National College of Ireland

Predicting Formula 1 Podium Finish Probabilities Using Machine Learning: A Comparative Study of Pre-Race and Within-Race Feature Scenarios

MSc Research Project
Data Analytics

Likhita Sargur Yathishkumar
Student ID: x23203951

School of Computing
National College of Ireland

Supervisor: Jaswinder Singh

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Likhita Sargur Yathishkumar
Student ID:	x23203951
Programme:	Data Analytics **Year:** 2025
Module:	MSc Research Practicum
Supervisor:	Jaswinder Singh
Submission Due Date:	27 Jan 2026
Project Title:	Predicting Formula 1 Podium Finish Probabilities Using Machine Learning: A Comparative Study of Pre-race and Within-race Feature Scenarios
Word Count:	6945 **Page Count:** 20 Including References

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Likhita Sargur YathishKumar

Date: 11/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Formula 1 Podium Finish Probabilities Using Machine Learning: A Comparative Study of Pre-Race and Within-Race Feature Scenarios

Likhita Sargur Yathishkumar
x23203951

Abstract

The unpredictability of the race, the sensitivity of the results to minor variations in performance, and the effect of the weather and strategy, as well as the form of the drivers, all these makes the predicting in Formula-1 a challenging task. However, precise and transparent the forecasting is, it is very useful for the teams, analysts and broadcasters who require information, based on predictions before the race and during a race. This thesis creates a machine learning model that predicts podiums finishers based on some pre-race data and within-race summaries for the performance. Multi-season Formula-1 data are combined and tested using chronological split to prevent future information leakage. Random Forest and LightGBM are the two models, evaluated and compared and later the winner that is chosen on validation of Macro-F1 and log-loss. The quality of probabilities is evaluated with Brier and macro one-vs-rest AUC. It is demonstrated that the informative, well-calibrated pre-race probabilities can be achieved and when the within-race aggregates are added, Macro-F1 and the calibration as an upper bound which is further benefited. SHAP analysis of the Pre-race LightGBM shows that grid position and recent form gives most contribution to podium probability. The system provides practical, probability-based, and explainable team, analyst and broadcast podium prediction.

1 Introduction

Formula One, is a thrilling sport that deals with speed, engineering and strategy which makes it exciting to watch but difficult to predict what could happen in the very next moment. F1 is one of the most data-intensive, competitive and technologically advanced sports in the world. The race involves a combination of team strategy, car performance, changing environmental changes, and skills of driver. Every race generates a wide range of data about drivers, teams, circuits, results, lap times, pit stops, and race events which can reveal valuable insights if it is analysed effectively. The huge amount of information produced, provides an opportunity to apply data analytics and machine learning to better understand and predict race outcomes. This complexity makes F1 more ideal to apply on data analytics and machine learning, which can help to discover patterns, forecast performance and also the strategic race decisions.

Figure 1: Formula One Cars on Racing Track (WallpaperSafari, 2026)

1.1 Background

The results of F1 comes from a combination of car performance, driver skill, team strategy, circuit characteristics and weather on race day. Previous works in motorsport has modelled race performance and driver vs constructor effects, (Bell et al., 2016) explored decision support as pit-stop timing but, most studies show winner, the points or even tactical problems rather than the podium probabilities with time aware validation without leakage. In sports analytics, the best practices focus on quality over raw accuracy for decision-making, and give warnings against the leakage, also recommending the evaluation (Gneiting and Raftery, 2007). These principles mainly motivates for the design of safe podium-probability models.

1.2 Motivation

Most of the previous works predicts the winners or focus on the sub problems like tyre performance, specific grand Prix, pit stop timing (El Haber et al., 2025; Heilmeier et al., 2020). In this study the need for analysts, audiences for probabilities that are reliable and comparable across races is shown. Predicting probabilities are more useful than that of single labels as they help in fair comparison across races and in decision making (Gneiting and Raftery, 2007). The pipeline gives calibrated class probabilities for the P1, P2, P3 and non-podium positions and aggregated podium finish probability for the driver ending up in any of among podiums. Here the pre-race that is deployable and the within-race scenario that is retrospective diagnostics are compared for better understanding.

1.3 Research Question and Objectives

The research question addressed in this study is: *"Can machine-learning models predict probabilities of podium finishers for Formula One Championship, and how does the performance differ between a pre-race and within-race feature set?"*
To answer the above set question, objectives are:

- To develop safe machine learning models that combines the Formula 1 race data with weather data to estimate interpretable podium finish probabilities.
- To compare the model performance between a pre-race and within-race features using two models Random Forest, Light GBM and evaluate them through metrics.

1.4 Limitations and Assumptions

In this study only F1 tables, and aggregated pit stops, and weather data are used no raw telemetry. Categorical encoding is limited to the constructorId and circuitId. Random Forest, Light GBM, are the only models which is tuned on validation, and hyperparameters are fixed before testing. The within-race scenario is retrospective which means that the pitstop and lap times aggregates are taken after the race has happened, not from live race. Weather is also simplified using average or sum, which may not capture sudden changes that will happen in the race. Both the models use class weights for imbalance whereas the final evaluation is done by macro-averaged to treat the podium classes P1, P2, P3 fairly.

Assumptions made in this study are using causal shifting to roughly measure the driver's and team's form. To summarise the lap and pit stop data to describe the performance without need for the use of raw telemetry. Weather gives a useful information of the conditions, though it changes in the race. Splitting the date by time is a good way to check if the model is working fine on other seasons.

1.5 Structure of the Report

The remaining structure of the report follows with literature review in Section 2 focusing on the previous studies in this field and how it helps this study. Section 3 is about the methodology and the framework used in this research and how the models were developed and trained. Section 4 is about the design specifications and Section 5 about the implementation of the models. In Section 6 the evaluation of the models and other comparison metrics used is discussed with required visuals to support the results. The Conclusion and Future Work section, 7 will show the aim of the paper, conclude the findings, evaluations and ends with future work.

2 Related Work

This section gives a detailed discussions of the previous studies related to Formula 1 performance modelling, machine learning prediction, within race strategy, data and telemetry systems, and probabilistic evaluation forecast. The aim of this is to show what are all achieved in the previous studies, what is missing in those and how can the gaps in these studies motive the research question for this study.

2.1 Statistical Foundations and Competition Modelling

Initial research in Formula One analytics is aimed at determining the factors of individual race results that depend on long-term factors instead of predicting individual race outcomes. In their analysis, (Garcia-del-Barrio and Reade, 2021) use several seasons of F1 outcomes and Google trend results and discover that in the cases where championships become predictable, people

lose interest since its already known. This brings out the reason why uncertainty in race outcomes are important and why predictions based on probability are worthwhile. With the help of statistical methods, such as correlation analysis and PCA, (Patil et al., 2023) investigate correlations between the qualifying position, tyre choice and race outcomes. They find that there are certain pre-race variables that are always sensitive in performance particularly in qualifying position. (Bell et al., 2016) advance this insight with a study of more than sixty years of data on race and demonstrate that driver ability and manufacturer power combine in a varying degree between seasons. More recently, (Erik–Jan van Kesteren and Bergkamp, 2023) apply a Bayesian model where the shown strengths may vary over time with the uncertainty that is additionally included in the estimates. These studies can contribute to the understanding of the organization of performance but they can more likely give descriptive contributions than prediction models. They are not generating predictions of race levels, nor are they making comparisons between what can be forecasted using the information available prior to the race, and what can be predicted using information created during the race. This imposes the desire of more predictive, uncertainty-sensitive modelling. These studies together have a good explanation of the structure of performance but on the other hand, they do not predict the outcome of individual races. They also fail to generate calibrated probabilities and compare the information before and during a race and this generates a need of more predictive work.

2.2 Machine Learning Prediction in Motorsport

The use of machine-learning in Formula One has been primarily used to determine a race winner or the identification of good-performing teams. (Priya Shelke et al., 2023) apply Support Vector Machines and other classifiers to determine the race winners according to the characteristics of the race. Their results affirm that machine-learning is capable of identifying patterns that affect performance, however, the models do not make probability estimates, they make winner labels. (El Haber et al., 2025) use the Random Forest and SHAP analysis to demonstrate the most significant features in predicting a race winner. Although this brings in transparency, it once again leads to deterministically predicted winners with no reporting of probability calibration or quality of uncertainty. On the team level, (Vidhya Shree M et al., 2025) contrast a number of models to identify the most successful constructors, and team success is also predictable. In these studies, the main predictions are mostly hard categories, and none of them assesses multiclass podium predictions, log-loss, Brier score, and reliability diagrams. They also fail to discuss the accuracy of prediction with transitioning to the pre-race features to within-race dynamic information. This creates a blank space that spurs more modelling on how racial dynamics affect probability predictions.

