Predictive Analytics in Formula 1

A Comparative Analysis of Linear, Ensemble, and Deep Learning Architectures for Point Scoring Prediction (2000–2024)

By Amaan Kara

Table of Contents

Introduction *3*

Research Questions 3

Data Acquisition and Scope *4*

Data Cleaning and Scope 4

Feature Engineering 4

Exploratory Data Analysis (EDA) *5*

Univariate analysis 5

Bivariate Analysis 6

Methodology *9*

Results and Discussion *9*

Overall Model Performance *10*

The Linearity of F1 11

Feature Importance 11

Conclusion *12*

Introduction

Formula 1 is the pinnacle of motorsport, a multi-billion-dollar industry defined by the interplay between engineering excellence and human skill. A common adage in the sport is that success is "90% car and 10% driver." However, quantifying this relationship remains a complex statistical challenge. Unlike varied sports such as football or basketball, Formula 1 is highly deterministic but plagued by significant noise factors, including mechanical failures, weather conditions, and frequent regulatory changes.

The primary objective of this project is to apply Machine Learning (ML) techniques to historical race data to separate the signal from the noise. Specifically, this report investigates whether the probability of a driver scoring championship points can be accurately predicted using only pre-race information. By comparing interpretable linear models against complex deep learning architectures, this study seeks to determine the underlying mathematical nature of the sport: is Formula 1 a linear physics problem, or a complex non-linear system?

Research Questions

To guide this investigation, three specific research questions were formulated:

1. Predictive Capability: Can the binary outcome of scoring points be accurately predicted using pre-race performance metrics and contextual features?
2. Methodological Comparison: Does a complex Deep Learning model (significantly outperform a Regularised Linear Model, or is the relationship between qualifying and finishing position largely linear?
3. Feature Value: Do engineered contextual features, specifically Recent Form and Teammate Comparisons , provide measurable predictive value over raw statistics?

Data Acquisition and Scope

The dataset for this project was sourced from Kaggle, which contained a comprehensive historical record of Formula 1 results. The raw data consisted of 14 relational CSV files including race results, lap times, and qualifying data.

Data Cleaning and Scope

To ensure statistical consistency, the analysis was restricted to the Modern Era (2000–2024). Historical data prior to 2000 was discarded to mitigate "regime noise" caused by radically different scoring systems (for example, only the top 6 scoring points versus the top 10 today) and reliability standards.

Data quality checks identified significant missingness in the Qualifying data (qualifying.csv). Drivers who failed to set a time were marked as \N. Rather than discarding these informative records, an imputation strategy was applied: missing qualifying positions were imputed to '25' (the back of the grid), preserving the signal that a driver encountered a mechanical issue or crash.

Feature Engineering

A core hypothesis of this study is that raw data (just using "Points Scored") is insufficient for prediction because it lacks context. A driver finishing 10th in a dominant car is a failure, whereas finishing 10th in the weakest car is a triumph. To address this, three contextual features were engineered:

1. **Driver Form (driver_form_3):** A rolling average of points scored in the previous three races. This feature captures momentum, a critical psychological and technical factor in sports.
2. **Teammate Skill Differential (teammate_quali_diff):** Calculated as the difference between a driver’s qualifying position and their teammate's. This effectively normalizes for car performance, isolating the driver's raw skill contribution.
3. **Team Strength (constructor_season_points):** The cumulative points scored by the constructor up to the current race, serving as a proxy for the mechanical quality of the car.

Exploratory Data Analysis (EDA)

Before modelling, a thorough EDA was conducted to characterize the dataset and justify the selection of specific algorithms.

