Predicting Formula 1 Race Results

Course:
CSE343 - Machine Learning

Presented By:
Ritvik Pendyala (2020378)
Tejdeep Chippa (2020253)
Jaideep Guntupalli (2020378)

TODAYS

Agenda

Introduction to Formula 1

Project Overview

Data Collection

Exploratory Data Analysis

Data Pre-Processing

Machine Learning Models

Application of Machine Learning Models

Conclusion

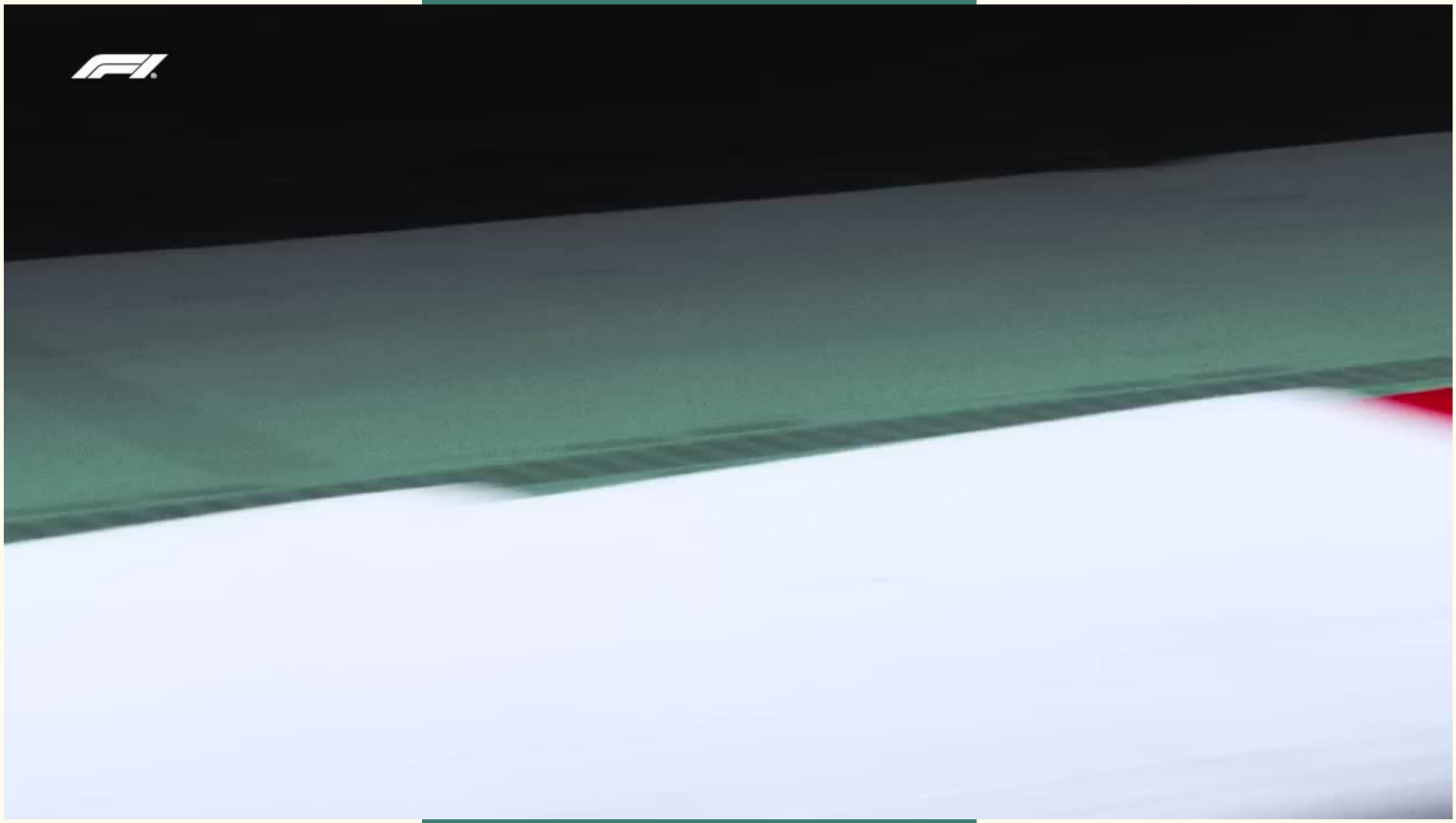
Introduction to Formula 1

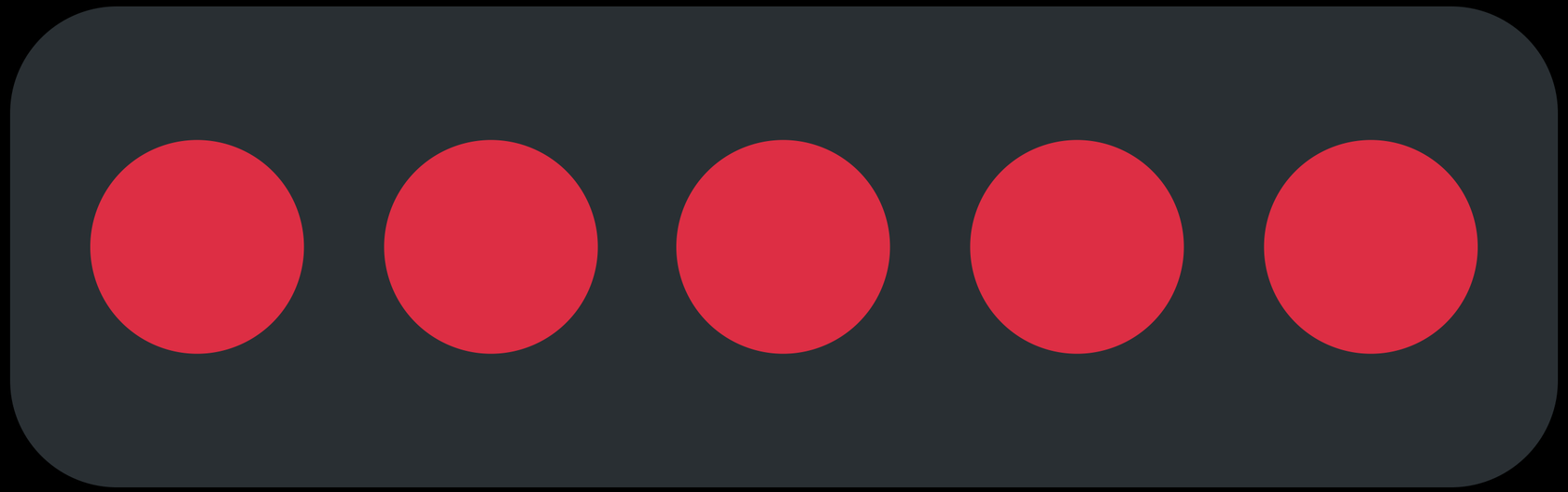
INTRO GUIDE

- 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 us; we're about to blow your mind!!!





ITS FIVE RED LIGHTS!

ITS LIGHTS OUT!

AND AWAY WE GO!

SHORT BRIEF

Project Overview

The primary objective was to propose a machine-learning approach to predict the following Formula 1 Grand Prix race winner. We aim to provide an accurate prediction considering various present and past factors to help fans, team managers, and betters make informed decisions. We aim 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

