Predicting Formula 1 Podium Finishes Using Ensemble Machine Learning Models

Anoop Lashiyal
Department of Computer Science
Georgia State University, Atlanta, USA

***Abstract*—Predicting podium finishes in Formula 1 (F1) is difficult due to strong class imbalance (only three podium positions per race), non-stationary team performance, and the complex interaction of qualifying, driver form, and track context. This paper presents a comparative study of two tree-based ensemble methods—XGBoost and Random Forest—for predicting whether a driver will finish on the podium in a given race. We evaluate the models on unseen season data to test generalization, reporting ROC-AUC, PR-AUC, and lift-based ranking performance. Both models substantially outperform a random baseline, with XGBoost showing a marginal but consistent advantage in ranking-sensitive metrics. The results support gradient-boosted decision trees as a strong approach for imbalanced, real-world sports analytics.**

***Index Terms*—Formula 1, podium prediction, imbalanced classification, XGBoost, Random Forest, ROC-AUC, PR-AUC, lift chart**

I. Introduction

Formula 1 (F1) race outcome prediction represents a challenging real-world machine learning problem due to its dynamic environment, evolving regulations, and strong class imbalance. Only three drivers finish on the podium in each race, while the remaining drivers constitute the majority class, making naïve prediction strategies ineffective. Furthermore, frequent changes in car regulations, driver line-ups, and team strategies introduce non-stationarity into the data distribution.

This project focuses on predicting whether a driver will achieve a podium finish in a given race using historical race data and ensemble machine learning models. The primary objective is not only to achieve high predictive accuracy, but also to produce well-calibrated probability estimates that allow meaningful ranking of drivers within each race.

A major motivation for this work stems from the practical applications of race prediction models, including motorsport analytics platforms, strategic simulations, and betting-oriented decision support systems. In such applications, ranking quality is often more important than raw classification accuracy, as financial or strategic value is derived from identifying high-probability candidates within a small subset of competitors.

Contributions:

- We formulate podium-finish prediction as an imbalanced binary classification problem on a driver–race dataset.
- We evaluate XGBoost and RF on unseen season data and compare discriminative and ranking performance.
- We include lift analysis to quantify practical gains when selecting top-ranked drivers.

II. Related Work

Sports outcome prediction has been studied using statistical modeling, classical machine learning, and deep learning. In motorsports, earlier approaches often relied on regression or handcrafted heuristics using qualifying and historical pace. More recent methods apply ensemble models to leverage feature interactions and nonlinearity. Gradient boosting has shown strong performance across structured prediction tasks, especially for rare-event classification and ranking.

III. Dataset and Feature Engineering

A. Data Source

The dataset used in this study was obtained from the Kaggle Formula 1 World Championship repository, which spans race seasons from 1950 to 2024. This dataset contains multiple relational tables describing drivers, constructors, circuits, qualifying results, race outcomes, weather conditions, sprint race data, and additional metadata.

Although comprehensive, the raw dataset contains a large number of features and historical records that are not directly relevant to modern race prediction. As a result, extensive data cleaning and preprocessing were required prior to model development.

B. Temporal Filtering and Regulation Consistency

One of the key challenges identified during preliminary exploration was the long temporal span of the dataset. Formula 1 has undergone multiple major regulation changes since 1950, resulting in substantial differences in car performance, aerodynamics, engine technology, and race dynamics. Models trained on such heterogeneous data risk learning spurious patterns that do not generalize to the modern era.

To address this issue, all race data prior to the 2014 season was removed. The year 2014 marks the introduction of the V6 turbo-hybrid engine regulations by the FIA (Fédération Internationale de l'Automobile), which established a more uniform technological baseline across teams. Restricting the dataset to post-2014 seasons ensures greater consistency in vehicle performance characteristics and removes a large number of retired drivers from the dataset.

This filtering step improves model validity by reducing temporal bias and ensuring that learned patterns reflect contemporary Formula 1 racing conditions.

C. Handling Rookie Drivers and Missing Historical Data

A critical challenge in race prediction arises when evaluating models on future or unseen seasons that include rookie drivers with limited or no historical race data. In the 2025 Formula 1 season, several drivers—including Andrea Kimi Antonelli, Oliver Bearman, Gabriel Bortoleto, Isack Hadjar, and Liam Lawson—entered the grid without extensive prior Formula 1 race histories.