Univariate analysis

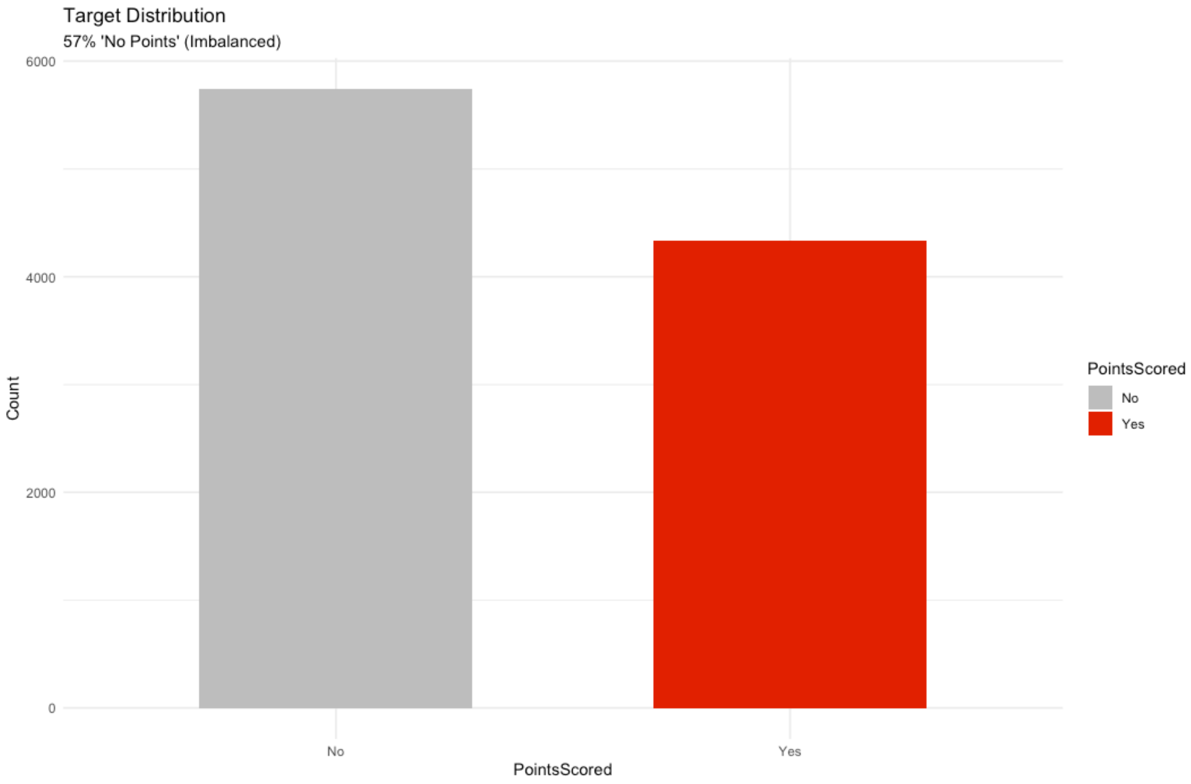

Figure 1: Target variable distribution

As shown in *Figure 1*, the dataset exhibits a moderate class imbalance, with approximately 57% of entries resulting in "No Points." This observation is critical for the evaluation strategy: while this represents a moderate class imbalance, it still necessitates the use of AUC over Accuracy to ensure robust evaluation.