Race MetaData

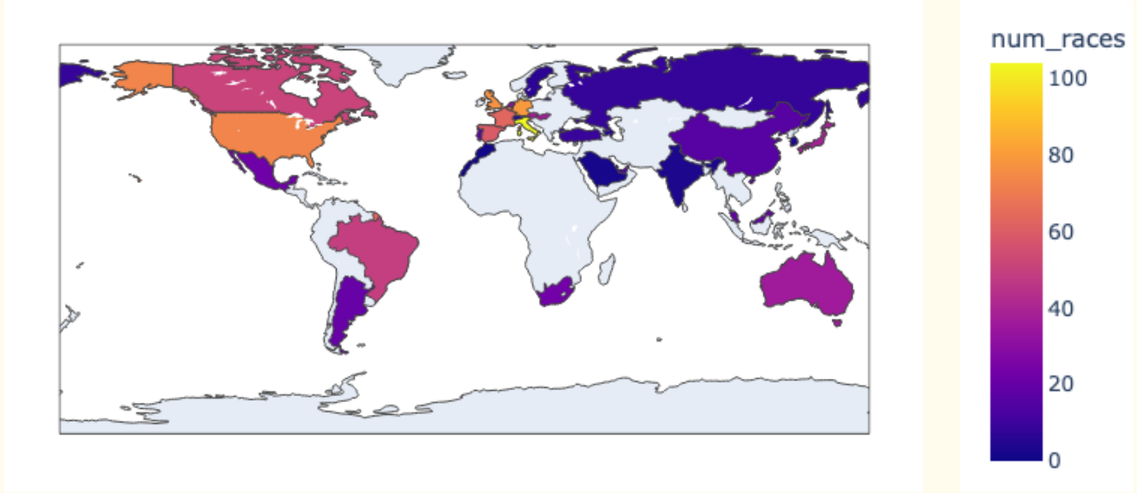
Qualifying Results

Driver & Constructor Standings

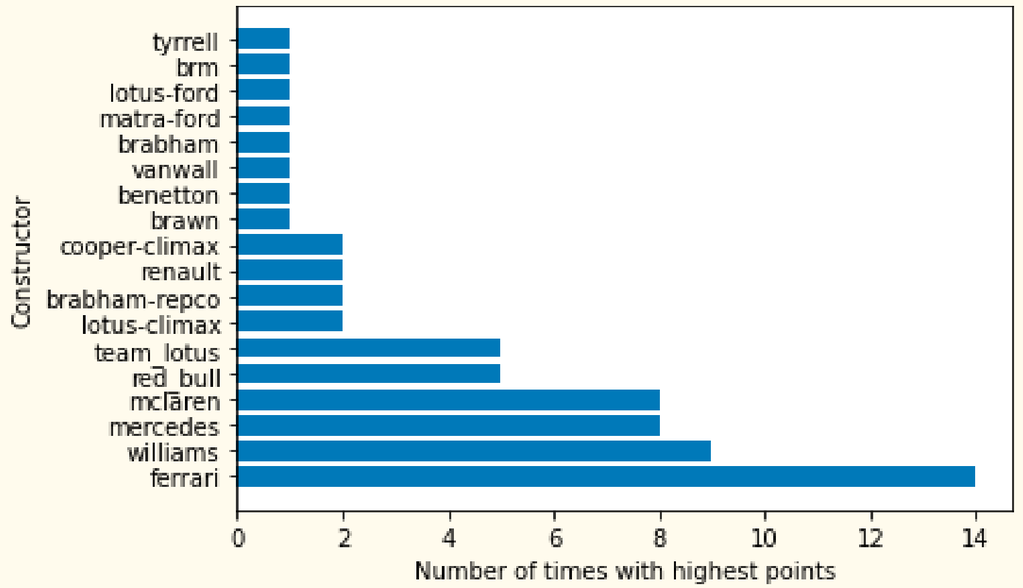
Race Stats

Exploratory Data Analysis

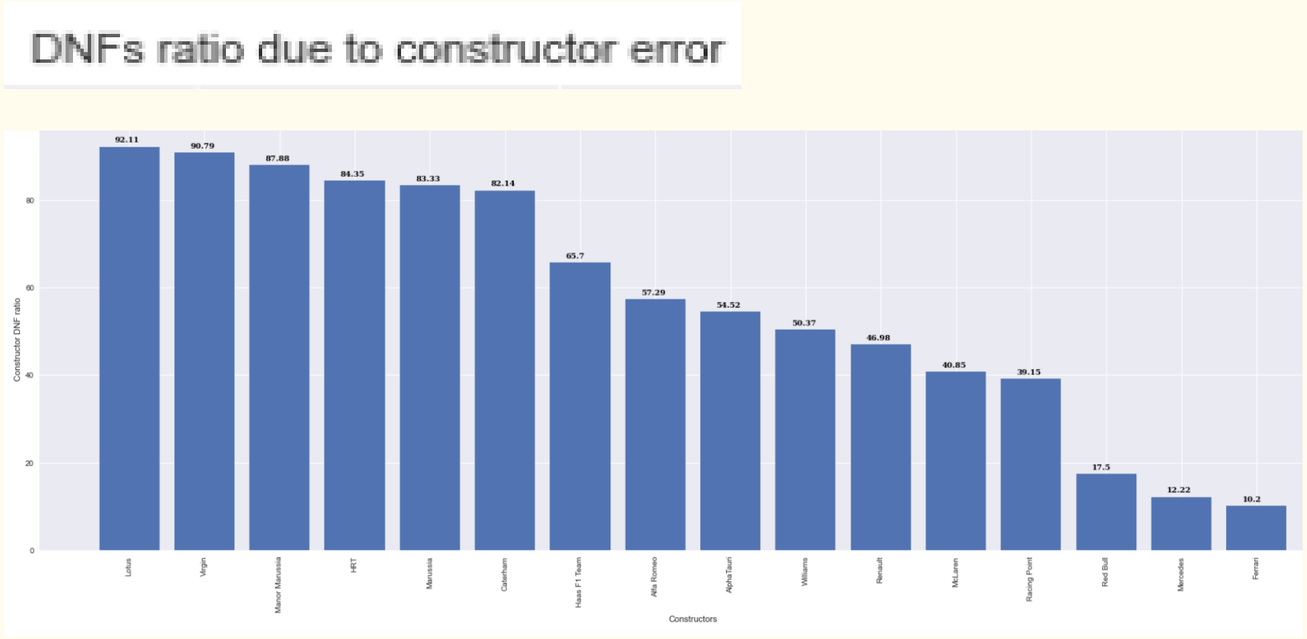
Number of races that have taken place in each country