The absence of historical data for such drivers creates feature sparsity that can degrade model performance if not handled carefully. To mitigate this issue, derived features were constructed by mapping rookies to statistically similar drivers based on junior series performance, team context, and early-season indicators. These derived features allow rookies to be meaningfully represented within the feature space, enabling the model to generate probability estimates rather than defaulting to majority-class predictions.

In real-world applications such as betting or ranking-based decision systems, accurate modeling of low-probability and high-variance outcomes is particularly valuable. Correctly identifying mid-field or unexpected high-performing drivers can yield significantly greater returns than consistently predicting dominant winners.

D. Supplementary Data Acquisition

To construct a complete dataset for recent and future seasons, multiple external data sources were evaluated. Initial attempts relied on telemetry-focused archives, such as the TracingInsights 2025 repository; however, these sources primarily contained lap-level telemetry data and lacked the structured race-level features required for this study.

Subsequently, the *formula1-datasets* repository was identified as a suitable source, providing race-level information consistent with the historical Kaggle format. This data was processed and converted into a unified CSV structure to ensure compatibility with the existing pipeline. The final dataset supports seamless training, validation, and evaluation across seasons.

E. Dataset Construction

Each observation corresponds to a *(driver, race)* pair. The label $y \in \{0, 1\}$ indicates whether the driver finished on the podium ($y = 1$) or not ($y = 0$). To evaluate generalization, we use a temporally separated test split consisting of an unseen season that includes rookies and new driver–team pairings.

F. Features

The feature set includes (representative examples):

- **Qualifying / grid**: qualifying position and/or grid position.
- **Driver form**: rolling averages of recent finishes, points, and consistency.
- **Team strength**: season-to-date team points/pace indicators.
- **Track context**: circuit-specific identifiers and/or engineered track difficulty proxies.

Missing values are handled using model-native strategies (e.g., split directions for tree models) or standard imputation where required.

G. Class Imbalance

Because podium events are rare, accuracy may be inflated by predicting the majority class. We therefore prioritize ROC-AUC, PR-AUC, and lift curves to evaluate ranking quality and usefulness for top-k selection.

IV. Methods

A. Problem Formulation

Given a feature vector x_i for a driver–race instance, the model outputs a probability $\hat{p}_i = P(y_i = 1 \mid x_i)$. The prediction can be used for:

- **Binary classification** via thresholding $\hat{p}_i$.
- **Ranking** drivers within a race by $\hat{p}_i$.

B. Random Forest

Random Forest is an ensemble of decision trees trained on bootstrapped samples with feature subsampling. RF is robust, easy to train, and provides strong baselines for tabular data.

C. XGBoost

XGBoost is a gradient-boosted tree model that sequentially builds trees to correct prior errors while optimizing a regularized objective. This often yields strong ranking performance and improved handling of rare-event signals.

V. Evaluation

We report:

- **Accuracy**: overall fraction correct (not sufficient alone under imbalance).
- **ROC-AUC**: threshold-independent discrimination ability.
- **PR-AUC**: precision–recall summary, informative for rare positive labels.
- **Lift**: gain over random selection when ranking by predicted probability.

A. Confusion-Matrix Notation

Let $y \in \{0, 1\}$ denote the ground-truth label where 1 indicates a podium finish. Given predicted label $\hat{y}$, we define:

$$TP = \sum_i \nVdash(y_i = 1 \wedge \hat{y}_i = 1), \tag{1}$$

$$FP = \sum_i \nVdash(y_i = 0 \wedge \hat{y}_i = 1), \tag{2}$$

$$TN = \sum_i \nVdash(y_i = 0 \wedge \hat{y}_i = 0), \tag{3}$$

$$FN = \sum_i \nVdash(y_i = 1 \wedge \hat{y}_i = 0), \tag{4}$$

where $\nVdash(\cdot)$ is the indicator function.