2.3 Operations, Strategy, and Within-Race Modelling

The other stream of research investigates in the role of decisions made in a race to affect eventual outcomes. (Heilmeier et al., 2020) create a Virtual Strategy Engineer based on neural networks that are trained on six F1 seasons. Their model suggests pit-stop time and tyre strategies depending on real-time information including stint procession and race position. This demonstrates that information inside the race is very predictive and it can be used to assist automated decisions at the strategic level. Nevertheless, this is a strategy optimisation study, not a podium predictive study. It does not evaluate multi-season multi-driver probability forecasts either. This shows that although within-race statistics is highly informative, there is no past literature that has integrated pre-race and in-race characteristics in a single probability-prediction model. This gives the incentive to use within-race features as a second scenario in this study to determine their predictive power in comparison with pre-race information.

2.4 Systems and Telemetry: Foundations for Predictive Modelling

Some of the studies review the technological infrastructures that facilitate advanced analytics in Formula One. According to (Belkovics, Krisztina Andr and István Takács, 2024), the F1 teams are gathering and sending real-time telemetry by high-frequency sensors, on-car devices, and edge computing systems around the circuit. The architecture introduced by (Kavitha and Kanishk, 2025) proposes that the incorporation of AI models into real-time strategy tools, are based on the distributed computing. These articles do not give the outcomes of the predictions themselves, yet they demonstrate that Formula One already has the infrastructure needed to implement the machine-learning models in real-time. That is to say that the biggest issue that is still awaiting them is not software or data acquisition but to establish correct, calibrated predictive software that will perform under pre-race and in-race conditions.

2.5 Probability-Forecast Evaluation

In the cases where models are not giving straightforward labels, probabilities that it generates should be assessed by using alternative statistical guidelines. The probabilistic forecasts are primarily evaluated using the framework (Gneiting and Raftery, 2007). They clarify that the predictions of good probability should be: Calibrated meaning that predicted probabilities are identical to those in the real world, and Sharp, that is, predictions are confident as you can make. They also bring the correct rules of scoring including the log-loss and Brier score which will encourage the truthful reporting of probabilities and punish overconfident or poorly calibrated forecasts. They also put emphasis on diagnostic tool like reliability diagram, and calibration curves. Evidence from sports analytics show that well calibrated probabilities is more than raw accuracy for decision-making where on the NBA betting data, choosing the model by calibration rather than accuracy will yield to better substantially returns, underscoring why the evaluation reports Brier or log-loss and reliability curves in this study alongside accuracy and F1 (Walsh and Joshi, 2024). The foundation directly applies to this study since the objective is to make predictions about podium, and not merely categorize winners. To evaluate the quality of probabilities, standard measures of accuracy are inadequate, and instead, evaluation must be done by calibration.

2.6 Research Gap

The current studies have managed to explain long-term performance trends, create winner classifiers, and illustrate within-race data usefulness in the optimisation of strategy. But, there are gaps in their studies that are summarised in below table 1.

Table 1: Summary of Related works and gaps in their study

Theme	What has Previous Work Achieved	What was missing	How this Study addresses the gap
Statistical F1 analyses	Long term factors, driver vs constructor effects	No race-level podium prediction or the probability evaluation	Podium probability models at race levels are built.
ML Prediction	Classified winners or teams. Shows important features	No multi class podium	Predict P1, P2, P3 with calibrated probabilities

Within-race Strategy study	Show in race data is highly predictive	There is no comparison of scenario predictions	Two scenario cases Pre-race and Within-race
Telemetry and Systems papers	It shows that real time data and compute are available	No deployable probability models.	It provides deployable pre-race pipeline, analyse within-race as retrospective.

These gaps in the previous studies helps to motivate the research question of this study.

3 Research Methodology

This section gives a detailed summary of the research methodology. It is always good to follow a proven industry framework as project guide to execute a safe machine learning research project. Therefore the chosen framework for this research project is CRISP-DM process model for data mining (Martinez-Plumed et al., 2020) as shown in the figure below.

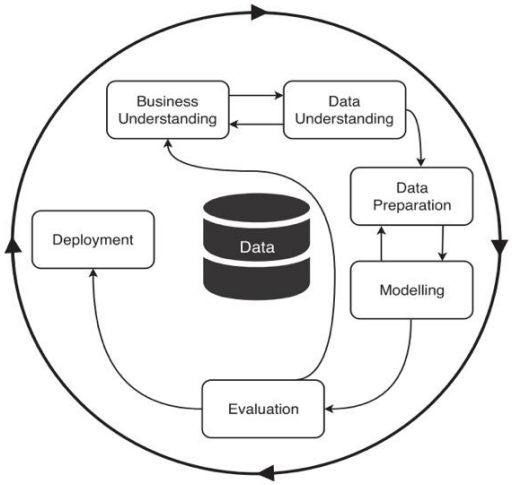

Figure 2: CRISP-DM process model

3.1 Data Understanding

In this study two datasets available publicly at Kaggle were used: The multi table www.kaggle.com. (n.d.). *Formula 1 World Championship (1950 - 2023)*. and the wsiqz (2024). *formula 1 weather info (1950-2024)*. are aggregated at the race level. The current page title of the dataset and the metadata contains from 1950-2024 whereas the URL is still showing till 2020. The data are joined using primary keys, each race (raceId) links to many driver results, every result is linked with one driver using driverId and one constructor (constructorId), and each race is hosted at a single circuit (circuitId). Qualifying results are used to get the starting position for driver-race pair. Laptimes and pitstops are summarised to get the race-level driver telemetry like laps completed, average and best laptimes, number and timing of pitstops, and others. All the weather information like temp, precipitation, wind are aggregated per race and linked using raceId.

3.2 Data Exploration

In the section all the Exploratory Data Analysis for the core datasets is shown. In the figure 3 it is seen that most drivers finish in mid to lower positions while only less number reach top.