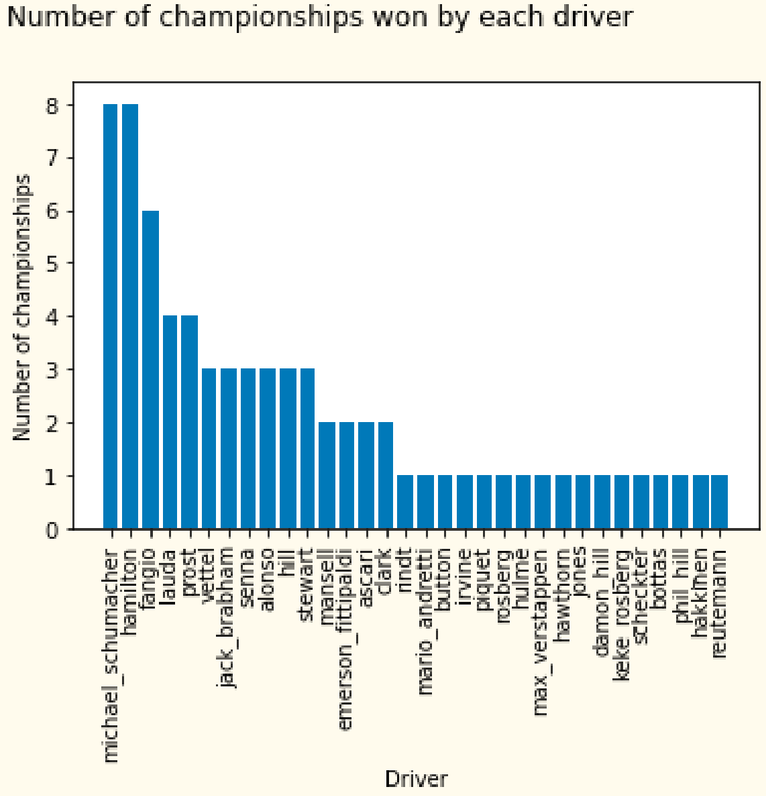


Number of times each constructor had the highest points at the end of each season



