Predicting Formula 1 Qualifying Order and Race Outcomes

Introduction

Good day,

The research proposal project is on predicting sports results. Specifically predicting Formula 1 (F1) qualifying order and race outcomes. The objective of this project is to use machine learning techniques and current relative literature throughout the project. This is done to maintain academic integrity and fulfil the project requirements. With the abundance of raw data F1 has to offer. The opportunity to present a research project on this domain is high, though not much research has been done on it. At least very little public data on predictive models that aim to accurately predict qualifying and race results.

F1 can be summarized as a team motor racing sport. Where a team develops a car that has to fit specific regulations. In 2025, 10 teams race 24 weekends a year with 2 vehicles each. Each team's goal is to gain as many points as possible to win prize money based on the teams finishing position at the end of the season. F1 weekends have free practice sessions to allow the teams and drivers to understand the track, make predictions for the rest of the weekend and adjust their cars accordingly. A qualifying session which allows teams to set the fastest lap times, where the faster the time, the closer to the front the team will start the race. Then the race, which is where points are handed out depending on the finishing position. With some weekends offering a second qualifying and race called a sprint qualifying and sprint race session. Where in these addition sessions, less points are offered than the main race but allows for more racing to be done over the race weekend. (F1, 2025)

I will be using the data from all the practise sessions to predict the qualifying order and the race outcome, as well as, use all the practice and qualifying sessions to predict the race outcome. The contribution I am looking to make is to determine the main factors that come into play in practise that determine the top qualifiers and finishes, as well as, show the significance qualifying makes compared to the other significant variables to the final outcome of the race. This will be done with two models being selected and using a single quantitative data source. FastF1 contains the lap times, car, tire and weather data as well as the session results. With data available as far back as the 2018 F1 championship. This is the best option for data due to being the only public, regularly updated and ready within 2 hours of any session. Limitations include no data prior to 2018 and no data on car aerodynamics or changes made throughout the race weekend by the team to their car. (FastF1 ,N.D)

The contribution I will make to the discipline of F1 and motor racing in general is how accurately can the qualifying order be predicted based on multiple variables of data collected during prior sessions. I want to determine what the biggest factors are that contribute to the predictions. With hundreds of millions of pounds being made and spent by each team, understanding key factors that go into out qualifying your opponents and winning races is the difference between profiting and dealing with financial struggles. (Franks, 2025)

Key literature related to the project

Three pieces of literature were found that inspired the proposal of this project.

Name of literature	Summary	How it relates to my Project
A Data-Driven Analysis of Formula 1 Car Races Outcome	5 years of formula 1 data from 2015 to 2019. Taking into account variables like average number of pit stops, tire usage, laps drivers spent in each position during the season and more driver related data. Using the variables the literature aims to predict the outcome of the race using linear regression models. (Patil et al .2023)	The data contains similar data like information on tire percentage usage and the use of many factors to predict race outcomes. The paper identifies factors that correlate to more points in a season and the correlation between those factors.
The Use of Machine Learning in Predicting Formula 1 Race Outcomes	Using deep learning models, the paper predicts driver finishing and team final positions in the championship. Using variables like qualifying positions, team driver data like number of laps and overtakes completed throughout the dataset. (Urdhwareshe. 2025)	The paper focuses on multiple factors using a deep learning model called TabNet to predict race outcomes. Though the paper lacks variables compared to the other key literature we will be using, the deep learning approach that is not commonly seen will help when comparing model performance.
Machine learning framework for formula 1 race winner and championship standing predictor	Using Ergast F1 API and weather data from F1 WIKI, the paper uses three models focusing on advanced regression and ensemble learning algorithms to predict the race winner and team standings. (Sicoie, 2022)	The paper uses Ergast F1 API which is the precursor to fastF1 and it integrates weather data which FastF1 contains. The model research done on these variables will assist decision making.

Aims and Objectives with research questions

Based on the literature and data available, the thesis will attempt to answer:

1. How accurately can qualifying order be predicted?
   This will be attempted using a machine learning model trained on past data from previous races and practice data from the current race weekend. The goal is to determine the qualifying order of all 20 drivers.
2. How much does post-qualifying information assist with race result predictions?
   After the first prediction is made on qualifying and the race. The next model will receive the qualifying results and make predictions based on those results. Demonstrating the significance of qualifying by the factor of change shown in the

race predictions compared to the new race predictions on accuracy to the actual race result.

