

Predicting Formula 1 Race Results



Presented By:
Yash Yegare

TODAY'S

Agenda

Introduction to Formula 1

Project Overview

Data Collection

Exploratory Data Analysis

Data Pre-Processing

Machine Learning Models

Application of Machine Learning Models

Conclusion



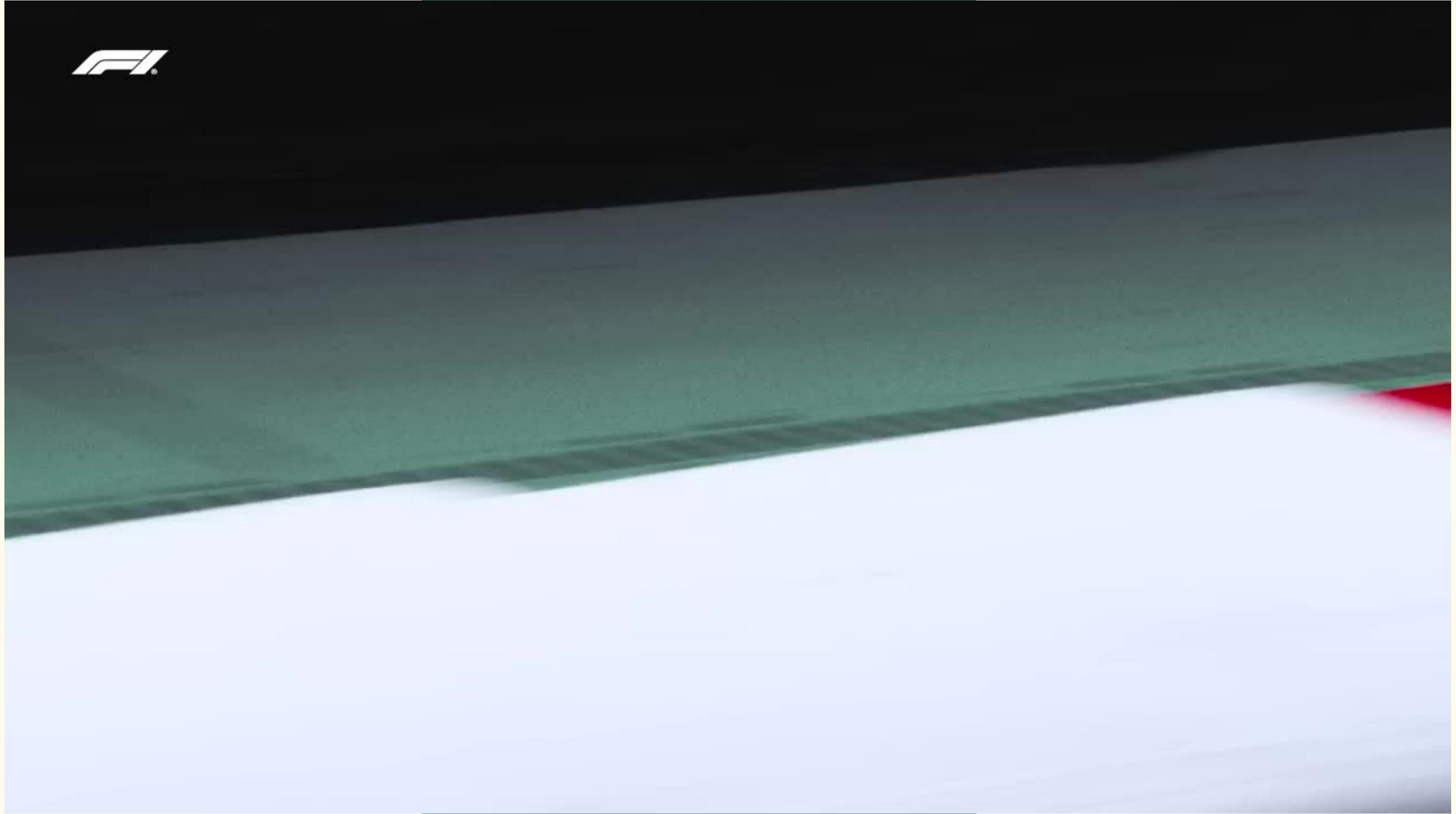


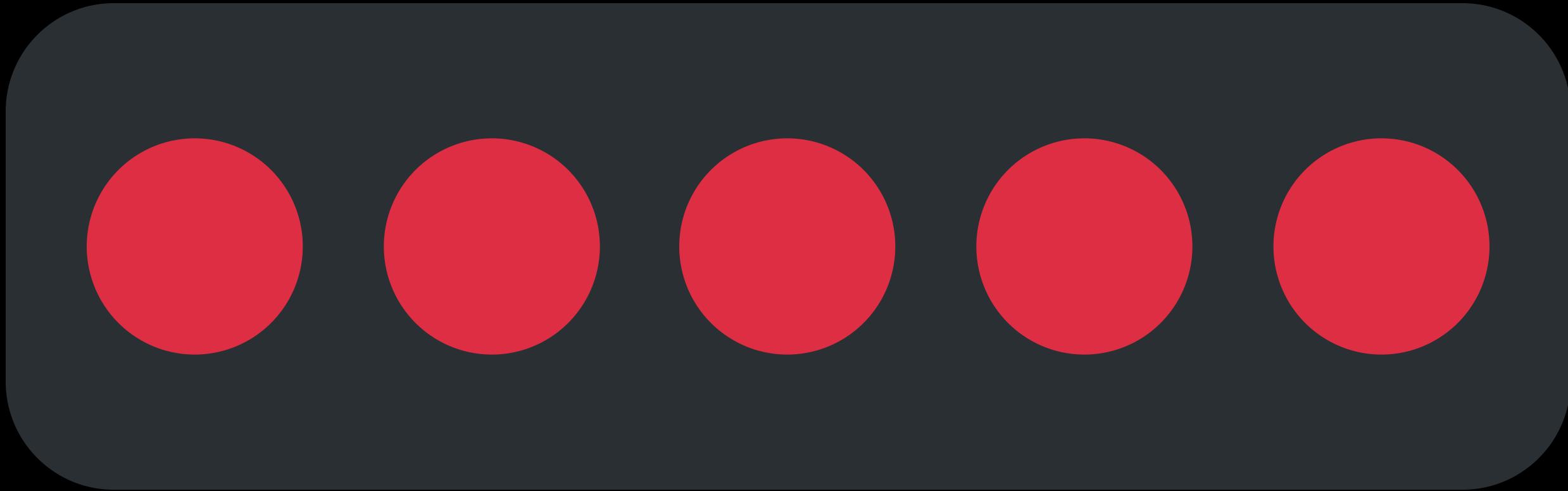
Introduction to Formula 1

- Formula 1 is widely considered the pinnacle of motorsport, boasting the fastest cars and most skilled drivers on the planet.
- The sport has a storied history dating back to the 1950s, with numerous iconic and majestic locations hosting races worldwide.
- Each race weekend is a three-day affair, featuring practice sessions where drivers can fine-tune their cars, culminating in a main qualifying session to determine the starting grid for the race on Sunday.
- Formula 1 has 20 drivers and 10 constructors competing for the World Driver and Constructor Championships, with points accumulated throughout the season to determine the champions.