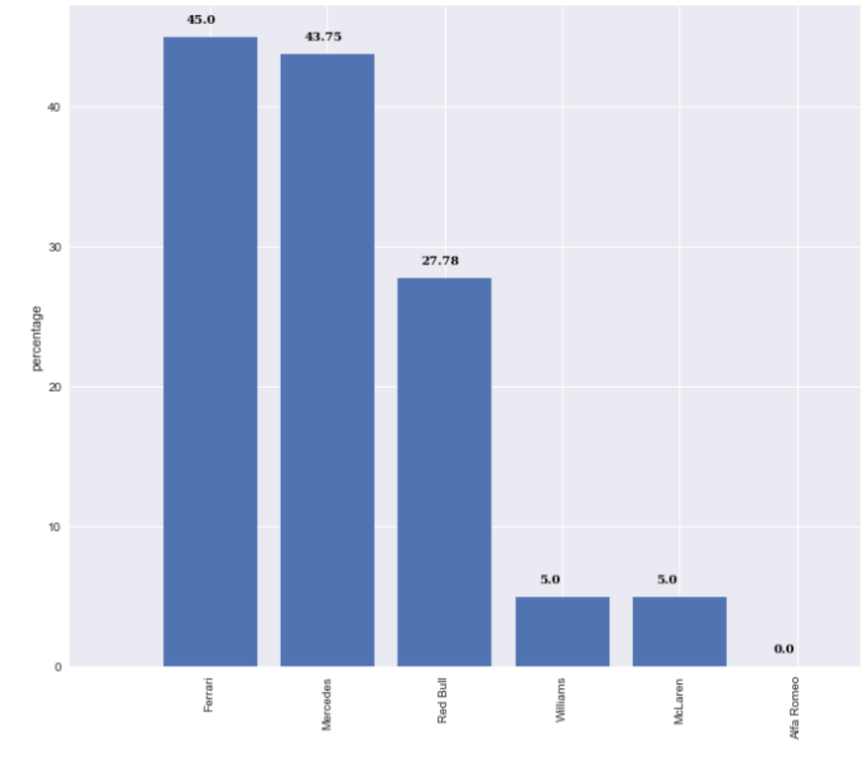
Exploratory Data Analysis



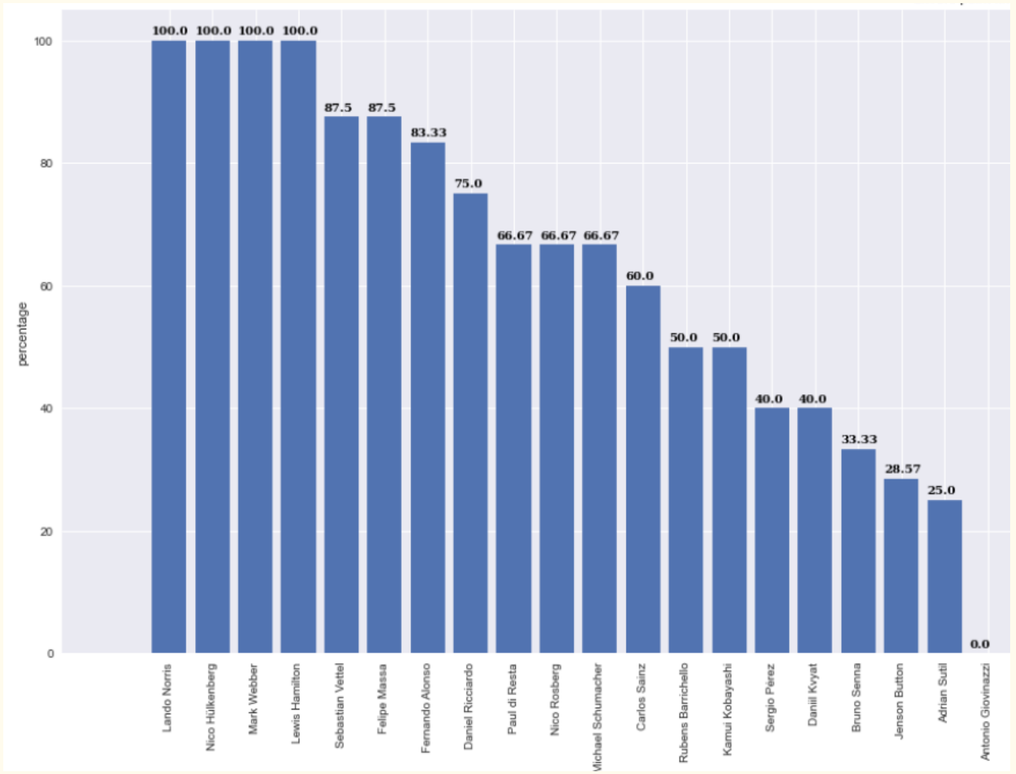


Exploratory Data Analysis

Constructor podium finish percentage at home race



drivers point finish percentage at home race



Data Preprocessing

Encoding the Data

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

Collated Data Frames

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

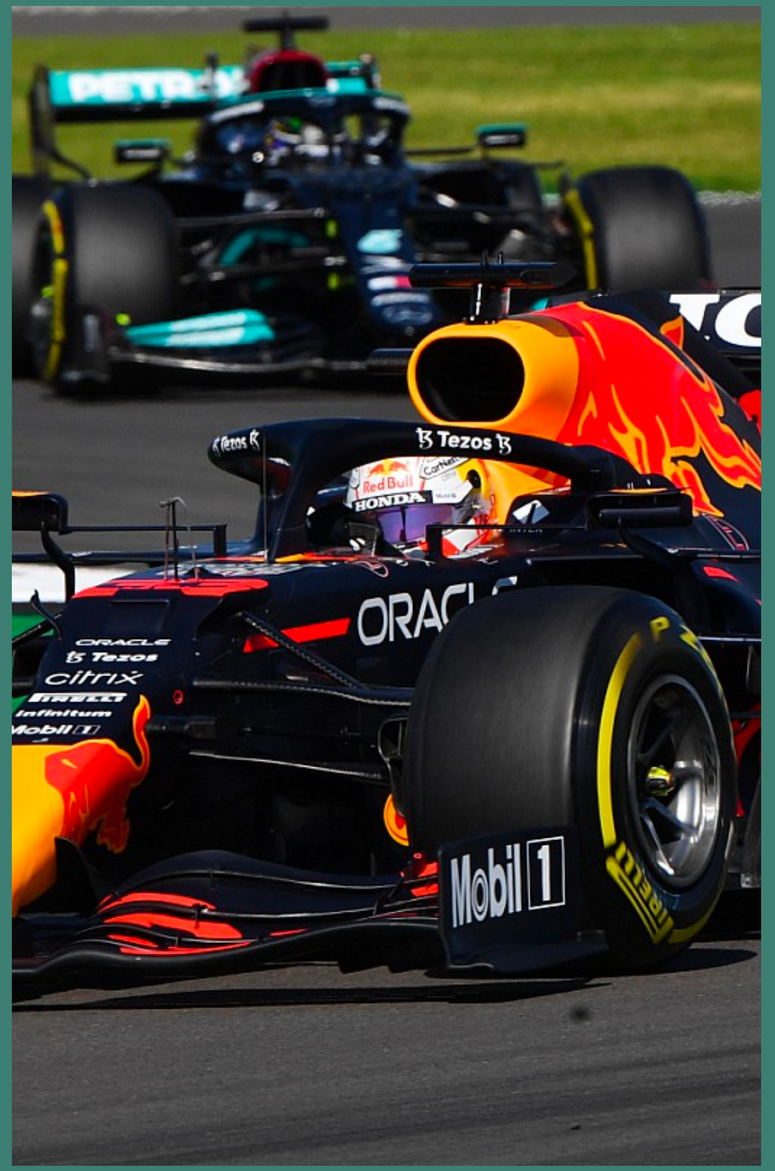
Scaled Data:

