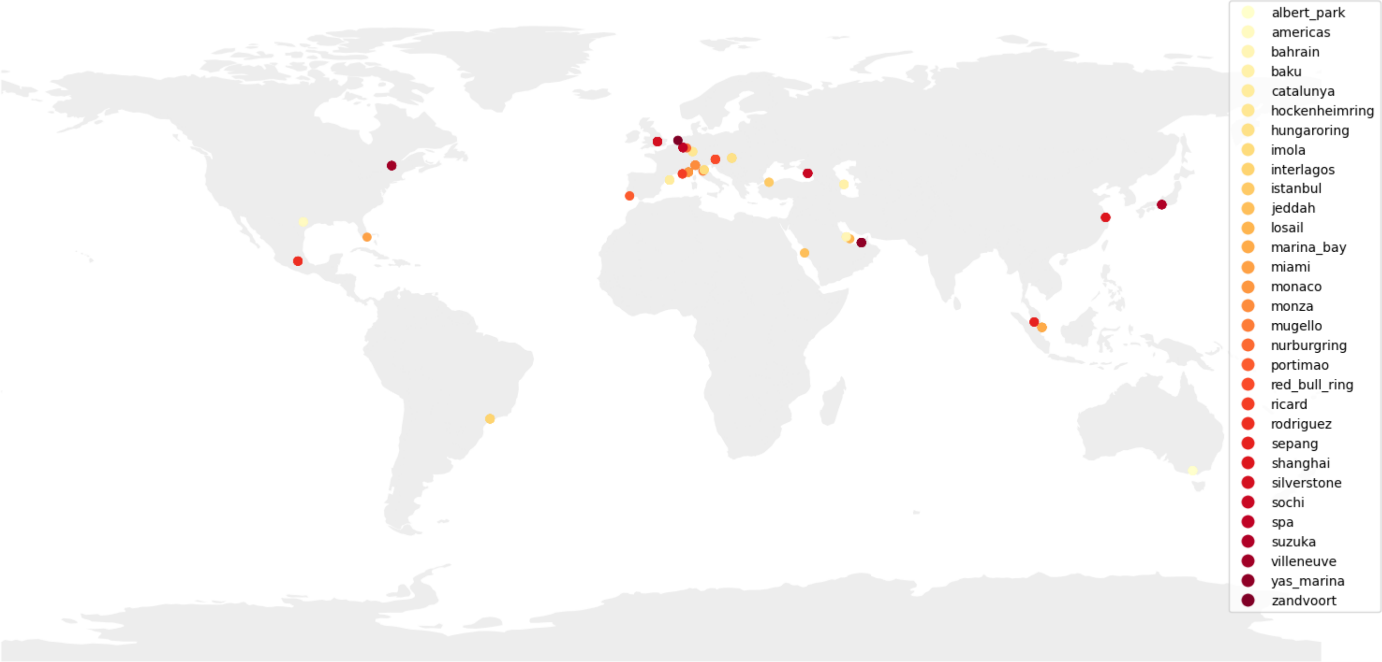


Secondly, the Failure Percentage graph (Figure 2) shows the records of teams in terms of reliability from 2014 - 2022. The graph indicates that some teams have a higher rate of failures than others, which can significantly impact their performance in races. The graph shows the lowest failure percentage for the most dominating team of the turbo hybrid era - Mercedes. This underscores the importance of reliability in Formula 1 racing, as it is a critical factor in achieving success.



Thirdly, the Incidents plot (Figure 3) shows a treemap of total incidents over the years (2014-2022) per circuit. This graph highlights the variability in the number of incidents per circuit over the years. The treemap visualization provides a clear comparison of the different circuits and the relative number of incidents that occurred over time. The graph in Figure 2 and Figure 3 states the importance of the "status of driver" feature in the results data as it suggests that circuits with more incidents or DNFs may have a higher likelihood of producing a random winner. The status of the driver contains categories of how the race ended for the driver: Finishing, Issue, Incident, Illness. The chances of failure of the driver's car or the likelihood of an incident at a circuit can increase the chances of a new winner. Pierre Gasly's win in 2019 Monza GP and Esteban Ocon's win in 2021 Hungaroring GP are a good example of a new winner because of incidents and failure to finish the race by other top drivers.