As people like to say - "It's just a sport where 20 drivers go around in circles for two hours."

Trust me; I am about to blow your minds!!!





ITS FIVE RED LIGHTS!

ITS LIGHTS OUT!

AND AWAY WE GO!

Project Overview

The primary objective was to propose a machine-learning approach to predict the following Formula 1 Grand Prix race winner. Aimed to provide an accurate prediction considering various present and past factors to help fans, team managers, and betters make informed decisions. Aimed to do robust data analysis and find the factors contributing to winners while also predicting the band of winners.



Data Collection

- Gathered historical data on Formula One from 1950 to 2023 from multiple sources, including the Ergast Data repository, the Formula 1 Website, Wikipedia and Open Weather.
- Obtained six individual data frames that contained information about all races, results, driver standings, constructor standings, qualifying standings, and weather conditions.
- Combined these data frames into a final dataset, keeping only the relevant features that influenced the race's outcome and removing redundant columns.

Weather

Qualifying
Results

Race Stats

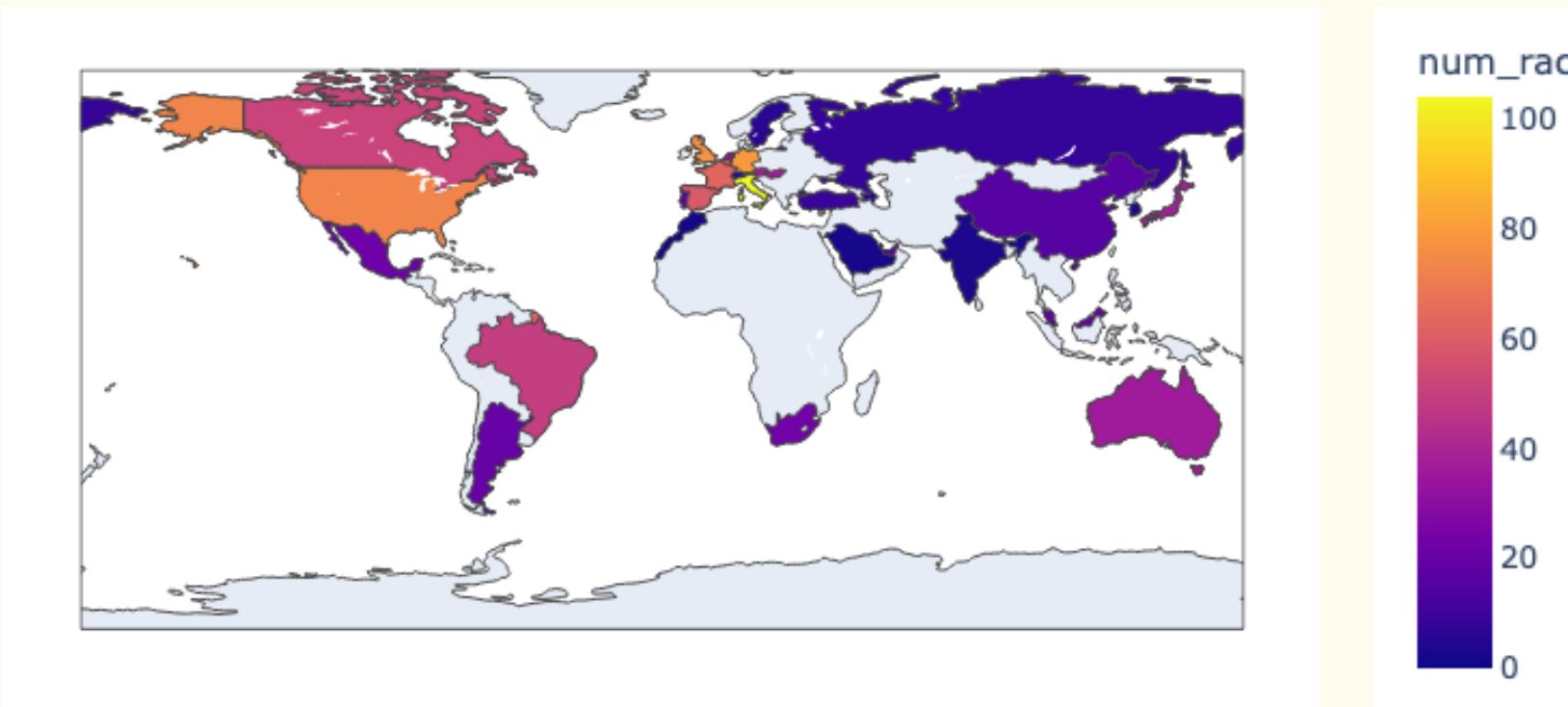
Race
MetaData

Driver &
Constructor
Standings

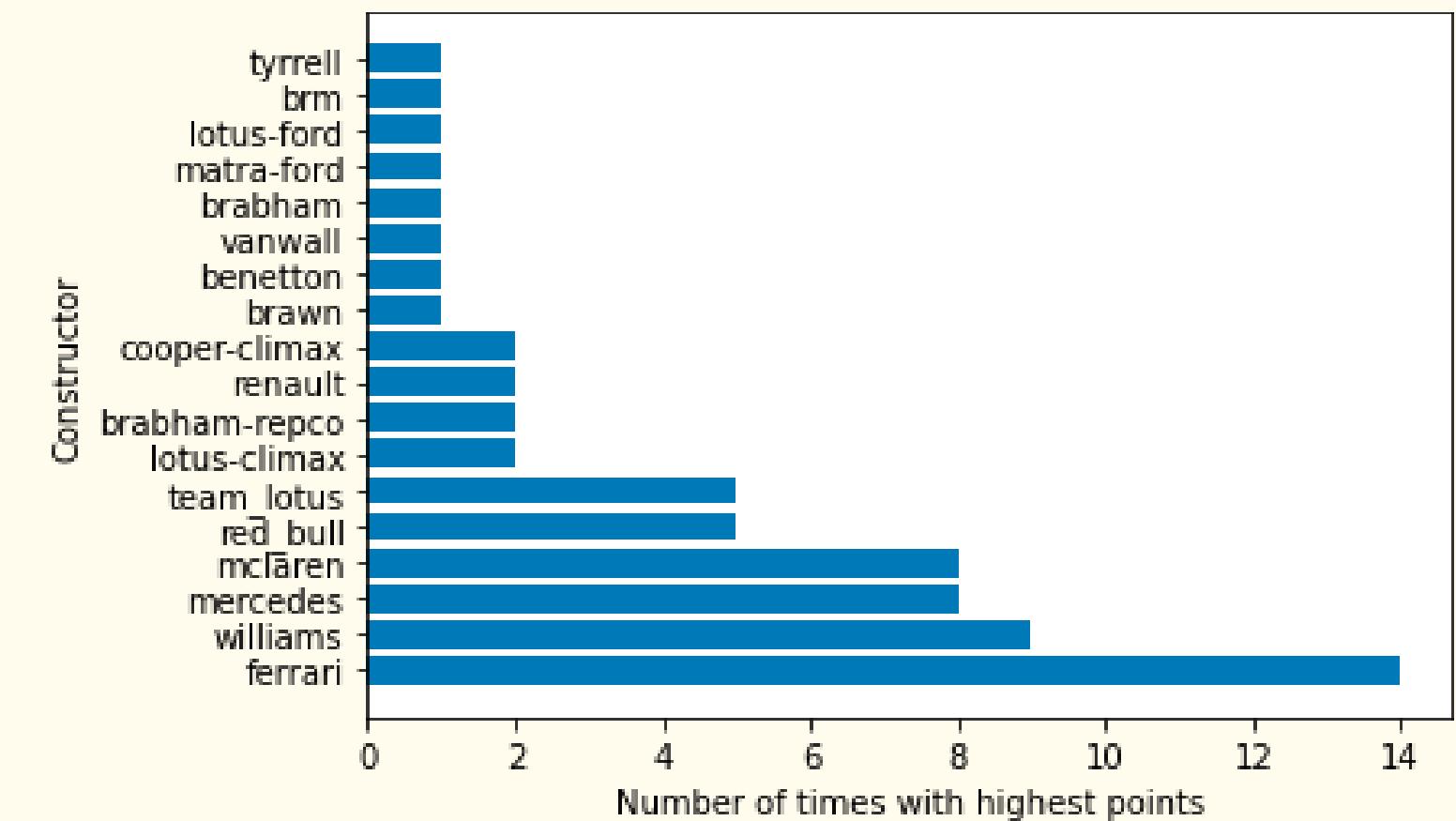


Exploratory Data Analysis

Number of races that have taken place in each country

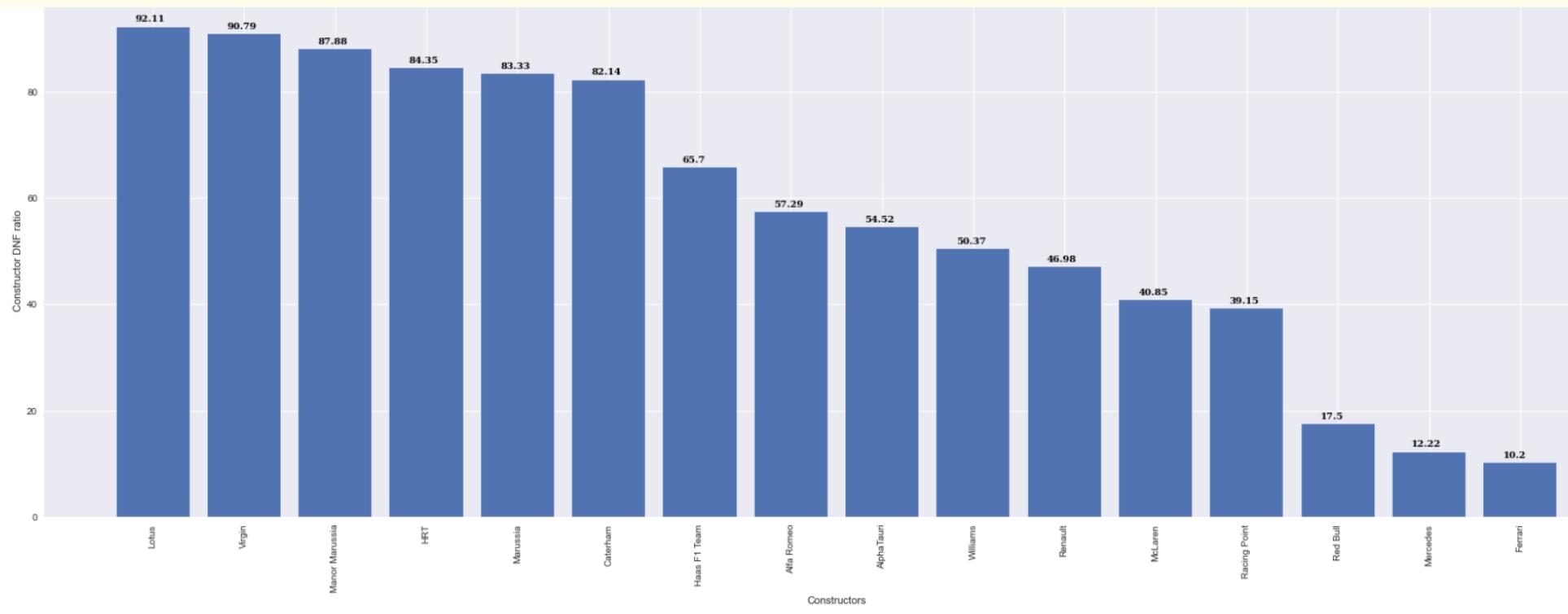


Number of times each constructor had the highest points at the end of each season