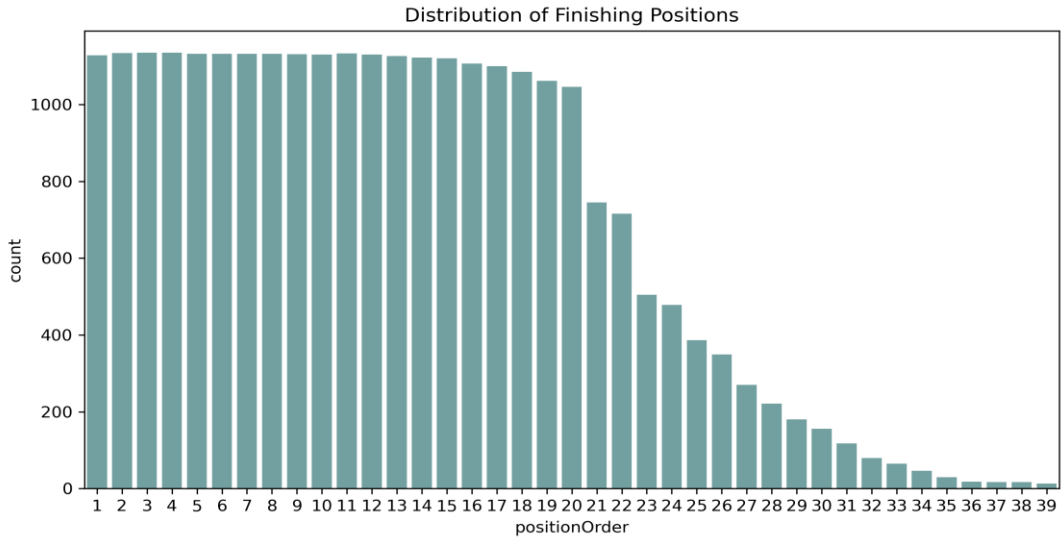

Figure 3: Distribution of finishing positions across races

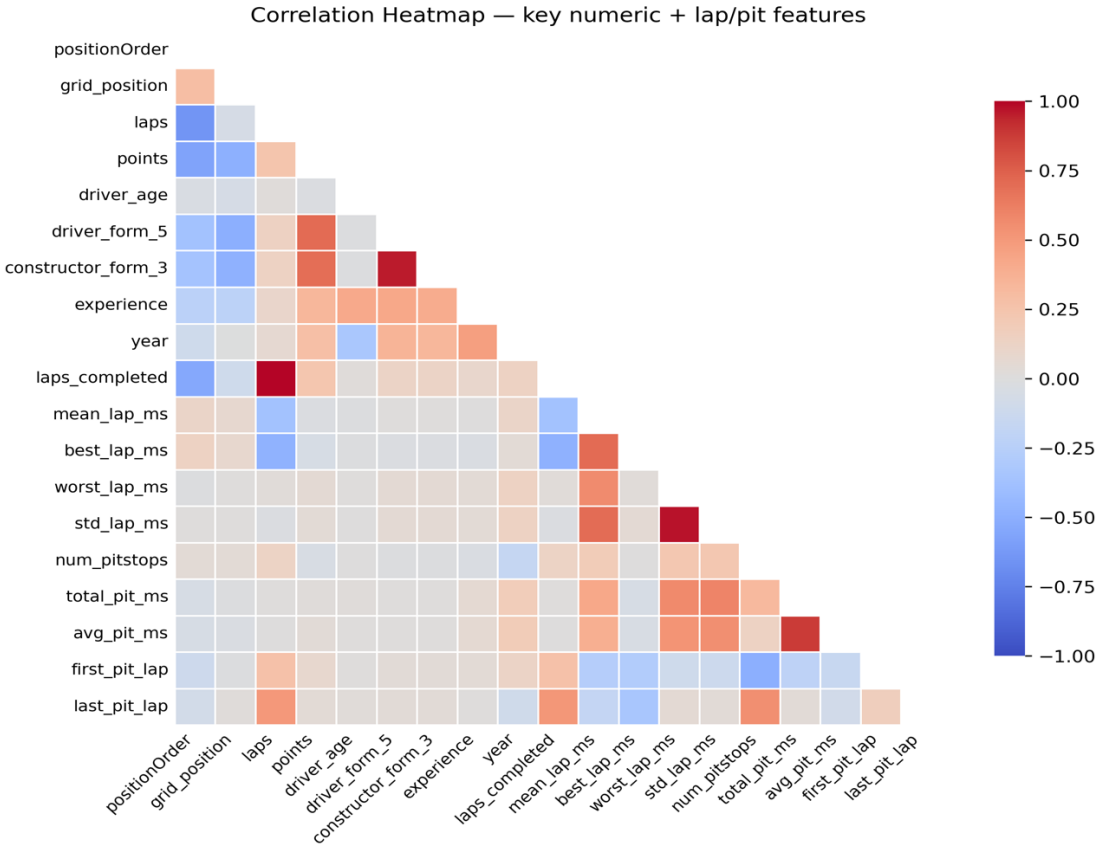

Figure 4: Correlation heatmaps of key numerics, lap-time and pit-stop features

Correlation heatmap shows that bright colours highlight the features that move together or against each other. Clear relationships with the positionOrder indicates those variables that

carry predictive signal. Highly correlated data warn about redundancy, guiding feature selection and model interpretation.

3.2.1 EDA for PRE-Race Features

The below figure 5 shows that the drivers who start at a closer to front usually finish in better positions. It is understood that downward trend in image means that qualifying performance is one of the most needed pre-race predictors of the race outcomes.

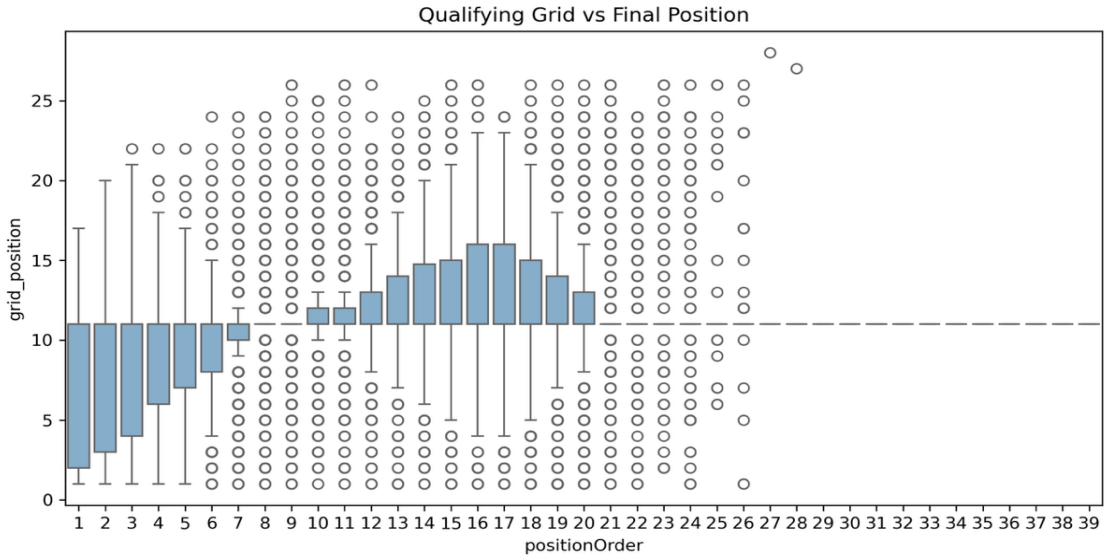

Figure 5: Relationship between qualifying (grid) position and finishing position

In this image below, 6 it is seen that how podium chances of drivers vary between dry, drizzle, and wet races. The differences show that how weather conditions can influence the race outcomes before the race begins, making this feature an important pre- race feature.

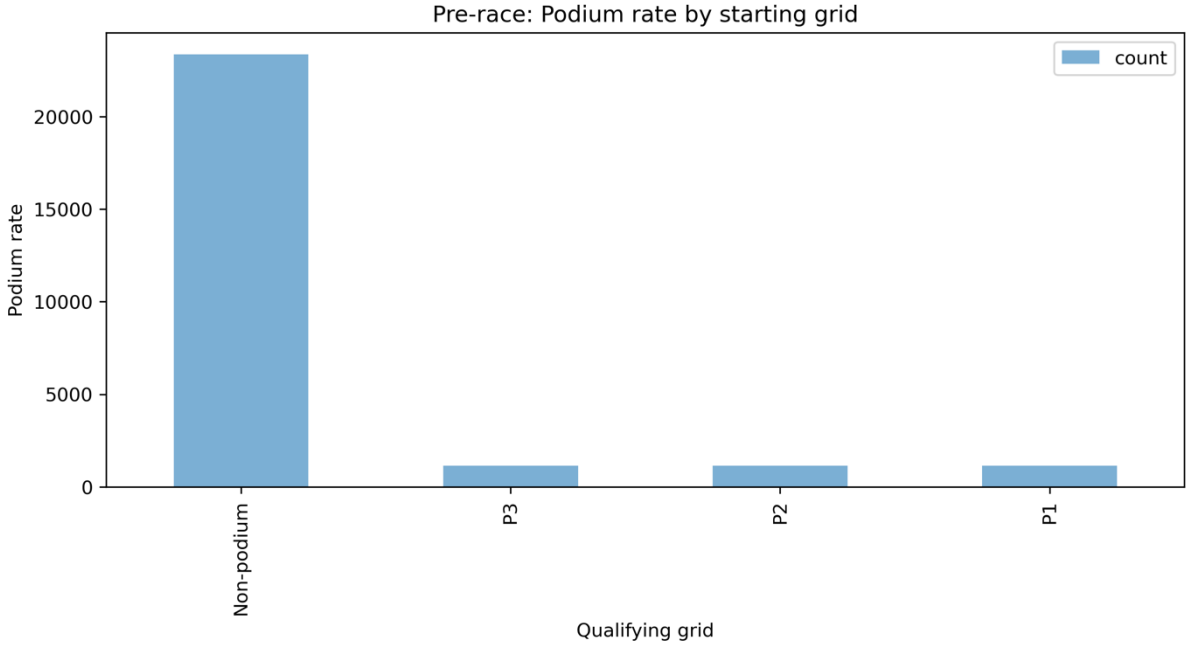

Figure 6: Podium rate under different pre-race precipitation conditions

3.2.2 EDA for Within-Race Features

The below figure shows graph of how driver's average lap time results during the race relates to the finishing position of them. It shows that faster mean lap times are linked with good finishing positions, also race pace is sustained which is a strong performance indicator. In this image 8, boxplot is used to compare lap time variability between podium and non-podium finishers. Drivers who finish on the podium show the lower lap time deviation, meaning they maintain more consistent pace throughout the race, which gives better results.