Figure 4: Circuit Locations



Lastly, the Circuit Location graph (Figure 4) provides an overview of the distribution of circuits on the world map. This graph shows that Formula 1 races are held in different parts of the world, with a concentration in Europe and Asia. The Americas and Australia also have a few circuits. The results of a race depends on the track specific layouts and their weather conditions. The weather is likely to be warm in Bahrain as compared to a race in Zandvoort. This highlights the importance and variance of Formula 1 circuits and their impact on the results.

4. Models

This section focuses on the model used for the prediction and follows up on its performance. The final dataset created after preprocessing is split into training and dataset. Two different sets of training and testing sets were made to better understand the difference. The first training set contained data from 2014-2021 data and was tested on 2022 year. The second set of data was from 2014 to the halfway of the 2022 year and was tested on the remaining half. Similar to the turbo hybrid regulation changes, the next big change of 2022 shuffled the order again and this difference is also seen in the prediction among the two sets. For predicting just the winner of the race, the finishing position of the dataset was converted to a binary data with 1 as the winner and rest of them as 0. This would simplify the data a bit for the model to predict the probabilities. To deal with the non-categorical non-numerical data, I transformed the data frame using the “StandardScaler” library.
In Nigro (2022) multiple models were used to predict the winner using similar data. The best-performing models were the Neural Network classifier and the SVM classifier model with a precision score of a little over 0.6. I decided to use the same model on my set of data to try and improve it.

4.1. Neural Network Classifier

The neural network classifier is a supervised learning algorithm that in this case is used for the classification of the race winners. I used a similar approach as Nigro (2022) to find the best parameters by training data on several models and using the model with the highest precision score. Hyperparameter testing is an important step to find the most effective models. The performance depends of a model depends on the hyperparameters. For the MLP Classifier, a grid search approach was used to try different models. Four parameters were being tuned: “hidden layer sizes”, “activation”, “solver”, and “alpha”. For the first set, the suggested hyperparameters were hidden layers having 75, 25, 50, and 10 neurons respectively, the activation parameter uses the identity function, and the solver to train the MLP is ‘lbfgs’ in this case, to prevent overfitting the alpha is set to 0.01623, and maximum iterations are set to 2000. For the second set of data, the parameters changed a lot. The hidden layer changed to 80,20, 40, and 5 neurons, the activation parameter used is tanh, while the solver and alpha remain the same. After getting the best possible parameters the data was trained again. After the final training, the model was tested on the respective sets to predict the race winners.

4.2. Support Vector Machine

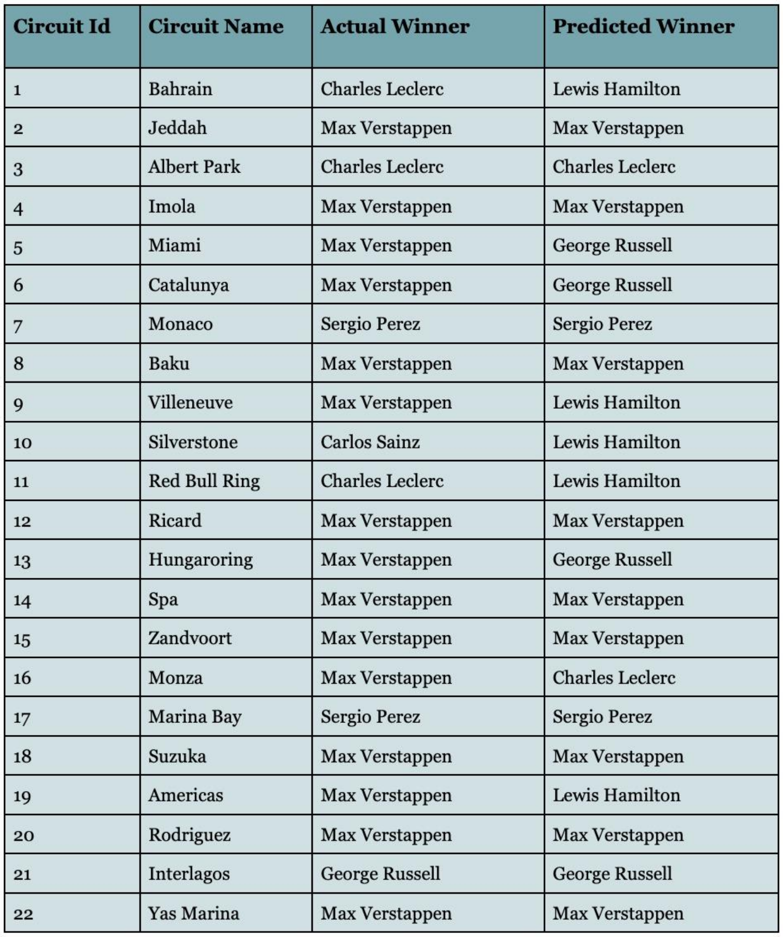
The other significant model that is used by Nigro (2020) was the Support vector machine classifier. A similar approach MLP classifier was used for this model as well. Different hyperparameters were tested using grid search approach. Two different models were created each for both of the training data sets. For the first set, the best fit hypeparameters were 0.1 gama to define the low influence of the training results, regularization parameter at 10 to control trade-off between achieving a low training error and a low testing error that is the ability of the model to generalize to unseen data, and used the sigmoid kernel to handle non-linearly seprable data.