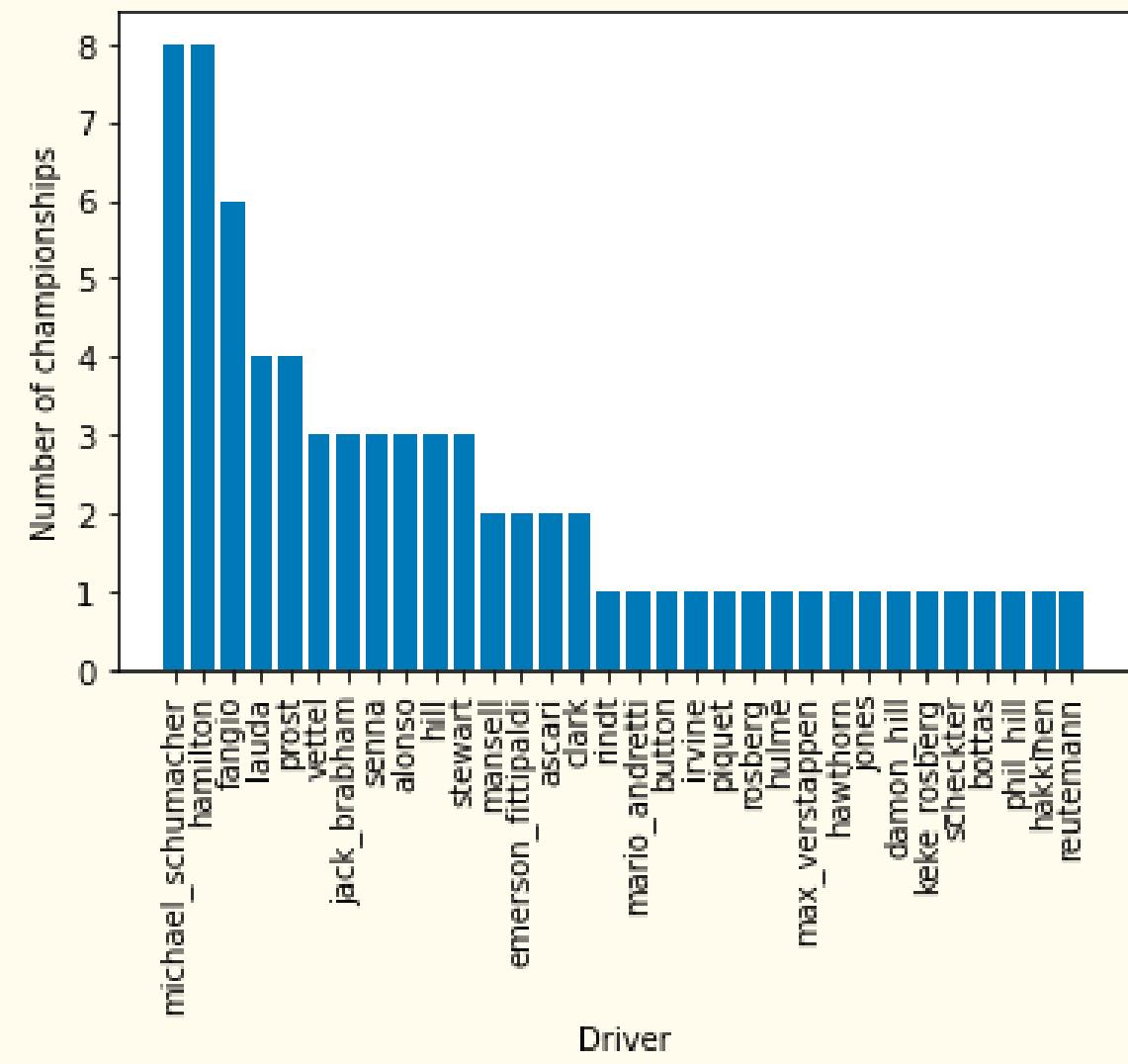


Exploratory Data Analysis

DNFs ratio due to constructor error

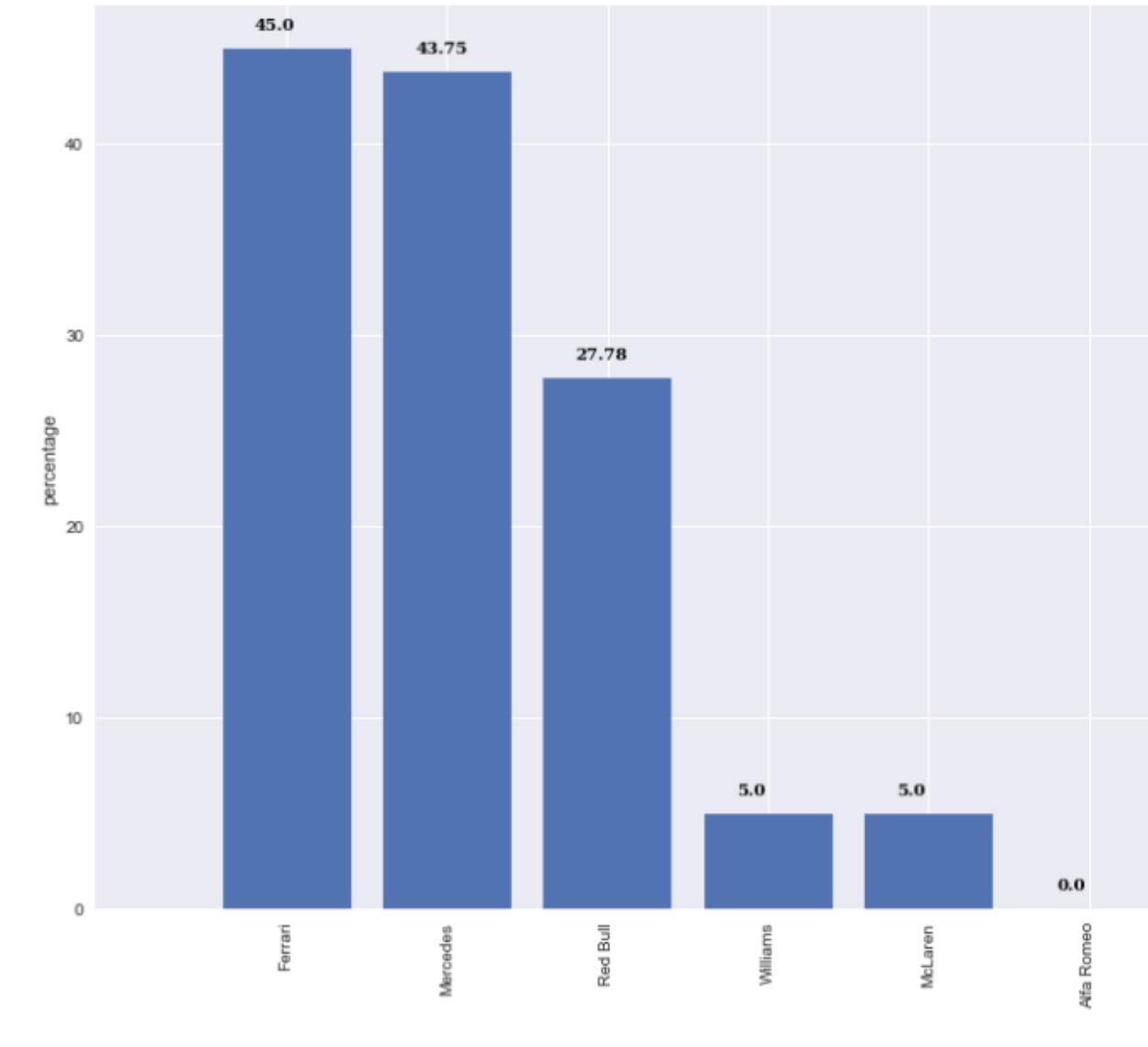


Number of championships won by each driver

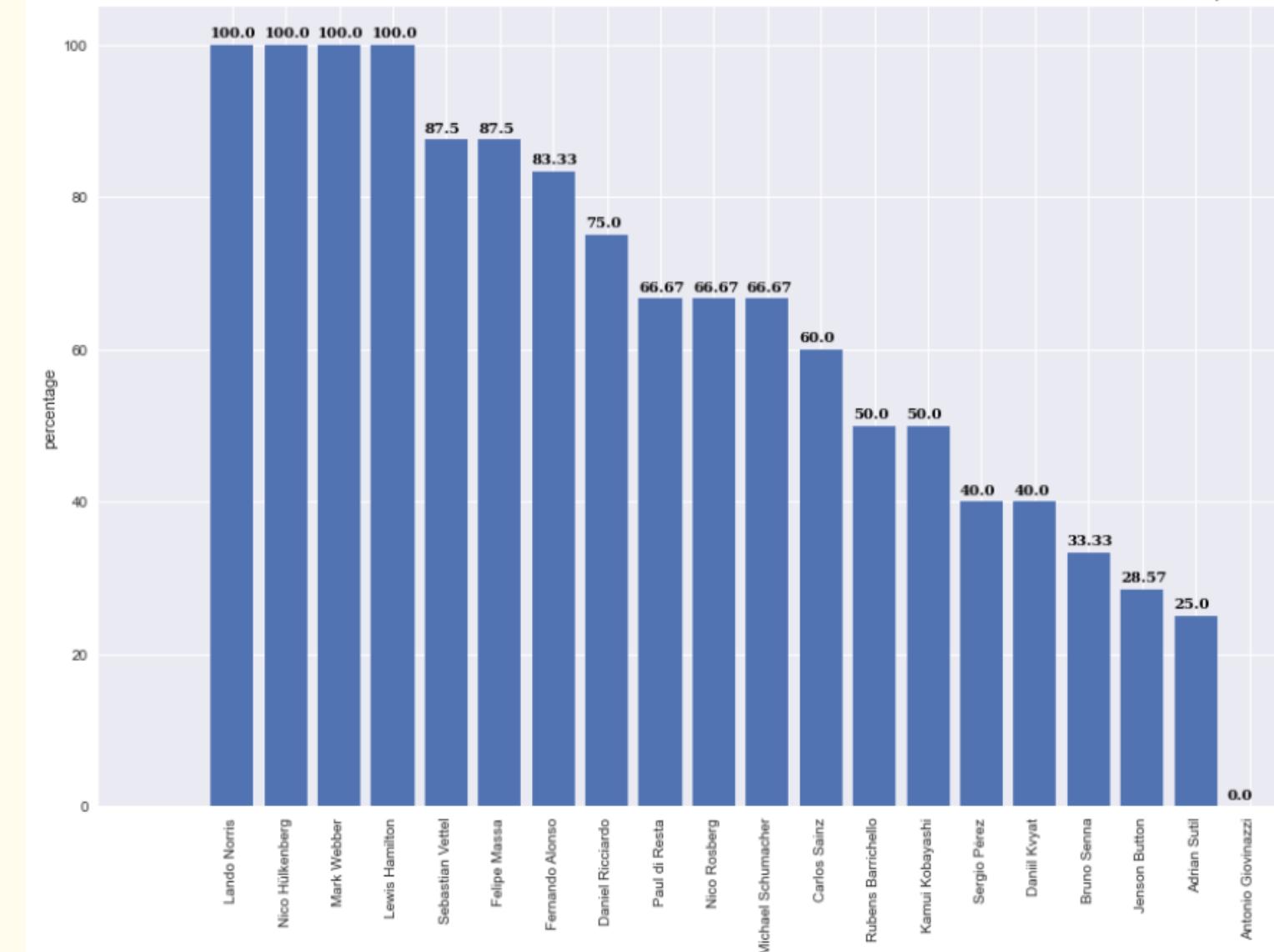


Exploratory Data Analysis

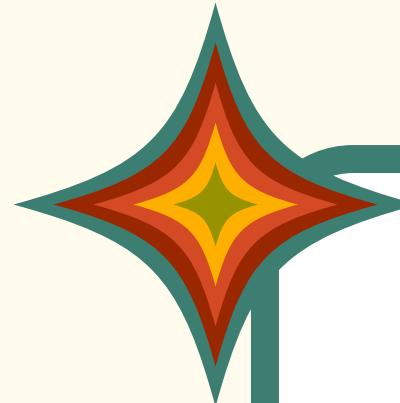
Constructor podium finish percentage at home race



drivers point finish percentage at home race



Data Preprocessing



Collated Data Frames

I collated these data frames into a conclusive dataset, retaining solely the significant features that impacted the race's outcome while eliminating any superfluous columns.