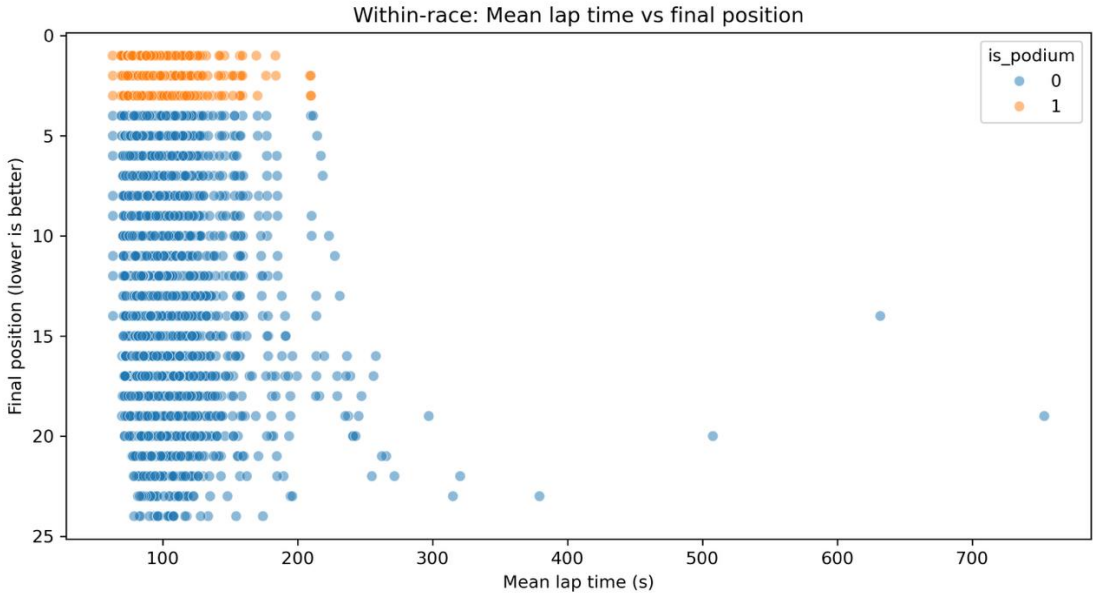

Figure 7: Mean lap-time vs Final position (Within-race)

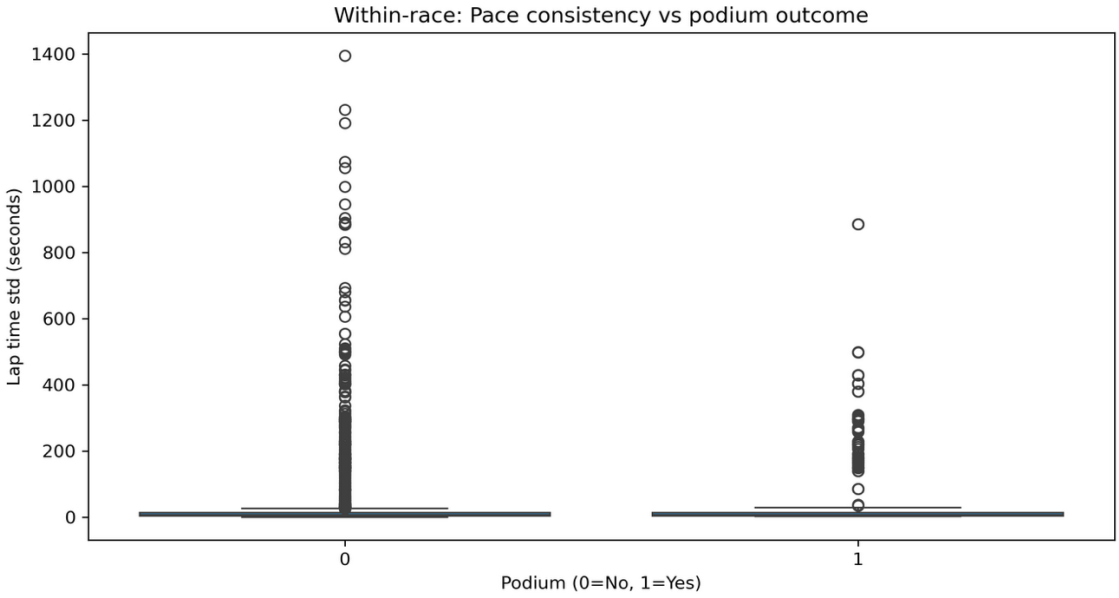

Figure 8: Pace Consistency vs Podium Outcome

3.3 Feature Engineering

Preparation The emphasis of preparation is to create a single, analysis-ready table on the level of a driver in a race. Any non-informative identifiers and fields of presentation are eliminated, and core sporting variables are enforced to be numeric. The missing values are treated in a systematic manner where the numeric data are median-imputed, text data that are categorical can default to the value 'Unknown' if there are no values and weather are set to median values

if there are no values. In order to respect temporal causality, only the prior races are used to compute driver and constructor features of form. In particular, rolling means of points along recent windows are shifted in such a way that no future information is given away on the current race. The driver_form_5 and constructor_form_3 are computed using groupby by driver or constructor, rolling windows, mean and one race shift(1) so that only past races will contribute. Driver experience is the count of prior stats. The experience of the drivers is recorded as the number of past race starts. Based on the official results, there are two targets, a binary podium flag and a four-class podium label with the distinction of P1, P2, P3, and non-podium. Aggregated observations are used to formulate weather because they mirror the environmental context. Strict chronological split of the data for train ≤2019**,** validation 2020–2022 and the test ≥2023, is done ensuring no leakage.

3.4 Scenario: Pre-race and Within-race features

In a bid to ensure that the evaluation is truthful and realistic, the study distinguishes two modelling scenarios in relation to the time when information is accessible. The Pre-Race scenario is based on information that can be known prior to the start of the race: qualifying grid position, driver and constructors recent form calculated based on previous races adjusted so that there is no leakage of future information, driver age and experience, season and circuit identifiers, and weather conditions at the race level. It does not use within-race telemetry. This is what a strategist or broadcaster may employ prior to lights out and it may be used live.
The Within-Race scenario incorporates all of Pre-Race and in-race aggregates retrieved by lap times and pit stops. These features come up after the race has been conducted hence in this research they are represented as post-race summaries. This causes the Within-Race setting to be a retrospective diagnostic indicating the extent to which additional predictive information is gained by the signals of race-event, and not something to implement prior to the race. In both cases, the training, as well as testing, will be performed using a similar chronological division to approximate real forecasting and prevent the ability to see into the future and make predictions. The Within-Race scenario uses aggregates like laptime, pitstop data that are available after the race or during the race is going on. Therefore it is a retrospective diagnostic upper bound, but not a deployable real-time forecast. The answer to two questions of how good the race features are known pre-race, how much better what happened during the race is also known, is clear in this design and information does not leak away and results can be easily interpreted.

4 Design Specification

Reproducibility is made possible by easy data automation and a configuration guide that will allow to re-run the entire pipeline using fixed seeds, scenario conditioning between the pre-race and within-race, and a strict chronological break that is train 2019, validation 2020-2022, and test 2023. It uses three levels, namely, the client tier that will execute the notebook file locally in Jupyter Notebook. To carry out the merge of the data, it goes to application tier, feature engineering, which is followed by the modelling layer, evaluation, and data or presentation layer, where the report knows the export tables. Local CSVs files are loaded to provide the databases with the races, results, drivers, constructors, qualifying, circuits, lap-times, pit-stops, and race-level weather. The storage is also flexible allowing the saving or storage of the intermediate tables in the memory. All of these are possible in the application tier: summary of the raw events, construct honest features, contrast pre-race and within-race setups. The two models are educated and tested. Results, such as figures, comparison tables, and 2024 season summaries can be stored to be used directly.