5. Results

5.1. Neural Network Classifier

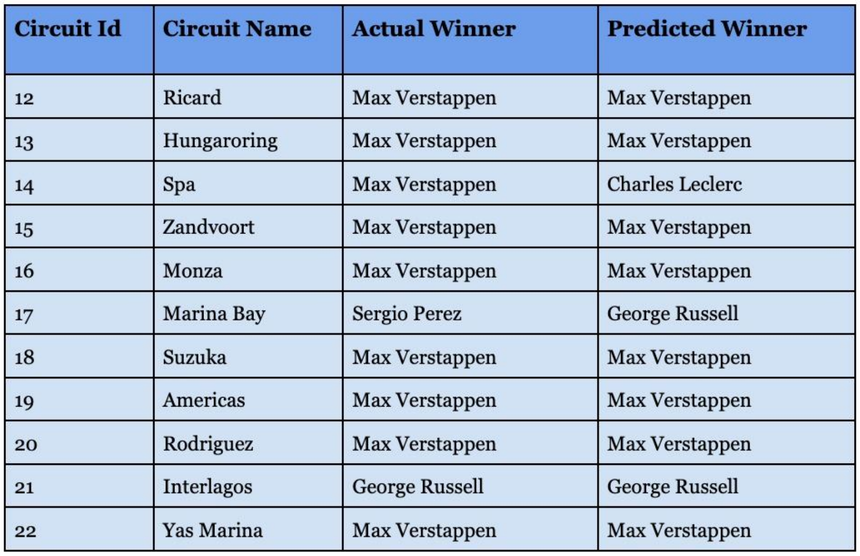
The initial precision score after testing it on the test data gave a score of 1.0. Looking at the prediction results the model learns to predict the first driver of the race round thus giving a high precision score. To further check I created a skewed dataset by shuffling the rows of the data frame before training the data and as expected the model predicted the first row as the winner with a new precision score of 0.15. To tackle the issue the classifier was function changed to how it predicts the winner. The new final model found with the training dataset of 2014-2021 predicted with a precision score of 0.6. The model as expected with this particular training dataset predicted for Mercedes to be a equally strong team as Red Bull and Ferrari and predicted for Hamilton and Russell to win a few races. The model is not trained for regulation changes and the chances crated of a new top team because of them. The second training dataset predicted a better result based on the training of 2014-2021 and the first half of 2022. The results were predicted with a precision score of 0.8. Although the score improved for the second model. The testing set is comparatively small and can lead to different results.

Neural Network Classifier (Training: 2014- 2021)



Circuit Id	Circuit Name	Actual Winner	Predicted Winner
1	Bahrain	Charles Leclerc	Lewis Hamilton
2	Jeddah	Max Verstappen	Max Verstappen
3	Albert Park	Charles Leclerc	Charles Leclerc
4	Imola	Max Verstappen	Max Verstappen
5	Miami	Max Verstappen	George Russell
6	Catalunya	Max Verstappen	George Russell
7	Monaco	Sergio Perez	Sergio Perez
8	Baku	Max Verstappen	Max Verstappen
9	Villeneuve	Max Verstappen	Lewis Hamilton
10	Silverstone	Carlos Sainz	Lewis Hamilton
11	Red Bull Ring	Charles Leclerc	Lewis Hamilton
12	Ricard	Max Verstappen	Max Verstappen
13	Hungaroring	Max Verstappen	George Russell
14	Spa	Max Verstappen	Max Verstappen
15	Zandvoort	Max Verstappen	Max Verstappen
16	Monza	Max Verstappen	Charles Leclerc
17	Marina Bay	Sergio Perez	Sergio Perez
18	Suzuka	Max Verstappen	Max Verstappen
19	Americas	Max Verstappen	Lewis Hamilton
20	Rodriguez	Max Verstappen	Max Verstappen
21	Interlagos	George Russell	George Russell
22	Yas Marina	Max Verstappen	Max Verstappen