TABLE I
PERFORMANCE COMPARISON OF XGBOOST AND RANDOM FOREST MODELS

Metric	XGBoost	Random Forest
Accuracy	0.8837	0.8860
ROC-AUC	0.9406	0.9379
PR-AUC	0.7895	0.7794
Top-3 Hit Rate	0.4011	0.4011
Precision (Podium)	0.71	0.70
Recall (Podium)	0.64	0.70
F1-score (Podium)	0.68	0.70

B. ROC Curve

Given a probability score $s_i \in [0, 1]$ and threshold τ, the predicted label is $\hat{y}_i(\tau) = \nVdash(s_i \geq \tau)$. The ROC curve plots True Positive Rate (TPR) against False Positive Rate (FPR) across all thresholds:

$$\mathrm{TPR}(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}, \tag{5}$$

$$\mathrm{FPR}(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)}. \tag{6}$$

C. Precision–Recall (PR) Curve

The PR curve plots Precision vs. Recall across thresholds τ:

$$\mathrm{Precision}(\tau) = \frac{TP(\tau)}{TP(\tau) + FP(\tau)}, \tag{7}$$

$$\mathrm{Recall}(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}. \tag{8}$$

D. Lift Curve

Let N be the number of instances, $P = \sum_i y_i$ the total number of positives, and $\pi = P/N$ the base positive rate. Sort instances by descending score s_i and let S_q denote the top fraction $q \in (0, 1]$ of instances. Let $P_q = \sum_{i \in S_q} y_i$ be the number of positives captured in the top q fraction. The lift at fraction q is:

$$\mathrm{Lift}(q) = \frac{P_q/(qN)}{P/N} = \frac{P_q}{qP}. \tag{9}$$

A lift of 1 corresponds to random selection; values greater than 1 indicate improvement.

VI. RESULTS

A. Overall Metrics

Fig. 1 summarizes the primary metrics for XGBoost and RF. Both achieve high ROC-AUC (≈ 0.94) and strong PR-AUC ($\approx$ 0.78–0.79), indicating effective discrimination and useful precision under imbalance.

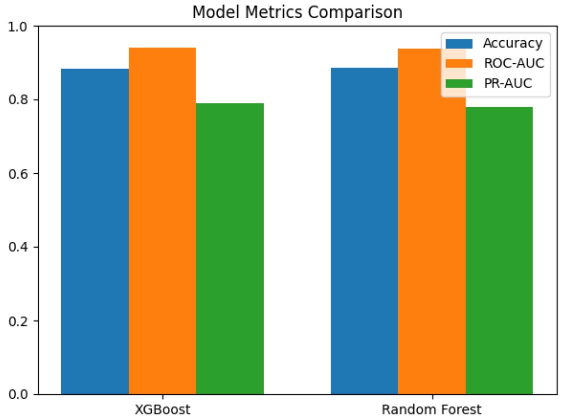

Fig. 1. Model metrics comparison for XGBoost and Random Forest (Accuracy, ROC-AUC, PR-AUC).

TABLE II
CONFUSION MATRIX FOR XGBOOST MODEL

Actual / Predicted	Non-Podium (0)	Podium (1)
Non-Podium (0)	328	21
Podium (1)	29	52

TABLE III
CONFUSION MATRIX FOR RANDOM FOREST MODEL

Actual / Predicted	Non-Podium (0)	Podium (1)
Non-Podium (0)	324	25
Podium (1)	24	57

B. Confusion Matrix Analysis

To provide a detailed view of classification behavior, we report confusion matrices for both models in Tables II and III. These matrices highlight the trade-off between false positives and false negatives when predicting podium finishes.

The XGBoost model exhibits fewer false positives, while the Random Forest model achieves a higher number of true positives for the podium class, resulting in slightly higher recall. This trade-off is consistent with the precision–recall behavior observed in Fig. 3.

C. Lift Analysis

Lift curves (Fig. 2) quantify practical ranking utility. In the top-ranked fraction of drivers, both models provide lift significantly above the random baseline (lift $=$ 1). This suggests strong value when selecting a small subset of likely podium candidates.

D. Precision–Recall Curve

The PR curves (Fig. 3) show performance under imbalance across recall levels. XGBoost achieves a slightly higher average precision (AP) than RF, indicating marginally better precision for comparable recall.