5 Implementation

In this section the architecture design that was considered was implemented. The two pipelines, complete predictive ones each corresponding to a different feature selection that is used for the podium prediction.

- Scenario A: Pre-Race scenario will generate the podium probabilities using only the data that are available before the start of the race.
- Scenario B: Within-Race is by extending the features also including information available internally or within-race such as lap time patterns and pit stop behaviours.

Both these scenarios form two final outputs of the implementation, which is a pre-race model that can be used before the race starts but features required should be known. Since it uses pre -race data alone its deployable. And within-race model that is diagnostic that also uses the data available before race and addition to that the extra data which makes it better so it shows upper bound on how the predictions can be better if the race events are allowed to watch.

5.1 Models Implemented

For each scenario, two end-to-end classifier pipelines were selected and finalised according to the nature of dataset. The F1 dataset is having tabular values, mixed numeric and categorical and is imbalanced. The below models both work well under these conditions and output probabilities for P1/P2/P3, which is needed for the ranking of finishers.

- Random Forest (RF): It builds many decision trees on various samples and then takes average of the votes. This will make it stable, robust to the outliers and also good at learning complex patterns without the need of heavy tuning. It is dependable and acts as baseline for all the tabular sports data and gives clear feature importance for interpretation.
- LightGBM (LGBM): This model adds all small trees one after the other, every time fixing the mistake of the previous one. This boosting process will give sharp probability ranking and also more accurate on tabular problems, specially when there are one-hot features. It is fast, memory efficient, performs better when the signals are subtle like interacting with race characteristics and handles class weighting.

These two models support class imbalance, which is crucial because podiums are very rare. They handle non-linear interactions that matter the most in racing like grid, weather, circuit and form. They produce well calibrated probability results which is needed for podium tables. Together they provide as a strong interpretable pair of baseline models with the transparency and for peak the performance and even better probability rankings. Even though both Pre-rcae and Within-race scenarios were compared, the winner for interpretation and calibration was selected as Pre-race scenario since these features are usable and deployable. The winning model is chosen using the validation rule as Macro f1 as primary and Log loss as secondary criterion. The best scenario model pair was found and evaluated on test (2023–2024). Macro-AUC (OVR) and Brier for probability quality is also noted.

5.2 Explainability with SHAP

The predictions of the model are explained using SHAP. SHAP represents the effect of every feature on the probability of the podium of a specific driver. This was calculated with respect to the best model that was selected namely LightGBM in this instance, background samples obtained during validation was used and specific plots were aimed at the podium class. This

created clear plots and tables that bring out the key drivers of the decisions of the model. In the Pre-Race stage, SHAP has generally indicated starting grid, recent driver/constructor form, and circuit/season context to be the most significant factors that increased the chances of podium. SHAP in the Within-Race environment was showing more weightage towards average lap time, lap-time consistency and pit-stop behaviour. These outputs allow readers to see why the model puts greater or lesser podium probability, in simple racing terms.

5.3 Probability Calibration

After choosing the winner model, isotonic calibration was applied in a one vs rest way, fitted on 2020-2022 seasons and then applied to 2023-2024 test. Calibration is independent of which class is selected but modifies the confidence fast that the predictions around 0.70 actually take place around 70 percent of the time. The outcome is a list of more reliable drivers ranking probabilities and risk communication. It is reported that the improvement is provided based on such probability centered metrics as lower log-loss or for the improved Brier score and gives calibrated podium-probability tables of the final seasons as the implementation outputs.

6 Evaluation

To provide a brief glance on the summary, the below Table 6.1 shows that the multi-class podium prediction performance for both scenario that is Pre-Race vs Within-Race (uses post-race aggregates, not a deployable real time forecast), using both the models - Random Forest vs LightGBM, and both splits that are the Validation 2020–2022; Test 2023–2024. This is before the calibration. Lower is the better for Log-loss and Brier, higher is the better for Macro-F1, Accuracy, and Macro-AUC.

Table 2: Performances by model, scenario and the split for podium prediction

Table — Overall performance for multi-class podium prediction

	Scenario	Split	Model	Macro-F1	Accuracy	Log-loss	Brier	Macro-AUC
0	Pre-Race	Test	LightGBM	0.432	0.797	0.536	0.067	0.749
1	Pre-Race	Test	Random Forest	0.428	0.803	0.610	0.075	0.777
2	Pre-Race	Validation	LightGBM	0.445	0.798	0.549	0.068	0.745
3	Pre-Race	Validation	Random Forest	0.413	0.802	0.632	0.078	0.750
4	Within-Race	Test	LightGBM	0.438	0.831	0.485	0.059	0.752
5	Within-Race	Test	Random Forest	0.465	0.849	0.526	0.064	0.766
6	Within-Race	Validation	LightGBM	0.436	0.837	0.475	0.056	0.730
7	Within-Race	Validation	Random Forest	0.420	0.861	0.517	0.062	0.746

The table results indicate that the overall generalisation is consistent across the seasons such as validation test. LightGBM (Pre-Race) is able to reach a Macro-F1= 0.445 (val) / 0.432 (test) with a Log-loss=0.549 / 0.536 and Brier=0.068 / 0.067, which suggests the existence of usable pre-race signal. In the Within-Race case, the quality of probability is better: the Log-loss of LightGBM is 0.475 (val) / 0.485 (test) and Brier is 0.056 / 0.059 (Accuracy 0.837 / 0.831). Whereas Random Forest achieves a slightly higher Macro-F1 on the Within-Race test (0.465

vs 0.438), LightGBM has better calibrated probabilities (lower Log-loss/Brier) in both cases. Since probability calibration quality in making a decision is a priority in this thesis work, LightGBM is chosen as the main model to be used in the future in the probability tables and SHAP analysis. Macro-F1, Log Loss, Brier Score, Accuracy and Macro-AUC are reported. Multi class AUC is computed one vs rest per class and consistent. For deployable case the winner model was selected that is Pre-Race scenario though Within-race Light GBM achieves lower Log loss / Brier, it depends on post-race features and is not deployable.

6.1 Case Study 1: Pre-Race Scenario Performance

The confusion matrix shows that the model classify non-podium outcomes reliably but face difficulty distinguishing between P1, P2 and P3 as shown in figure 10, an expected challenge in motorsport where podium finishes are determined by fine margins and stochastic events.

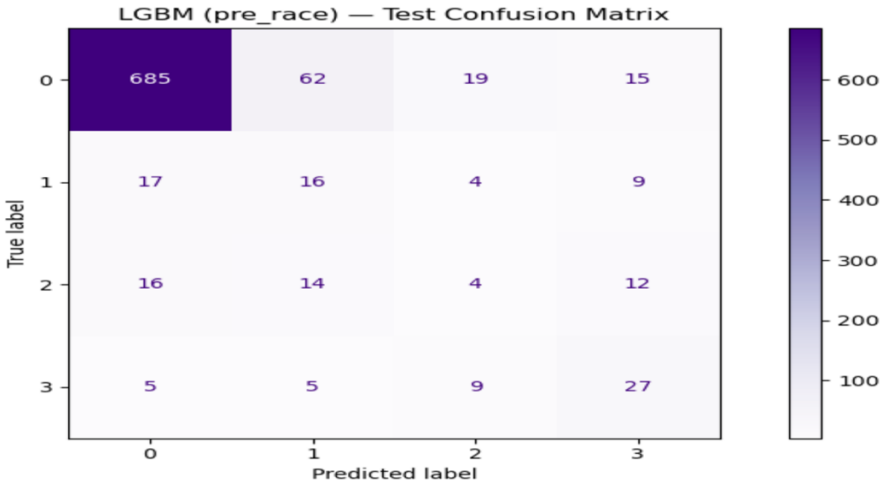

Figure 10: Pre-Race Confusion Matrix(LGBM)