3. What factors contribute the most to qualifying and race results?
   Of all the data used in training the models on predicting the qualifying and race result. Which factors played the biggest role in predicting the results.
4. What else can be predicted? And future research
   During Exploratory Data Analysis (EDA), as the data is further understood. An attempt will be made to identify additional questions or insights that could be further investigated in future research. This includes predicating the qualifying pole lap time or the lap the cars would make a pit stop.

Methodology/Development strategy/Research Design

The FastF1 site provides detailed guides on how to accurately read the data provided using Python. (FastF1, N.D) Including examples of how to pull specific data and plot the data. The first step of the project will be an EDA to understand the data first hand. This falls under the data understanding section of the CRISP-DM style workflow. (IBM, 2021) Which will be the workflow we will be following throughout this process. Since the data is unfamiliar, a large portion of work will go towards data understanding. Missing data will be handled after understanding why they are missing. Patil et al (2023) replaced missing values with 0's as it found most missing values were due to drivers not participating in the event after a crash or other issue that prevent the driver from carrying on the race. Where Urdhwareshe (2025) filled in missing values in telemetry data using methods like k-nearest neighbour and predictive mean matching so the telemetry data flows and is not disrupted by 0's which could cause outliers or skew the data. More data preparation methods will occur as the data is further understood.

The next step is to model the data. The key literature used linear, deep learning and advanced regression models. The linear models used all data to train the model to explain what caused points to be scored. The deep learning and regressions models trained on previous years to predict the most recent years results. We will use a similar method as the deep learning and regression models to train on past data to predict unknown results.

The models evaluation methods included several widely used regression methods like R squared method which gave 99% accuracy to the linear model, around 70% for the deep learning models and around 90% for the advanced regression models. This suggests using advanced regression models when predicting qualifying and race results is best, as the linear models had no unseen data. Using SHAP evaluation techniques on the advanced regression models would provide further insight into model performance and understanding how each feature contributed to the models predictions. Which will assist in answering research questions 1, 2 and 3. (ApX, N.D). Like Sicoie (2022), Multiple advanced regression models will be used and compared to each other to find the best model to predict the most accurate results. The benefit of this is maybe one model is better at predicting qualifying results based on practise session data while another is better at predicting race results based on practise and qualifying data. That is the advantage of having multiple types of advanced regression models and parameters. The limitations are that deep learning models may hold more accuracy if I test parameters further than in the key literature. While deep learning models have the disadvantage of requiring large data sets. FastF1 data is a relatively small dataset with 24 races per season and 20 drivers. 480 samples per year and only 8 years of data. Making the dataset size better suited for regression models as seen in the prediction results of key literature.

Ensuring reliability and validity throughout the workflow needs to be held. Reliability will be achieved by providing all code and methods used as well as explanations with them. The methods used will be done with widely used tools like Python, standardised libraries in Python, and model evaluation methods like SHAP. This allows the study to be repeated and understood when run in the future. The only domain out of my control is the accessibility to FastF1. The only data source. Which is currently free to use if used for educational purposes.

Validity will be upheld by training and testing the data on the splits the project states they will be on and not leaking any test data into the training data. Mirroring real-world forecasting on qualifying and race sessions.

Ethical considerations and risk assessment

FastF1 is released under the MIT license. Which allows free use, modifications and redistribution for research purposes.

Although FastF1 dataset contains no personal or sensitive information which means no ethical issue could arise from presenting all data publicly. Dominant teams like multiple world team champions Mercedes and Red Bull may distort predictions and hence be biased. This will be acknowledged in the EDA.

While the research is academic. With artefacts such as predictions being posted publicly before official results are released. The research may be misused for gambling or betting contexts. Clear disclaimers will be included on artefacts to demonstrate this is strictly for educational and research purposes only.

Therefore, this data is low risk but has the potential to cause harm if misused.

Description of artefacts that will be created

Basic artefacts include the predictive models being built, the SHAP results and other evaluation methods as well as multiple graphs.

Documentation and interactive demonstrators using GitHub and web-based dashboards. Since these features are free and easily done using features on GitHub. Documentation boost's reliability and validity and web-based dashboards could allow easy accessibility to certain races to view predictions againts actual results. As well as show graphs on the best predicted races or best overall predictions per each season and model. Overall improving the analysis and understanding of the project outcomes and performance.