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

SHORT BRIEF

Machine Learning Models

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

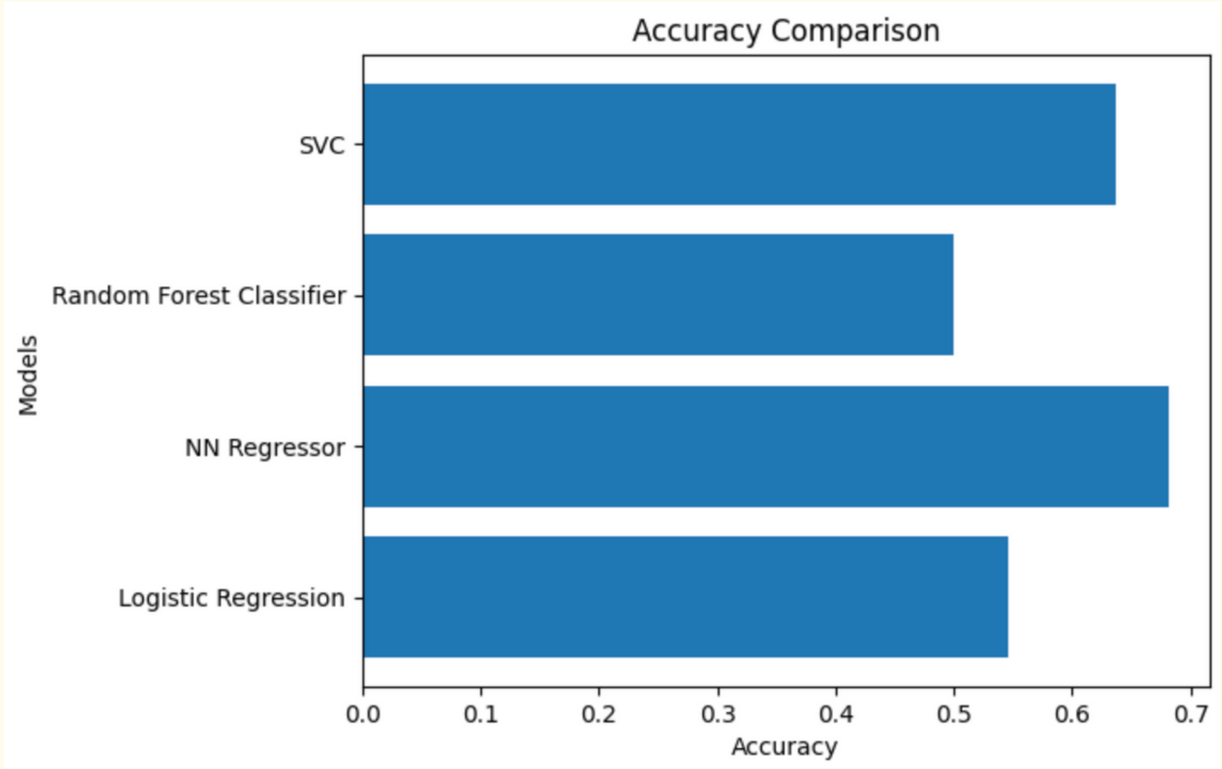
- We explored two approaches to solve the problem:
  - **Classification:** We built a model that categorizes the drivers into podium range, points range, or no points range.
  - **Regression:** We sorted the values according to the predicted values and placed them into the appropriate buckets