Neural Network Classifier (Training: 2014- 2022(first 11 races))



Circuit Id	Circuit Name	Actual Winner	Predicted Winner
12	Ricard	Max Verstappen	Max Verstappen
13	Hungaroring	Max Verstappen	Max Verstappen
14	Spa	Max Verstappen	Charles Leclerc
15	Zandvoort	Max Verstappen	Max Verstappen
16	Monza	Max Verstappen	Max Verstappen
17	Marina Bay	Sergio Perez	George Russell
18	Suzuka	Max Verstappen	Max Verstappen
19	Americas	Max Verstappen	Max Verstappen
20	Rodriguez	Max Verstappen	Max Verstappen
21	Interlagos	George Russell	George Russell
22	Yas Marina	Max Verstappen	Max Verstappen

5.2. Support Vector Machine Classifier

As compared to the Neural Network Classifier, the Support Vector Machine produced similar results with the training set of 2014 - 2021. The model still predicted for a strong Mercedes team with both of the drivers winning. The precision score of this first model was 0.5. The second model with the testing set of second half of 2022 season again produced a similar result as the Neural network classifier with a score of 0.81. The model was able to predict a strong second half of 2022 by max verstappen. This model still has the same concerns of a small testing set. To check if both the Neural Network classifier and Support Vector Machine Classifier is not just memorizing the data or finding a loop hole in the data. I tried the model with a skewed data to check on the results. The model still predicted the same results with a slight change in the probabilities of winning. Testing it on the new 2023 season would be something to look forward to on how this models perform.

6. Conclusion

In conclusion, this thesis has presented a comprehensive study on predicting the winner of a Formula One (F1) race using machine learning (ML) techniques. The paper attempts to show the importance and possibilities of data science in a sport like F1 where multiple unpredictable factors are responsible for a race result. The study began with data collection from various reliable sources which included driver and constructor statistics, weather report and race results. In this paper, we used Neural Network Classifier and Support vector Machine to try and predict the race winner of a formula 1 race over a season. There is still a lot to work further on this topic to try and improve the prediction rate. The models can be trained from 2006 - 2021 instead of the one used for this paper. This would help model train on two different regulation eras (2006-2013 and 2014-2021). There is also data that was not considered which plays a big role in a race's rsult. Race strategies, track conditions, and one lap qualifying pace can also contribute to the results. For future research, the features used in the model can be modified to define weights based on their importance. The ML models used can also be explored more with different hyperparameters and ML techniques such as increasing number neural network layers or using deep learning to improve the accuracy of F1 race winner prediction.

References

van Kesteren, E.-J., & Bergkamp, T. (2022, March 17). *Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage*. Retrieved April 17, 2023, from https://arxiv.org/pdf/2203.08489v1

Sicoie, H. (2022, January 14). *Machine Learning Framework for formula 1 - tilburg university*. Retrieved April 17, 2023, from http://arno.uvt.nl/show.cgi?fid=157635

Nigro, V. (2020, June 11). *Formula 1 race predictor*. Medium. Retrieved April 17, 2023, from https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da

Fédération Internationale de l'Automobile. (2014, February 28). 2014 FORMULA ONE SPORTING REGULATIONS.

Cooney, C. (2020, July 18). Formula One: Extracting and analysing historical results. Medium. Retrieved April 17, 2023, from https://towardsdatascience.com/formula-one-extracting-and-analysing-historical-results-19c950cda1d1

A racing revolution? understanding 2014's technical regulations. Formula 1® - The Official F1® Website. (2014, January 24). Retrieved April 17, 2023, from https://www.formula1.com/en/latest/features/2014/1/A-racing-revolution-Understanding-2014s-technical-regulations.html

F1. (2022, January 31). *7 key rule changes for the 2022 season: Formula 1®*. Formula 1. Retrieved April 17, 2023, from https://www.formula1.com/en/latest/article.7-key-rule-changes-for-the-2022-season.2E7JH9MywymU8xxw6r5yDS.html

F1 Results and Statistics. Formula 1® - The Official F1® Website. (n.d.). Retrieved April 18, 2023, from https://www.formula1.com/en/results.html

Eargast Developer API. Ergast developer API. (n.d.). Retrieved April 18, 2023, from https://ergast.com/mrd/