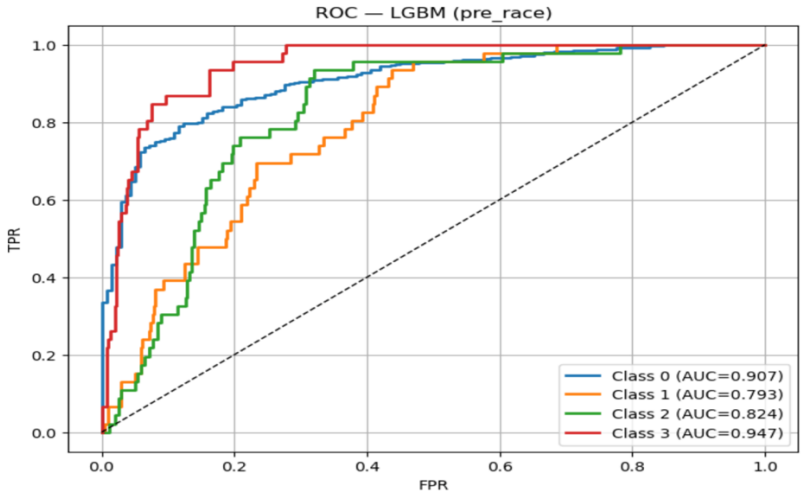

Figure 11: ROC Curve for Pre-Race Scenario

ROC curves in figure 11 explains that for each podium class, however, demonstrate AUC values meaningfully above 0.5, which indicates that the model extracts signal rather than guessing. In this case study it is shown that podium predictions are feasible with pre-race information, but with expected limitations that apply. There is a meaningful predictive signal,

but uncertainty remains high due to in-race factors that the model cannot know before the race starts.

Top-10 drivers (LGBM, pre_race)

	driverId	surname	forename	podium_prob	model
7	830	Verstappen	Max	0.886704	LGBM
8	832	Sainz	Carlos	0.807008	LGBM
12	844	Leclerc	Charles	0.787079	LGBM
13	846	Norris	Lando	0.772083	LGBM
14	847	Russell	George	0.585486	LGBM
18	857	Piastri	Oscar	0.344367	LGBM
3	815	Pérez	Sergio	0.317371	LGBM
0	1	Hamilton	Lewis	0.181677	LGBM
10	840	Stroll	Lance	0.074574	LGBM
11	842	Gasly	Pierre	0.073880	LGBM

Figure 12: Top 10 Predicted Podium Probabilities

6.2 Case Study 2: Within-Race Scenario Performance

This case study is helpful in directly showing that the predictability improves dramatically when within-race information is included. This contrasts with the pre-race scenario and illustrates that much of the outcome variance in Formula-1 emerges from race-time factors such as strategy, execution and pace stability.

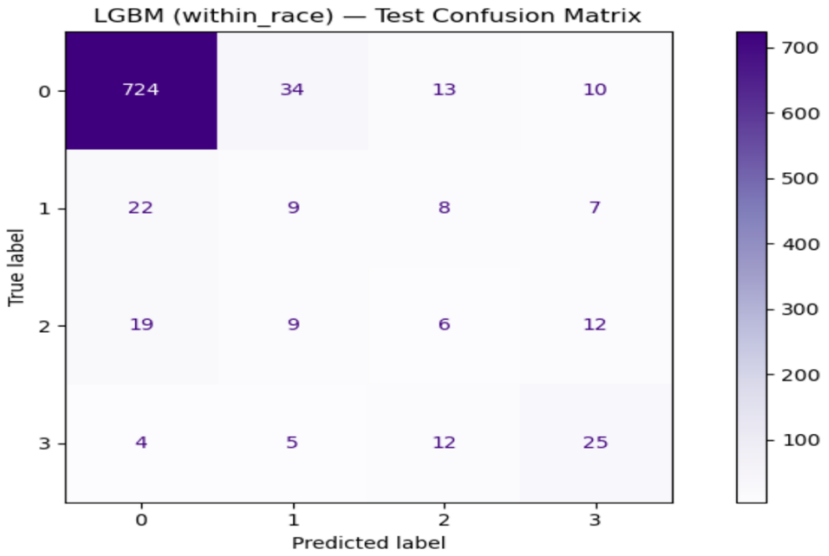

Figure 13: Within-Race Confusion Matrix(LGBM)

ROC curves in this scenario as shown in the figure 14, has consistently higher AUCs across all podium classes compared to the Pre-Race Scenario, supporting the conclusion that in-race factors significantly enhance the predictability.

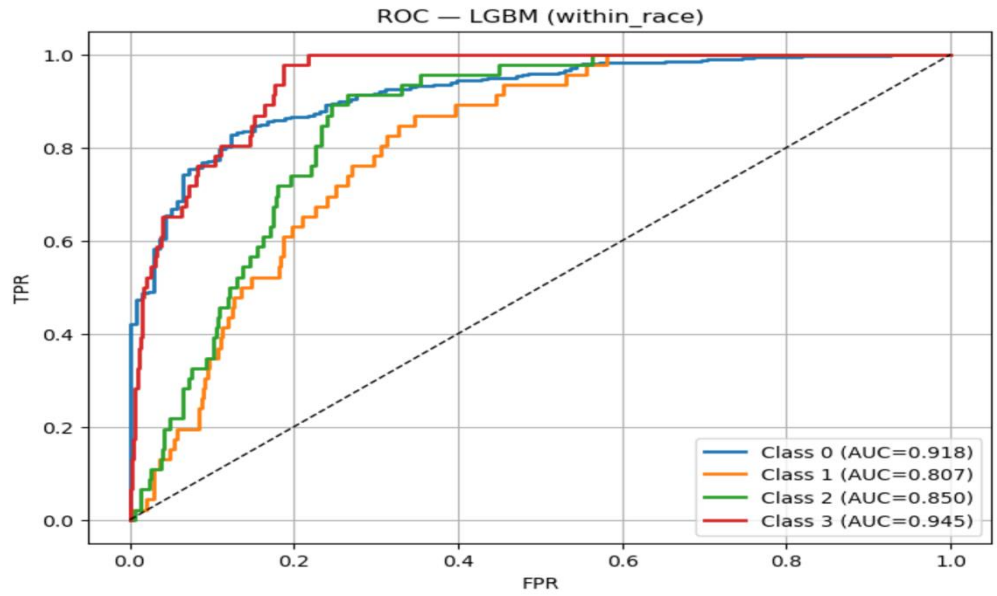

Figure 14: ROC Curve for Within-Race Scenario

Across both scenarios, LightGBM provides consistently better probability quality than Random Forest, achieving lower log-loss and Brier scores on both validation and test splits. Random Forest occasionally attains a slightly higher Macro-F1 particularly in the Within-Race test set but its probability estimates are less well calibrated. Because this study prioritises probability calibration quality over raw classification labels, LightGBM is selected as the final model for interpretation, SHAP analysis, and podium-probability reporting.

Top-10 drivers (LGBM, within_race)

	driverId	surname	forename	podium_prob	model
7	830	Verstappen	Max	0.731103	LGBM
13	846	Norris	Lando	0.709063	LGBM
8	832	Sainz	Carlos	0.685912	LGBM
12	844	Leclerc	Charles	0.643343	LGBM
14	847	Russell	George	0.454147	LGBM
18	857	Piastri	Oscar	0.355087	LGBM
3	815	Pérez	Sergio	0.239589	LGBM
0	1	Hamilton	Lewis	0.177827	LGBM
1	4	Alonso	Fernando	0.031859	LGBM
16	852	Tsunoda	Yuki	0.024847	LGBM

Figure 15: Top 10 Predicted Podium Probabilities for LightGBM showing Within-race scenario

6.3 Case Study 3: Feature Important Analysis

Explainability analysis was conducted on the model that was chosen between the two scenarios. The values of SHAP can be interpreted as a clear and understandable picture of the contribution of features to the probability of a podium finish that is predicted. In the Pre-Race scenario, was selected as it is deployable winner. SHAP summary graph validated that grid position poses the most positive effect whereas driver form and constructor momentum are always greater

chances of being on the podium. Such attributions are very similar to the correlations realized during EDA which confirm that the model reasoning works in line with the known knowledge and is not based on artefacts or noise. This helps both literature and intuitive domain knowledge: podium contestants are expected to have a consistent lap time, minimal delays, and be efficient in terms of strategy.