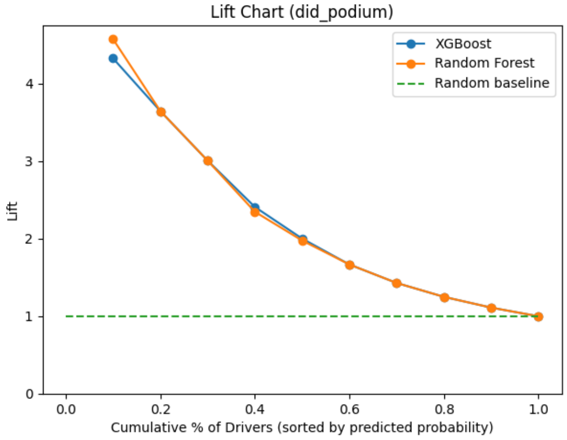

Fig. 2. Lift chart for podium prediction. Lift >1 indicates improvement over random ranking.

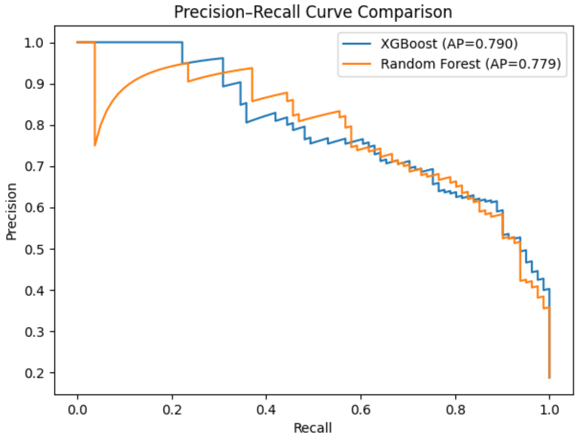

Fig. 3. Precision–Recall curves for XGBoost and Random Forest with average precision (AP).

E. ROC Curve

ROC curves (Fig. 4) demonstrate strong separability for both models. XGBoost achieves a slightly higher AUC, consistent with its boost-based optimization.

F. Top-3 Hit Rate Analysis

To evaluate race-level ranking performance, we compute the Top-3 Hit Rate, defined as the fraction of races in which at least one of the actual podium finishers appears among the model's top three ranked drivers. This evaluation is performed on historical seasons only, excluding the 2025 dataset, to avoid bias from derived rookie features.

G. Top-3 Hit Rate (Per Race)

For each race $r \in \{1, \dots, R\}$, let $\mathcal{D}_r$ be the set of drivers in that race. Let $\text{Top3Pred}(r)$ be the set of 3 drivers with the highest predicted probabilities, and $\text{Podium}(r)$ be the set of true podium finishers. We define an indicator of success for race r as:

$$h_r = \nVdash \left(|\text{Top3Pred}(r) \cap \text{Podium}(r)| \geq 1 \right). \tag{10}$$

The Top-3 Hit Rate is then:

$$\text{Top3HitRate} = \frac{1}{R} \sum_{r=1}^{R} h_r. \tag{11}$$

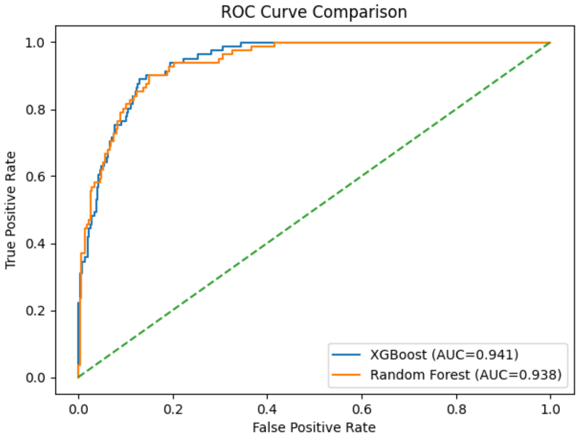

Fig. 4. ROC curves for XGBoost and Random Forest with AUC values.

TABLE IV
TOP-3 HIT RATE PER RACE (EVALUATION EXCLUDING 2025 SEASON)

Model	**Top-3 Hit Rate**
XGBoost	0.4011
Random Forest	0.4011

As shown in Table IV, both XGBoost and Random Forest achieve a Top-3 Hit Rate of 0.4011. This indicates that in approximately 40% of races, the models successfully identify at least one true podium finisher within their top three ranked predictions. The identical performance suggests that both models capture similar high-level ranking signals at the race level, despite differences observed in instance-level metrics such as PR-AUC and ROC-AUC.