Timeline of proposed activities

Since this is my first thesis, the preliminary timeline will be 8 months, based on the rough timeline given by the institute I am studying with. The module should begin around November 2025 and the thesis is expected June 2026.

Time	Task	Description
Month 1 (Week 1-3)	Further literature review and feedback from lecturers on this research proposal.	Using the feedback given I will use further literature to make improvements to those sections that require it.
Month 1 (Week 4)	Set up project infrastructure on GitHub repo.	Spend time working on pipelines and notebook skeletons necessary.
Month 2	Data collection, understanding and preparation.	Using FastF1, I will spend time on the data and preparing the EDA and any further ethical or Bias assessments.
Month 3	Modelling and evaluation	Modelling of the data will take place. Looping through multiple parameters to find the best models through evaluation techniques like SHAP
Month 4	Advanced comparisons	Further evaluation comparing models and truly understanding the output
Month 5	Re-evaluation	Reevaluate the project, and evaluations compared to the research questions and project overview. Making sure no strays off paths occurred or misalignment to the final expected project.
Month 6	Conclusion and finalization	Conclude and finalize the project. Provide a solid conclusion, evaluating the project as a whole and make sure it looks and reads like a top academic thesis.
Month 7 and 8	Buffer	Any task could take longer and the 2 month buffer is providing to add week/s to any task in case of issues with the project or personal ones.

Conclusion

The proposal provides a framework, timeline and backed research for predicting Formula 1 qualifying order and race outcomes using FastF1. By addressing the research questions, understanding the data and the outputs of the models. The final submission should provide new insights into performance determinants in F1. Including research questions based on future research, giving inspiration to future papers made by myself or others on the theoretical never-ending questions to be answered by big data like F1 data.

The project integrity not only lies in the application of multiple modelling strategies, or the double barrel prediction method, but also in the use of SHAP model evaluation technique which will assist in understanding, based on the data provided, what key public data metrics give insight to which drivers will be on top in qualifying and on top on the race. In doing so, the project shows how predictive analytics can extend beyond descriptive race statistics or assumptions made during the early weekend hours of the practice sessions.

Ultimately, the project aims to describe the key data in practise and qualifying that harnesses measurable accuracy. While deepening understanding into the dynamics that drive success in Formula 1. By achieving this, I hope the project makes a meaning contribution to both academic study of predictive modelling and sport analytics. Perhaps the stepping stone required to one day work on specific team data only accessible to an employee of an F1 team.

References

ApX (N.D) Hands-on Practical: Calculating SHAP Values. Available from: https://apxml.com/courses/model-interpretability-explainability/chapter-3-shap-additive-explanations/shap-hands-on-practical [Accessed 4 October 2025]

F1 (2025) Everything you need to know about F1 – Drivers, teams, cars, circuits and more. Available from: https://www.formula1.com/en/latest/article/drivers-teams-cars-circuits-and-more-everything-you-need-to-know-about.7iQfL3Rivf1comzdqV5jwc [Accessed 27 September 2025]

FastF1 (N.D) FastF1 For the passionate F1 nerds. Available from: https://docs.fastf1.dev/ [Accessed 27 September 2025]

Franks, T. (2025) The Economics of a Formula 1 Team. Available from: https://www.grandprix247.com/2025/03/28/the-economics-of-a-formula-1-team/ [Accessed 27 September 2025]

IBM (2021) CRISP-DM Help Overview Available from: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview [Accessed 12 September 2025]

Patil, A., Jain, N., Agrahari, R., Hossari, M., Orlandi, F. & Dev, S (2023) A Data-Driven Analysis of Formula 1 Car Races Outcome. Available from: https://link.springer.com/chapter/10.1007/978-3-031-26438-2_11 [Accessed 21 September 2025]

Sicoie, H. (2022) MACHINE LEARNING FRAME WORK FOR FORMULA 1 RACE WINNER AND CHAMPIONSHIP STANDINGS PREDICTOR. Available from: http://arno.uvt.nl/show.cgi?fid=157635 [Accessed 21 September 2025]

Urdhwareshe, A. (2025) The Use of Machine Learning in Predicting Formula 1 Race Outcomes. Available from: https://www.preprints.org/manuscript/202504.1471/v1 [Accessed 21 September 2025]