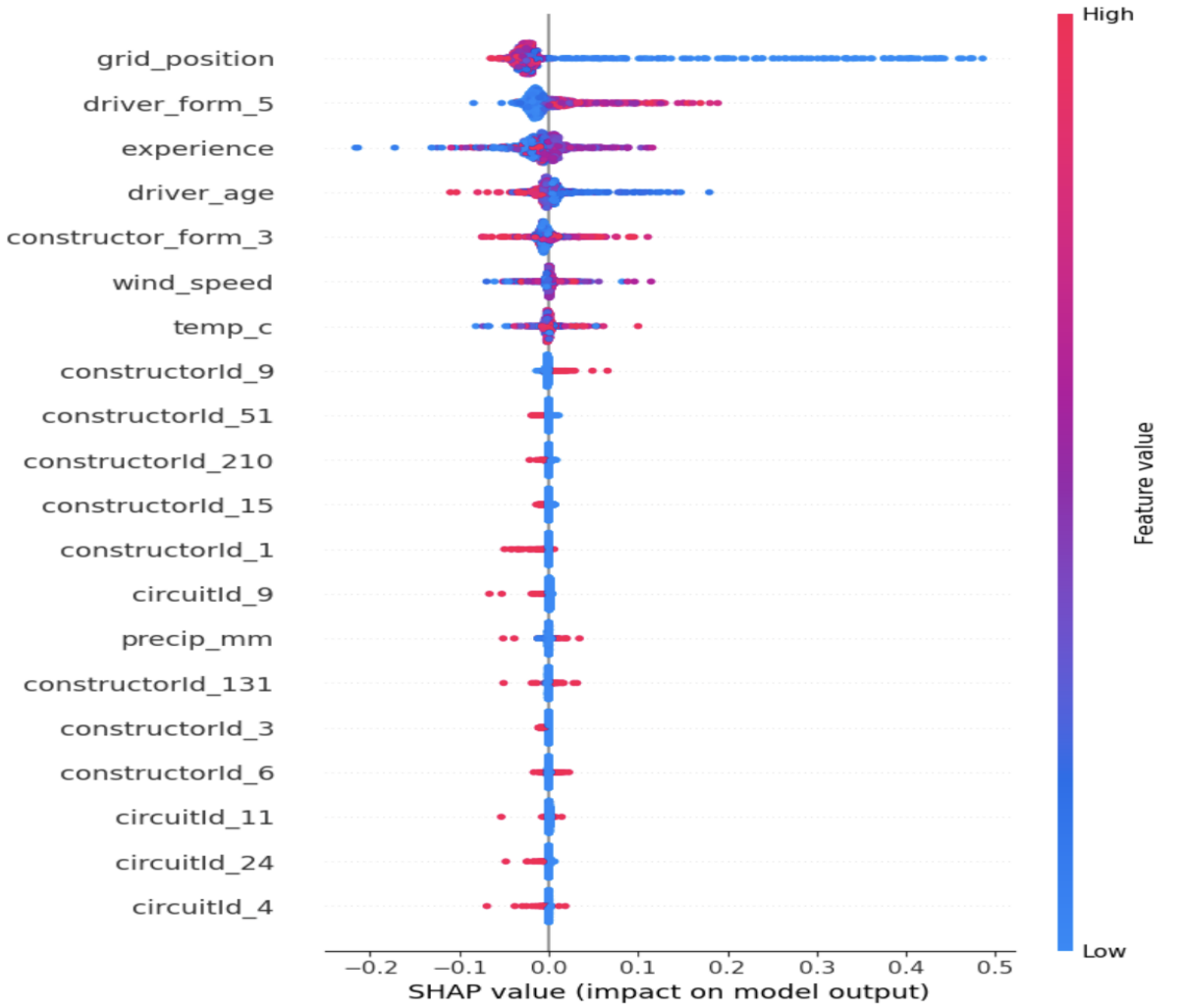

Figure 16: SHAP Feature importance for the Light GBM

Overall, the SHAP importance structure closely mirrors the correlations observed during EDA particularly the strong monotonic link between grid position and finishing order, as well as the negative relationship between lap-time variability and podium likelihood. This alignment confirms that the model is learning meaningful racing dynamics rather than picking up noise or artefacts. Within-Race feature importance were not plotted as it is retrospective scenario and not used in the final visuals.

Probability Calibration Results

To ensure that the predicted probabilities are corresponding with the real-world frequencies, calibration was applied with Isotonic OVR on validation seasons and not on test to preserve the chronological split that was done. Calibration was applied to Pre-race LightGBM, which is further selected as deployable model. It improved for the scenario, Pre-Race LightGBM,

model's probability quality and aligned much better with the observed outcomes, as the log-loss improved from 0.5361 to 0.5105**,** Brier from 0.0675 to 0.0507, and the ECE on the test set dropped from 0.0975 to 0.0243. Calibration does not change the Macro-F1, but it definitely makes the probability estimates more trustworthy for interpretation and downstream usage.

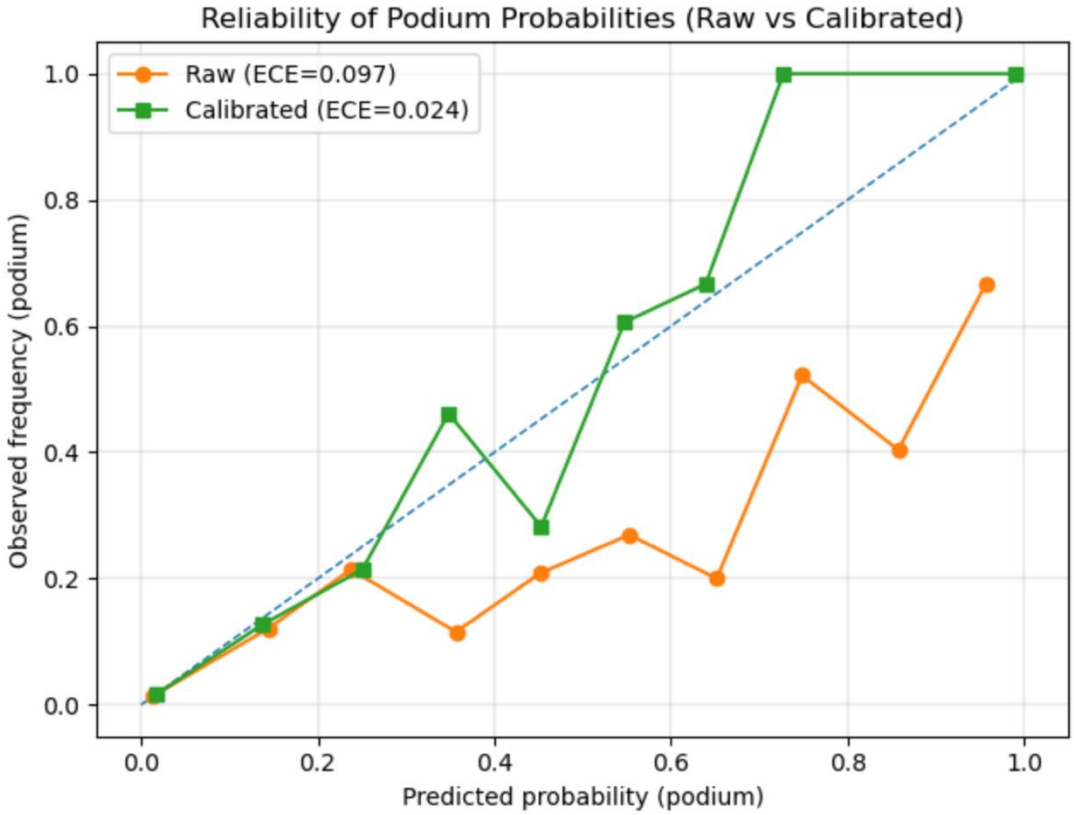

Figure 17: Reliability of Podium Probabilities

Calibrated predictions track the 45° line more closely and showing that ECE improves from 0.0975 to 0.0243, lower is the better.