H. Hyperparameter Sensitivity Analysis

1) Effect of Ensemble Size: To evaluate the impact of ensemble size on model performance, we trained both XGBoost and Random Forest classifiers using two different numbers of estimators, $n_{\text{estimators}} = 200$ and $10{,}000$. This analysis examines whether increasing ensemble size improves predictive performance for Formula 1 podium classification, particularly under severe class imbalance.

For XGBoost, increasing the number of estimators beyond 200 did not lead to performance improvements. While ROC-AUC remained comparable across both settings, accuracy decreased from 0.8907 to 0.8767 and PR-AUC declined from 0.7795 to 0.7602 when using $10{,}000$ estimators. These results indicate diminishing returns and potential overfitting when excessively large ensembles are employed.

Random Forest exhibited different behavior. With 10,000 estimators, the model achieved its strongest overall performance, obtaining the highest PR-AUC (0.7867) and improved recall for podium finishes. Compared to the smaller ensemble, the larger Random Forest more effectively captured complex nonlinear interactions, leading to improved minority-class ranking.

At $n_{\text{estimators}} = 200$, XGBoost slightly outperformed Random Forest in terms of accuracy and podium-class precision, while both models achieved comparable ROC-AUC and PR-AUC values. These findings highlight XGBoost's ability to achieve competitive performance with substantially fewer estimators, offering improved computational efficiency.

2) Impact of Tree Depth on Model Performance: To analyze the effect of model complexity on podium prediction, we evaluated both XGBoost and Random Forest using two maximum tree depths, max_depth = 3 and max_depth = 6, while keeping all other hyperparameters fixed. This experiment assesses the bias–variance tradeoff and robustness under varying depth constraints.

For XGBoost, increasing tree depth improved ranking performance without affecting overall accuracy. Accuracy remained constant at 0.8837, while ROC-AUC increased from 0.9349 to 0.9406 and PR-AUC improved from 0.7830 to 0.7895 when using max_depth = 6. This suggests that moderately deeper trees allow XGBoost to capture more complex feature interactions relevant to podium outcomes.

Random Forest showed greater sensitivity to tree depth. With max_depth = 3, the model achieved high recall but suffered reduced accuracy due to increased false positives. Increasing depth to 6 improved overall performance, raising accuracy to 0.8791 and PR-AUC to 0.7894 while maintaining strong minority-class recall.

Overall, these findings demonstrate that controlling tree depth is more impactful than excessively increasing ensemble size. Moderate tree depths provide an effective balance between expressiveness and generalization for Formula 1 podium prediction, particularly when computational efficiency is considered.

VII. Discussion

Across metrics, both ensembles perform strongly for podium-finish prediction. The lift analysis is particularly relevant in deployment settings: when the application cares about the top-ranked drivers (e.g., top 10–20%), both models produce large gains over random selection. The marginal advantage of XGBoost in PR-AUC and AUC suggests better ranking stability under imbalance.

A key takeaway is that accuracy alone is insufficient: in imbalanced settings, models can achieve high accuracy while failing to identify podium events. Ranking-based metrics and PR analysis provide a more realistic view of practical utility.

VIII. Limitations and Future Work

This work uses structured historical features and does not incorporate live telemetry, tire degradation, incidents, or weather. Future work can:

- Incorporate context such as weather, safety cars, and tire strategy.
- Extend to **multi-class** prediction (P1, P2, P3) or per-race top-3 ranking.
- Explore calibration (e.g., reliability curves) and race-level constraints (only 3 podiums).

IX. Conclusion

We compared XGBoost and Random Forest for predicting F1 podium finishes on unseen season data. Both models substantially outperform random baselines, with XGBoost providing a small but consistent improvement in ranking-sensitive metrics. These results support gradient-boosted tree ensembles as a strong baseline for imbalanced F1 outcome prediction.

The code and processed datasets used in this study are publicly available on GitHub [5].

References

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.

[3] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. ICML*, 2006, pp. 233–240.

[4] FIA Formula One World Championship, "Historical race data" accessed 2025.

[5] A. Lashiyal, "F1 Podium Prediction Using Machine Learning" GitHub repository, 2025. [Online]. Available: https://github.com/Anoop-cpu/F1project