Formula 1 Qualifying Result Prediction

Ben Palmer – benpalmer470@gmail.com
Brain Station Data Science Capstone Project
November, 2022

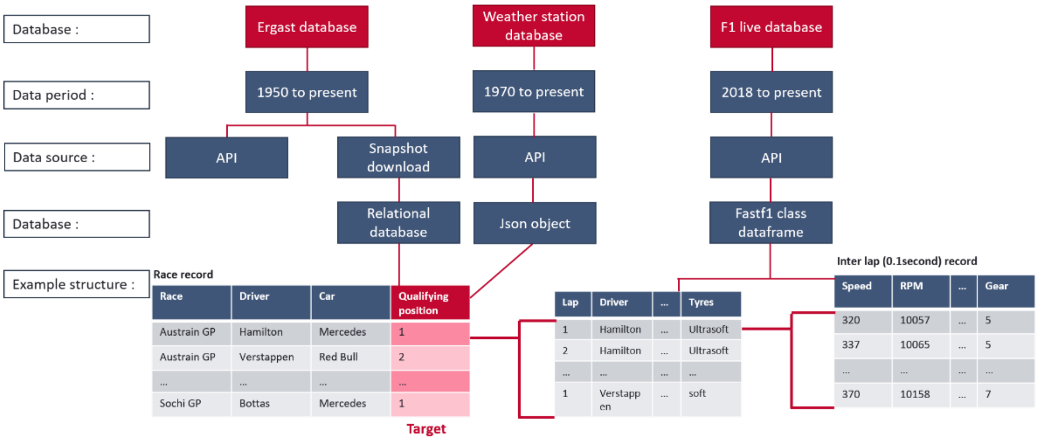
Introduction & Background

The aim of this project is to predict the qualifying results for Formula1 races. Formula1 is the pinnacle of motorsport where drivers and teams compete against each other over multiple races around the world, to be crowned world champion. For each grand prix the drivers start the race from a standing position, where the order is based on their results from the qualifying session, on the previous day. The teams and drivers seek to maximise their position in the qualifying session in order to give themselves the best chance of performing well in the grand prix. Therefore, having the knowledge of what car and driver features are likely to maximise their performance, for a given track, could be advantageous to the team. In addition, the ability to predict the team's performance based on their car and driver combination can be useful for the team to focus on their areas for improvement.

Formula1 is a popular sport, and a few Data Science projects have been completed previously, as showed in Table 1.This project is unique in relation to previous ones due to its innovative approach: prediction of qualifying results using a combination of race results, from 1950 to present, and telemetry data, car sensor data, from the F1 live database.

Data Acquisition and Description

The data used in this project is from three primary sources, with different data structures and databases. The data structure adopted in this project is summarised in Figure 1.



Race record

Race	Driver	Car	Qualifying position
Austrain GP	Hamilton	Mercedes	1
Austrain GP	Verstappen	Red Bull	2
...	...	...	...
Sochi GP	Bottas	Mercedes	1

Lap	Driver	...	Tyres
1	Hamilton	...	Ultrasoft
2	Hamilton	...	Ultrasoft
...	...	...	...
1	Verstappen	...	soft

Inter lap (0.1second) record

Speed	RPM	...	Gear
320	10057	...	5
337	10065	...	5
...	...	...	...
370	10158	...	7

Figure 1: Illustration of the 3 primary sources of data used in the project and the database structure

The Ergast API contains legacy results of Formula1 races and qualifying sessions, from 1950 to present. This is stored in a relational database, of which a snapshot was downloaded on the 24th September, 2022. The database can also be accessed via an API, and the database is continuously updated after the latest races are completed, allowing for further analysis. The Ergast database was supplement with weather data from a weather API and web scrapping from Wikipedia, used for the initial analysis. Since 2018, Formula 1 hosts live sensor data from the cars around the track. This data was downloaded using the FastF1 python library written by Philipp Schaefer et al. The telemetry data from F1 live consists of lap and inter-lap sensor data, which is a different record level to the Ergast race results data.

This project focused on the combination of the F1 live telemetry data with the Ergast race results, during their concurrent period of 2018 to present, to predict qualifying race results.

Interesting Insights from Exploratory Data Analysis

A number of interesting insights were gained from the Exploratory Data Analysis. Some examples are showed in Figure 2.