Encoding the Data

As necessary, I also ensured the complete one-hot encoding of the data, thereby transforming all non-numerical data into fully numeric tabular data. Additionally, conducted label encoding of the drivers, since one-hot encoding was not a feasible option, and I deemed it imperative to be cognizant of the variables which I encoded them to.

Scaled Data:

Finally, to ensure that all the numerical features on the dataset were on a comparable scale, I implemented Standard Scaler.



Machine Learning Models

Now after perfecting the dataset to the maximum extent, at least in my opinion, we're moving on to the heart of our project's success. In the Machine Learning models part, I trained, tested and predicted the likelihood of the driver's finishing position from their qualifying result. From which I bring out and conclude ample inferences as well as many applications.



PART 1

First Attempt

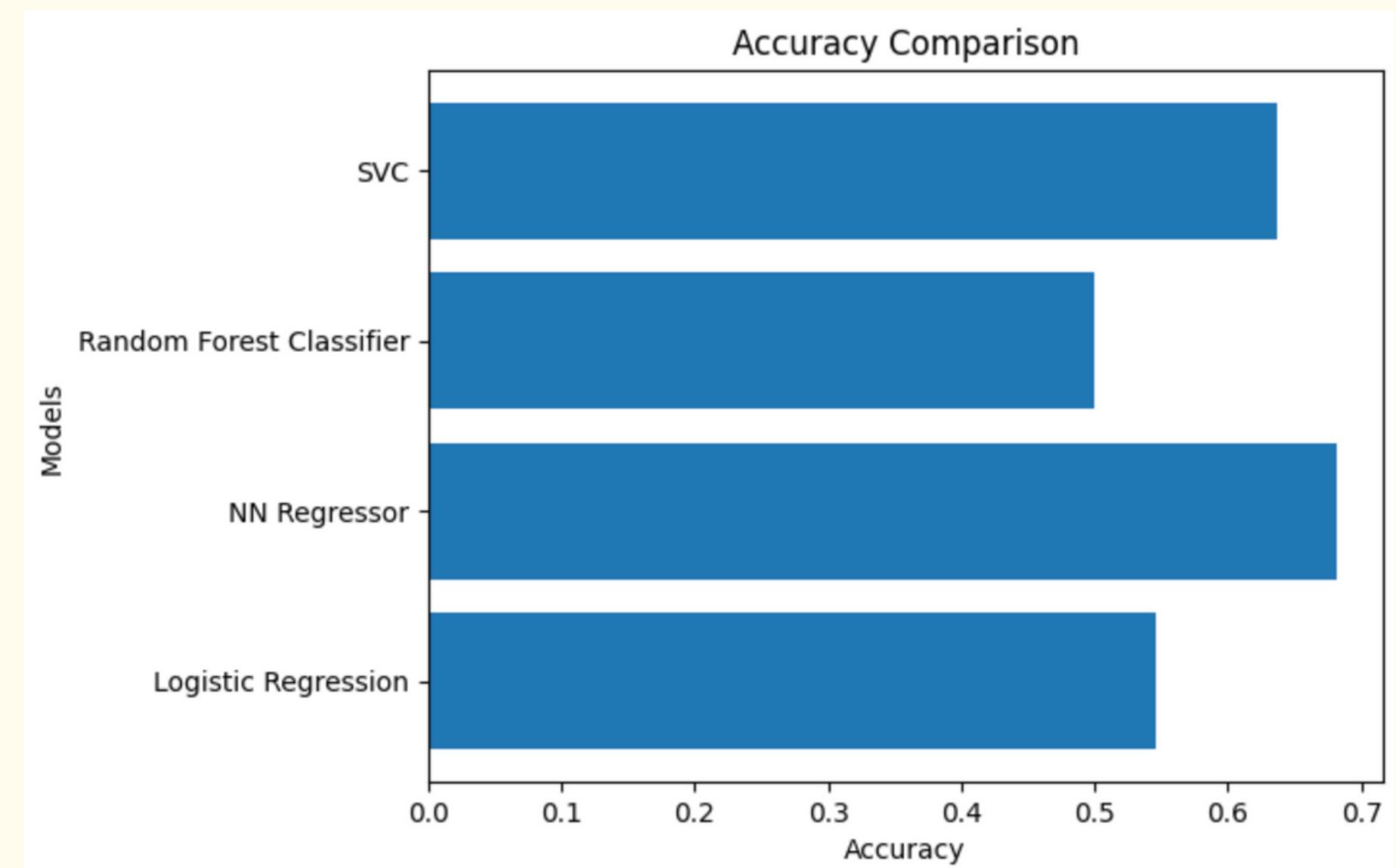


Methods and Discussions

- Important features discovered through data analysis:
 - Weather
 - Constructor & Driver standing in the season
 - Qualifying race lap times, and others.
- Explored two approaches to solve the problem:
 - **Classification:** Built a model that categorizes the drivers into podium range, points range, or no points range.
 - **Regression:** Sorted the values according to the predicted values and placed them into the appropriate buckets.
- Tuning the hyperparameters of the models was crucial for achieving high accuracy. After several attempts, I used a for loop to iterate through all possible combinations and selected the model with the best accuracy.
- My selection included a range of models such as logistic regressor, linear regressor, neural network, SVM classifier, SVM regressor, random forest classifier, and regressor.
- The best models yielded an accuracy of around 60%, which was lower than the initial expectations. This led me to investigate the approach further and identify areas where I could improve the model's performance.



Preliminary Analysis



What Did Not Go Well

During the preliminary run of models!

Identified unnecessary factors:

Through the analysis, I discovered that I was considering several factors that were not relevant to the predictive models, such as driver age and constructor points. Essentially I realised the assumptions were wrong.

Importance of DNF ratios:

I realized that I had overlooked important factors such as driver DNF and constructor DNF ratios that could have a significant impact on race outcomes.

Issues with hyperparameter tuning:

In the initial approach, I relied on a for loop to tune the hyperparameters of the models. However, I found that this approach did not provide logical conclusions, and I was unable to fine-tune the models as effectively as I had hoped.