- Tuning the hyperparameters of our models was crucial for achieving high accuracy. After several attempts, we used a for loop to iterate through all possible combinations and selected the model with the best accuracy.

- Our selection included a range of models such as logistic regressor, linear regressor, neural network, SVM classifier, SVM regressor, random forest classifier, and regressor.

- Our best models yielded an accuracy of around 60%, which was lower than our initial expectations. This led us to investigate our approach further and identify areas where we could improve our model's performance.

Preliminary Analysis



What Did Not Go Well

During our preliminary run of models!

Importance of DNF ratios:

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

Identified unnecessary factors:

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

Issues with hyperparameter tuning:

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

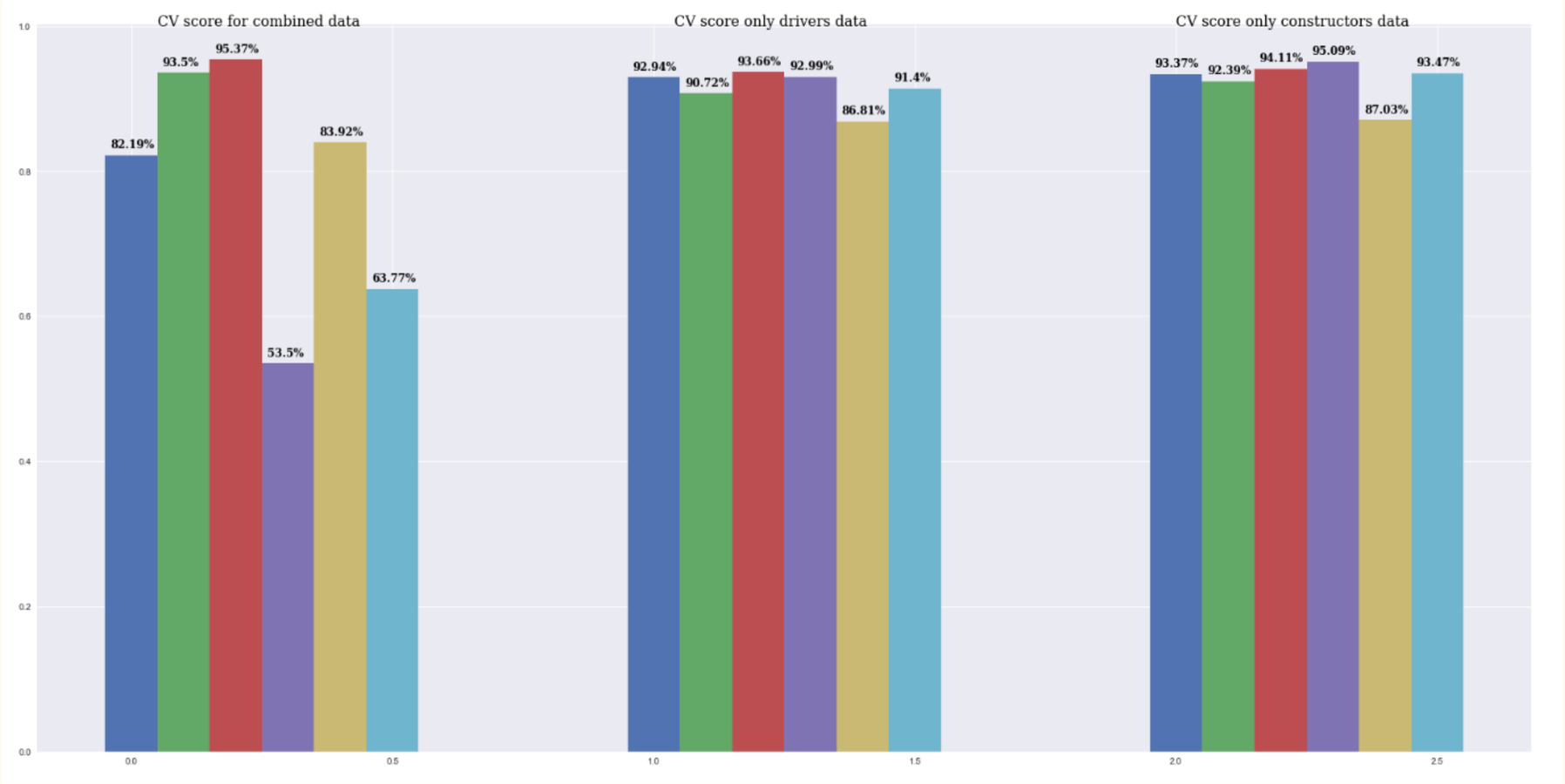
PART 2

Feature Engineering Modifications

Methods and Discussions

- Data analysis identified DNF ratios and home team advantage as important factors not considered in our previous approach. We aimed to evaluate the impact of these factors on our 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. We 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, we 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, we 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, we focused on predicting the exact race positions of drivers.**
- We discovered that race positions were highly dependent on qualifying lap times, starting positions, and other factors considered in the previous analysis.
- We updated our 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, we were satisfied with the performance of our model.

What Did Not Go Well

During our next iteration!

Impact of missing data on model predictions

Our 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

Our 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 we could not access this data before the weekend.

PART 4

Predicting Qualifying Results

WILLIAMS RACING

Methods and Discussions

- We 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 we realized a random forest would still be better