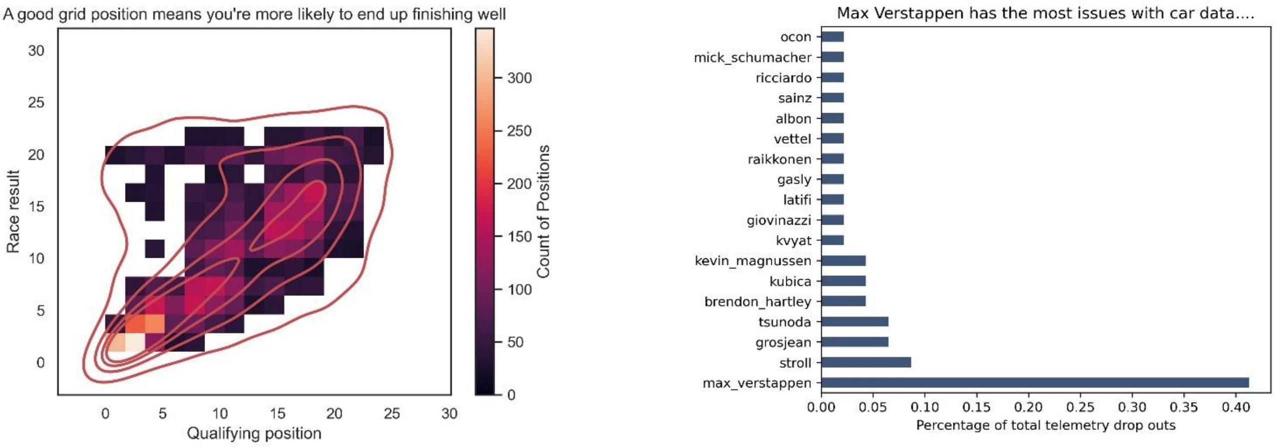


Figure 2 Two dimensional histogram of qualifying position vs race result (left). Percentage of qualifying sessions where no car data was available, per driver, on qualifying day (right).

The two dimensional histogram shows the strong correlation between qualifying position and race result, especially for the top 5 results. From the legacy data, it was observed that starting in first position gives the drivers a 50% chance of winning the race. Whilst outside the top 3, the drivers have less than 15% chance of winning. Hence, it is beneficial to maximise qualifying position. Another interesting insight is the null days observed in the telemetry database, i.e., when a car did not record any telemetry data for that day. Max Verstappen, one of the main protagonists of Formula1, suspiciously stands out as the driver with most null days.

Feature Aggregations

As mentioned previously, the two main data sources used in this project are at different record levels. To be able to predict qualifying results at a race and driver input level, the telemetry data needs to be aggregated into a race and driver record. This is a critical step of the project, as the predictive power of the model is controlled by the ability of the feature aggregations to accurately represent the car and driver performance in the lap. During this step, 80+ features were created to represent the telemetry data, such as: time spent braking in a lap, variance of engine's RPM, and minimum speed in the corner. The feature aggregations were first run on a sample dataset, to ensure features were predictive, before running it on every Grand Prix, from 2018 to present day.

Modelling & Conclusions

The data consists of all qualifying races from 2018 to present day at a driver record resulting in 1900 records (20 drivers * 20 races * 5 years). This is a small dataset for machine learning to predict, especially when considering the large amount of features generated (80+). Therefore, the project had a great emphasis on feature importance and dimensionality reduction. A summary of the key findings is the following:

- The regression models XGboost and Random Forest performed the best when using a subset of the most important features selected by the Random Forest feature importance method. These models outperformed the classification models since the Mean Absolute Error (MAE) was smaller than the bin size of the classification models.
- Across the test dataset, i.e., the 4 Grand Prix's which occurred during the project, the best performing models had a R^2 score ranging from 0.5 to 0.6, with lowest MAE of 2.8 for qualifying position and lowest MAE of 0.8 s for lap time delta.
- The models indicated a number of interesting features that impact performance, including:
    - The variance of the engines Revolutions Per Minute (RPM) on the straights
    - The maximum speed the drivers carried in the corners
    - The distance a driver spent on the brakes in a given lap
- The models had a tendency to overfit, likely due to the small dataset and high dimensionality
- In some circuits the models did not perform well, especially on lap time delta predictions. For example, the prediction of lap time delta for the Singapore Grand Prix presented a high MAE, likely due to this race being a street circuit which is very different to the other circuits, which are mainly open race tracks.

Recommendations

Overall, the project has developed good initial models to predict Formula's 1 performance, and it has identified a number of interesting insights on the important features related to Formula's 1 qualifying results. The models did not predict the performance results accurately (low MAE) most likely due to the features variance being predominantly caused by the different circuits rather than the driver's performance. Therefore, it is recommended to iterate on feature aggregation to normalise the features in relation to the circuits. In addition, it is also recommended to investigate alternative techniques of feature aggregation, such as, building an auto encoder to extract features related to lap performance.

This project focused on using telemetry data from the qualifying sessions themselves to predict the qualifying result, which is useful to illustrate the important features that contribute to a good performance. To use the model as a prediction ahead of the qualifying session, the features need to be generated based on likely similarity to past performance. An alternative, for a future project, is to use the telemetry data from Practice 3 sessions, ahead of the qualifying session, to predict the results.

Appendix & References

Project link	Database	Prediction
https://www.f1-predictor.com/	Unknown	Race Results
https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da	Ergast	Race Results
https://towardsdatascience.com/reinforcement-learning-for-formula-1-race-strategy-7f29c966472a	Simulator	Tyre Strategy
https://medium.com/@michael45684568/formula-1-and-modeling-e610f837d3ac	Ergast	Pit stop strategy

Table 1: Previous Data Science projects completed with Formula 1 data