PART 2

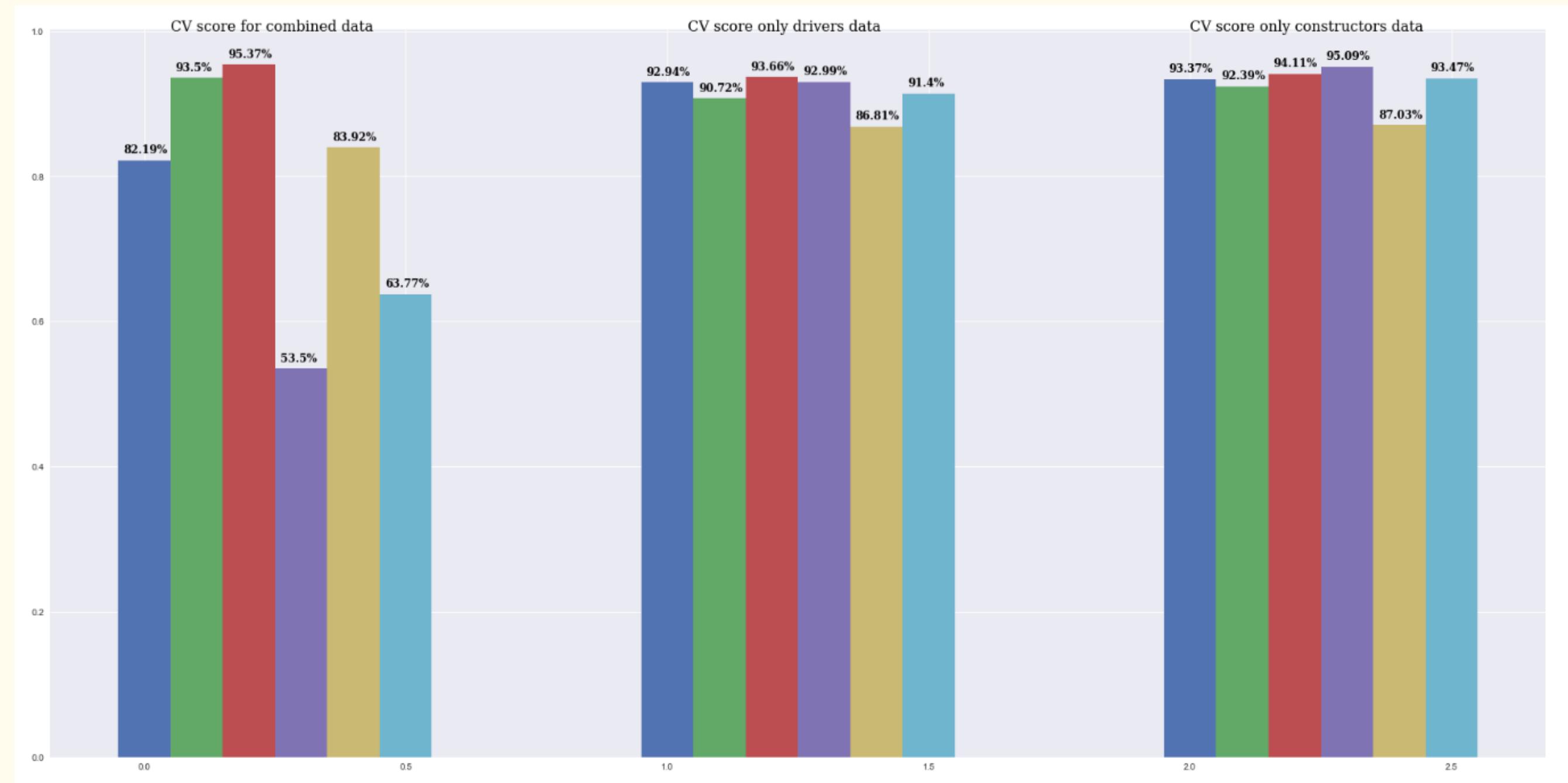
Feature Engineering Modifications



Methods and Discussions

- Data analysis identified DNF ratios and home team advantage as important factors not considered in the previous approach. I aimed to evaluate the impact of these factors on the predictions.
- New features were created by calculating driver and constructor DNF ratios and home team advantages through computations on the existing columns.
- This revealed that driver DNF ratios could be used to determine driver confidence, constructor reliability, and the impact of home team advantage on drivers unless they were underperforming.
- Analysis of previous driver and constructor statistics showed their impact on current race position. I developed separate models for drivers, constructors, and a combined approach.
- Although the combined approach had better accuracy, the individual models were also effective. After experimenting with different models, I found that random forest and SVM classifiers performed well, achieving approximately 90% accuracy in predicting driver race range.
- With the successful prediction of driver ranges, I aimed for a bigger challenge of predicting the actual positions of the drivers.

Test Accuracies



PART 3

Predicting Exact Race Positions



Methods and Discussions

- Taking on a harder challenge, I focused on predicting the exact race positions of drivers.
- Discovered that race positions were highly dependent on qualifying lap times, starting positions, and other factors considered in the previous analysis.
- Updated the dataset accordingly and used the same models with the same hyperparameters since the input format remained the same with adding a few columns.
- The model was able to predict around 60% of the drivers correctly. However, some drivers were still not being predicted due to unexpected accidents and other factors, like safety cars whose data was unavailable even from the online APIs and F1 official site.
- Despite these limitations, I was satisfied with the performance of the model.



What Did Not Go Well

During the next iteration!

Impact of missing data on model predictions

The model was impacted by the unavailability of data on safety cars and accidents, which can significantly affect mid-race changes in position, leading to erroneous predictions in some cases.

Dependency on qualifying race data for accurate predictions

The model's dependence on the qualifying race means it cannot predict the outcomes more than 24 hours before the race. This is because qualifying races occur at least 24 hours before the main race, and I could not access this data before the weekend.



PART 4

Predicting Qualifying Results

WILLIAMS
RACING

Methods and Discussions

- I went on to predict the qualifying results as well. This was the same as race results, and the formats of data were also the same except for the qualifying data not being there, so I realized a random forest would still be better
- Reduces overfitting: Random Forest helps to reduce overfitting by using multiple decision trees instead of a single decision tree. This makes the model more robust and better able to generalize to new data.
- Can handle missing values: One of the advantages of using Random Forest for F1 prediction is its ability to handle missing values. This is crucial since there may be missing data for certain races or teams, and a model that can handle missing values without significant loss in accuracy is highly desirable.
- But the results were shocking as the accuracy was very low.