- Handling of non-linear relationships: Random Forest can handle non-linear relationships between features and the target variable, which is vital in case of F1 prediction as there are likely to be non-linear relationships between factors such as track type, car performance, and weather conditions, and the final outcome of the race.

- 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. We realized that predicting qualifying results requires more research and data.

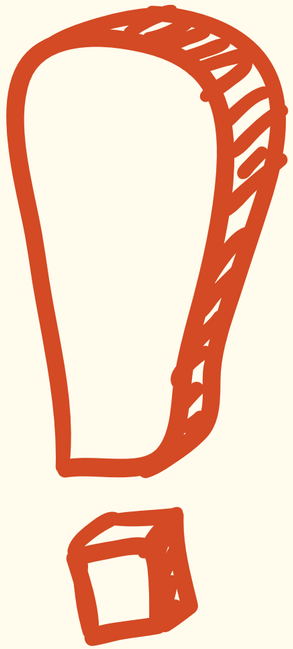
- Formula 1 is an unpredictable sport with constantly changing variables:
  - Qualifying results can provide accurate data on car and driver performance.
  - More data, expertise, and research are needed to predict qualifying results effectively.

- A key realization we 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, we 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 us to predict their positions in the race more accurately.

WHAT SEPARATES US

Novelty



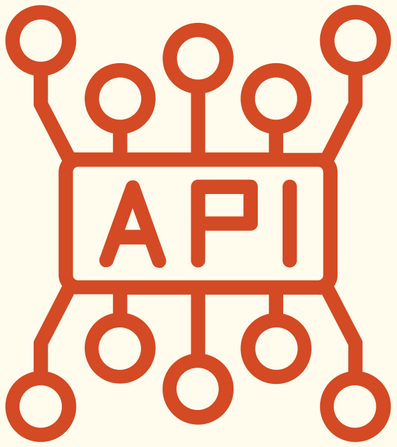
Our project applies data analysis and machine learning techniques to predict driver performance in Formula One races, providing a unique and data-driven approach to predicting outcomes.

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

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

WHAT WE WANT TO ACHIEVE

Applications



Our 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

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

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

Our 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.

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

FAIR AND SQAURE

Team Contribution

All of the team members have been an integral part in all of the stages of the project and have put their hard work into it at equal levels.

- **Jaideep Guntupalli (2020378):**
  Data Collection , Modeling & Inferences , Documentation.

- **Ritvik Pendyala (2020096):**
  Data Collection, Modeling & Inferences, Documentation.

- **Tejdeep Chippa (2020253):**
  Data Preprocessing & EDA, Modeling & Inferences, Documentation.

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 do.