6.4 Discussion

All three case studies are united by the fact that the results of a podium in Formula 1 can be predicted to some degree prior to the start of a race and that such predictability significantly grows when the within race information appears. The Pre-Race scenario demonstrates that such features as grid position, recent driver and constructor form and circuit context already hold informative signal. Nonetheless, Within-Race scenario is more inclusive of extra variance explained by facts of real race dynamics, including consistency in paces, pit-stop performance, and cumulative race-time behaviour, which greatly increases the probability accuracy and calibration quality. The within-race scenario uses post-race aggregates, so it is not deployable forecast. The increase in these two situations demonstrates the essential ambiguity created by live race events: although pre-race factors create some baseline expectation, in-race strategic and operational variables account for much of the rest of the variation in the probability of being on the podium. SHAP analysis proves these findings by indicating that the model is based on domain-consistent attributes. The grid position and rolling form are the most dominant in

pre-race prediction and the mean lap time, lap-time stability, and pit-stop duration are the most dominant in within-race prediction. SHAP values are also consistent with the trends in the EDA, which proves that the model is also learning the true racing dynamics in lieu of spurious correlations. In general, the difference between scenarios measures the extent to which forecasts can only be realized after the race has taken place, depicting the realistic constraints of pre-race forecasting.

7 Conclusion and Future Work

7.1 Conclusion

This study shows that Formula 1 podium finishes can be predicted in a useful way before the race, and that predictions improve further when in-race signals are added. A clean driver-per-race dataset was built from multiple F1 tables. Two feature sets were engineered: Pre-Race having features like grid position, recent form from past races only, circuit/season context, and weather and Within-Race scenario having features like lap-time and pit-stop summaries. Two models were used in this study Random Forest and LightGBM were trained in reproducible pipelines with chronological splits using train ≤2019, validate 2020–2022 and the test ≥2023. Form features were computed with a one-race shift so that only the past information was used, further preventing from any leakage that can occur.

Calibration was done with the assistance of the isotonic calibration to predict probabilities that would enhance the quality of the probabilities. It reports smaller log-loss as well as Brier score and more reliable curves. What drives the predictions were explained by SHAP grid position and recent form are strongest before the race and then lap-time consistency and pit strategy come in clearly with that of the second scenario within-race. Findings indicate that the Pre-Race scenario already offers realistic podium probabilities that can be deployed where there is access to broadcast insights or strategy support that can derive information out of this is a more valid diagnostic upper bound due to the fact that it utilizes within-race aggregates which offers more weightage to the prediction. LightGBM produces better quality probability estimates across seasons, as compared to Random Forest. On the whole, the project provides a leakage-safe, time-conscious, and reproducible pipeline responding to the research question: pre-race podium prediction is practical and helpful and in-race information also enhances explanatory power and can be implemented with the information being available so far during the race. Since Within-Race scenario features require post-race features they are not deployable. Hence all calibrated probability results and SHAP explanations in this study are based on the Pre-Race LightGBM the best performing deployable model. Although concrete metric results will depend on the model and season, the trend remains the same: LightGBM usually outperforms Random Forest in the quality of the probability predictions, and Within-Race scenario outperforms Pre-Race by a reasonable margin, which attests to the usefulness of the telemetry. The within-race scenario is retrospective and serves as a diagnostic upper bound rather than a deployable real-time model.

7.2 Future Work

The future work directions can be is a way such that the in-race aggregates should be transformed into lap-by-lap cumulative features such that the Within-Race scenario is modelled as a live forecasting say that the measure performance at laps 10,30,50 and not a post-race summary. To include probability calibration like isotonic calibration on validation seasons and to display calibration curves to enhance decision-making in the face of uncertainty. Also,

addition of feature engineering with regulation and tyre-era indicators, safety car or even virtual safety car signs where feasible and enhanced weather dynamics. To investigate prioritization of goals and pairwise loss functions such as LightGBM ranker specific to podium ranking and can also think about conformal prediction to add accurate uncertainty constraints to podium calls. It also has a commercial potential. Pre-Race model is capable of driving broadcast graphics, fan engagement application, and odds-setting systems with clear probabilities of the driver on the podium, at the starting line. The season-level breakdowns can be used by teams and other interested media partners to make comparisons on consistency across circuits or conditions. Having the live cumulative features available, the Within-Race model can be an in-race decision aid, updating podium and top-six or top-10 likelihoods during the race, marking risk windows in which either undercut or overcut strategies can be used with impurity, and marking the drivers whose pace is becoming less consistent. Due to the modularity of the pipeline and leakage-safety, the pipeline can be embedded into a lightweight service that consumes results that are qualifying, weather forecasts, and incremental lap data to generate real-time explainable predictions.

References

Belkovics, L., Krisztina Andr and István Takács (2024). Formula 1 - Information Solutions on and off the Racetrack. *Formula 1 - Information Solutions on and off the Racetrack*, pp.1–6. doi:https://doi.org/10.1109/saci60582.2024.10619918.

Bell, A., Smith, J., Sabel, C.E. and Jones , K. (2016). This is an author produced version of Formula for success: Multilevel modelling of. *Journal of Quantitaive Analysis in Sports*, [online] 12(2). doi:https://doi.org/10.1515/jqas-2015-0050.

El Haber, E., Sawaya, E., Attieh, M., Tannous, A., Ghazaly, W. and Owayjan, M. (2025). Formula 1 Race Winner Prediction Using Random Forest and SHAP Analysis. *2025 International Conference on Control, Automation, and Instrumentation (IC2AI)*, [online] pp.1270–1274. doi:https://doi.org/10.1109/ic2ai62984.2025.10932140.

Erik–Jan van Kesteren and Bergkamp, T. (2023). Bayesian analysis of Formula One race results: disentangling driver skill and constructor advantage. *Journal of Quantitative Analysis in Sports*, 0(0). doi:https://doi.org/10.1515/jqas-2022-0021.

Garcia-del-Barrio, P. and Reade, J.J. (2021). Does certainty on the winner diminish the interest in sport competitions? The case of formula one. *Empirical Economics*, 63, pp.1059–1079. doi:https://doi.org/10.1007/s00181-021-02147-8.

Gneiting, T. and Raftery, A.E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), pp.359–378. doi:https://doi.org/10.1198/016214506000001437.

Heilmeier, A., Thomaser, A., Graf, M. and Betz, J. (2020). Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport. *Applied Sciences*, 10(21), p.7805. doi:https://doi.org/10.3390/app10217805.

Kavitha, R.K. and Kanishk, S. (2025). Formula One: Racing Intelligence in the Age of AI. *2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, [online] pp.1–7. doi:https://doi.org/10.1109/icaeca63854.2025.11012647.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez Orallo, J., Kull, M., Lachiche, N., Ramirez Quintana, M.J. and Flach, P.A. (2020). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, [online] 33(8), p.1. Available at: https://research-information.bris.ac.uk/ws/portalfiles/portal/220614618/TKDE_Data_Science_Trajectories_PF.pdf.

Patil, A., Jain, N., Rahul Agrahari, Murhaf Hossari, Orlandi, F. and Dev, S. (2023). A Data-Driven Analysis of Formula 1 Car Races Outcome. *Artificial Intelligence and Cognitive Science (AICS)*, 1662, pp.134–146. doi:https://doi.org/10.1007/978-3-031-26438-2_11.

Priya Shelke, Riddhi Mirajkar, Pande, A., Kale, S. and Yash Paralikar (2023). F1 Race Winner Predictor. *International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. doi:https://doi.org/10.1109/iccubea58933.2023.10392224.

Vidhya Shree M, L, T.S., N, A. and Jayati Bhadra (2025). Predictive Modeling of Formula 1 Constructor Performance: Discovering the Top Team. *International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, [online] pp.2131–2138. doi:https://doi.org/10.1109/idciot64235.2025.10915082.

WallpaperSafari. (2026). *F1 Wallpapers High Resolution on WallpaperSafari*. [online] Available at: https://wallpapersafari.com/f1-wallpapers-high-resolution/#google_vignette [Accessed 27 Jan. 2026].

Walsh, C. and Joshi, A. (2024). Machine learning for sports betting: Should model selection be based on accuracy or calibration? *Machine Learning with Applications*, [online] 16, p.100539. doi:https://doi.org/10.1016/j.mlwa.2024.100539.

wsiqz (2024). *formula 1 weather info (1950-2024)*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/mariyakostyrya/formula-1-weather-info-1950-2024 [Accessed 2025].

www.kaggle.com. (n.d.). *Formula 1 World Championship (1950 - 2023)*. [online] Available at: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/data.