Inferences

Predicting Qualifying Results

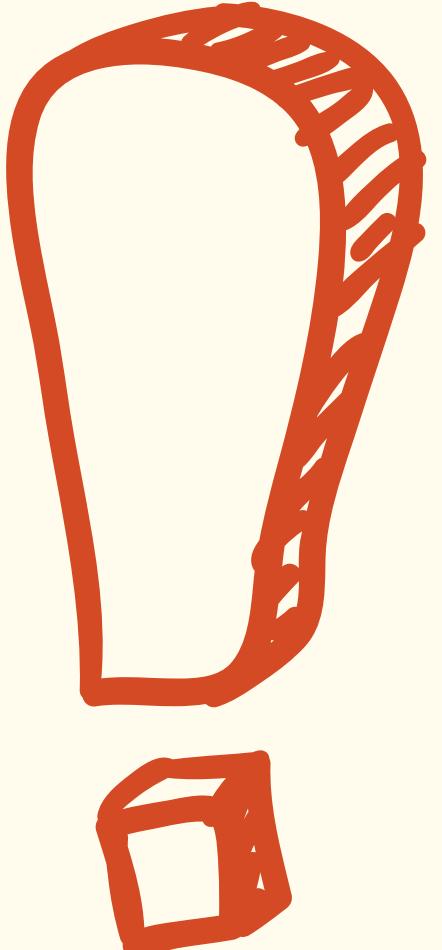
- Qualifying is crucial to F1, setting the tone for the entire weekend. Teams evaluate and adjust their car metrics based on practice races, which can greatly impact race position. I realized that predicting qualifying results requires more research and data.
- Formula 1 is an unpredictable sport with constantly changing variables:
 - Qualifying results can provide accurate data on car and driver performance.
 - More data, expertise, and research are needed to predict qualifying results effectively.
- A key realization I had was that Formula One is an unpredictable sport with many constantly changing factors, such as **accidents** and **pit strategies**, making it difficult to predict race outcomes accurately.
- However, I found that using the qualifying round as a yardstick is an effective way to gauge the level of both the car and the driver. Qualifying provides a clear understanding of the car's and driver's performance levels, which allows me to predict their positions in the race more accurately.





WHAT SEPARATES US

Novelty

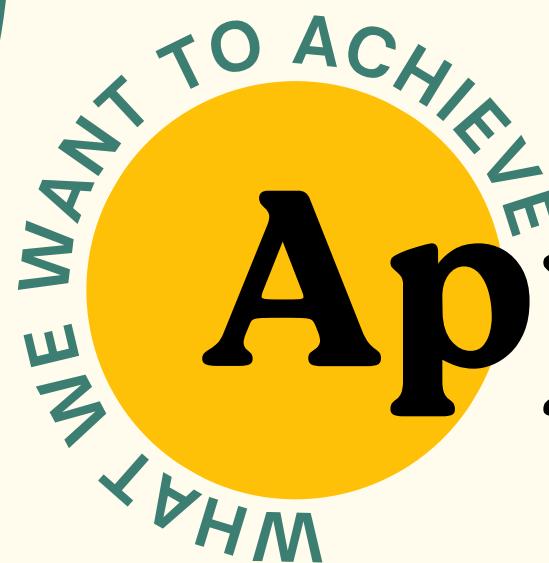


The project applies data analysis and machine learning techniques to predict driver performance in Formula One races, providing a unique and data-driven approach to predicting outcomes.

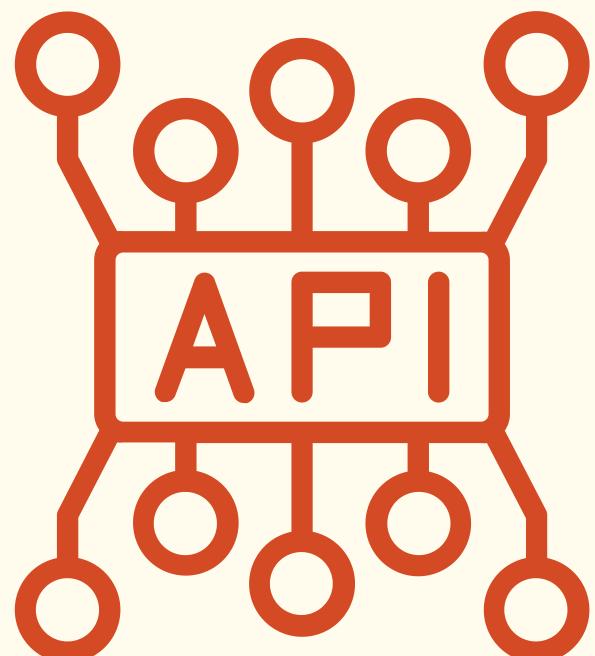
I got to introduce novel features, such as driver and constructor confidence, home team effects, DNF ratios, and DNF index, allowing to capture the nuances of motorsport performance better.

The project has important applications in sports betting, team management, and driver development and the contribution represents a significant advancement in motorsport analytics and showcases the potential of data-driven approaches in predicting and understanding motorsport performance.





Applications



The project would help bettors to make more informed decisions and improve their chances of winning by using data-driven models to predict if a driver would end up on the podium, in points, outside the points or even DNFs.

F1 teams could benefit from our predictions as they would be enabled to make strategic decisions that is in regard to the race strategy, by providing insights into the factors that would contribute to both the driver and constructor success.

This has massive potential even in the field of fantasy F1 by selecting the top-performing drivers and constructors within the constrained limited budget, maximizing the chances of scoring the maximum number of points in the league.





AT THE END

Conclusion

The project successfully applied data analysis and machine learning techniques to predict driver performance in Formula One races.

I used exploratory data analysis, feature engineering, feature selection, and predictive modeling to identify key factors that contribute to driver and constructor success.

The models accurately predicted podium and points positions, taking into account variables such as DNF index, home team effect, circuit analysis, race history, and driver nationality.

Identified key winning factors such as driver experience, circuit characteristics, and team performance. The project has demonstrated the potential of data analysis and machine learning in providing actionable insights and making predictions.



References

Page

<https://www.f1-predictor.com/category/data-science/>

https://www.researchgate.net/publication/359277496_Bayesian_Analysis_of_Formula_One_Race_Results_Disentangling_Driver_Skill_and_Constructor_Advantage

https://eprints.whiterose.ac.uk/96995/14/WRRO_96995.pdf

<https://www.sciencedirect.com/science/article/pii/S2210832717301485>

Thank you! for listening!

Hopefully, you will start to share the same love for
Formula 1 as we (Myself & Formula 1 fans) do.