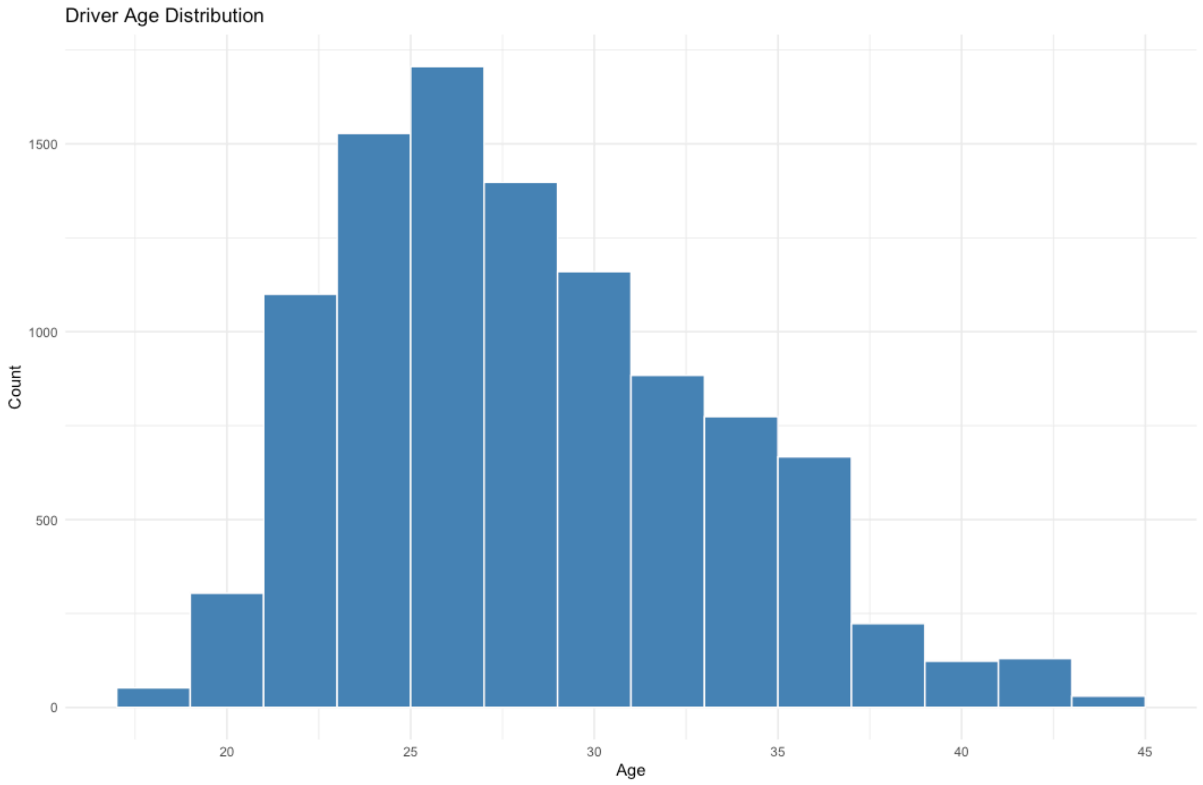

Figure 2: Driver Age Distribution

Figure 2 illustrates the demographic profile of the dataset. The age distribution is right-skewed, with most drivers aged between 22 and 30. This confirms that F1 is a sport dominated by youth, though the long tail of drivers over 35 suggests that experience allows some athletes to compete well beyond their physical prime.

Bivariate Analysis

To select the appropriate machine learning models, we visualised the relationship between key predictors and the target variable.

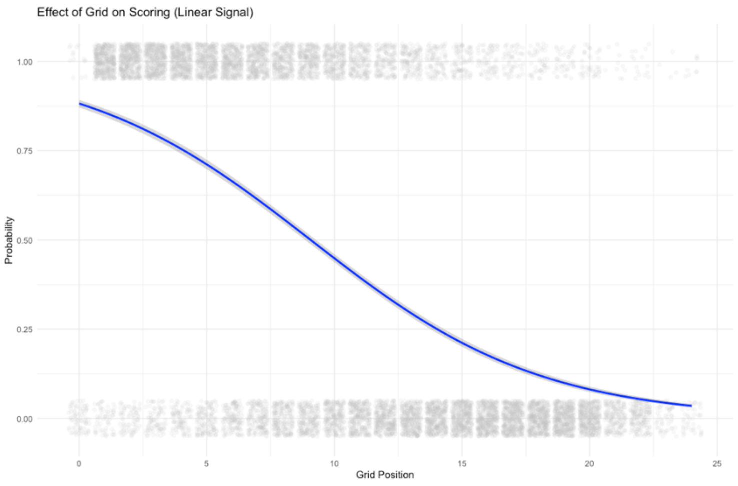

Figure 3a: S-Curve

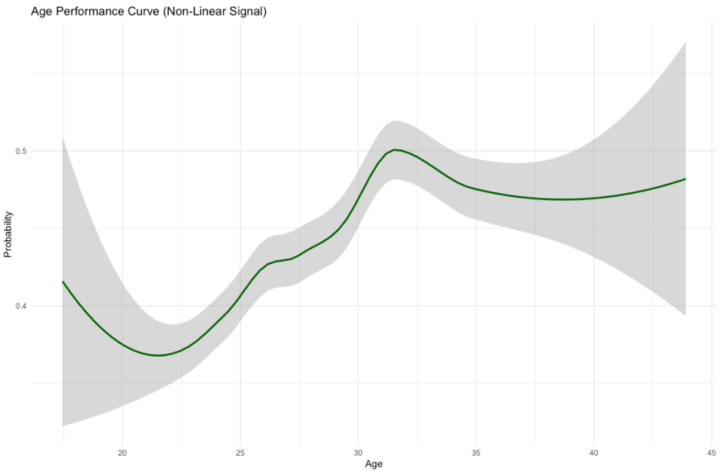

Figure 3b: Non- Linearity Check

Figure 3a displays the relationship between Starting Grid Position and Scoring Probability. The distinct S-Curve demonstrates a strong, monotonic, negative relationship. This linear signal suggests that a **Generalized Linear Model (GLM)**, specifically Logistic Regression, serves as an excellent baseline candidate.

Conversely, *Figure 3b* illustrates the relationship between Driver Age and performance. The curve reveals a non-linear performance peak in the mid-20s, with lower probabilities for rookies (<21) and declining veterans (>35). A standard linear model would fail to capture this inverted U-shape. This empirical evidence justifies the inclusion of **Tree-based ensembles (Random Forest, GBM)** and **Neural Networks**, which can model such non-linear discontinuities.

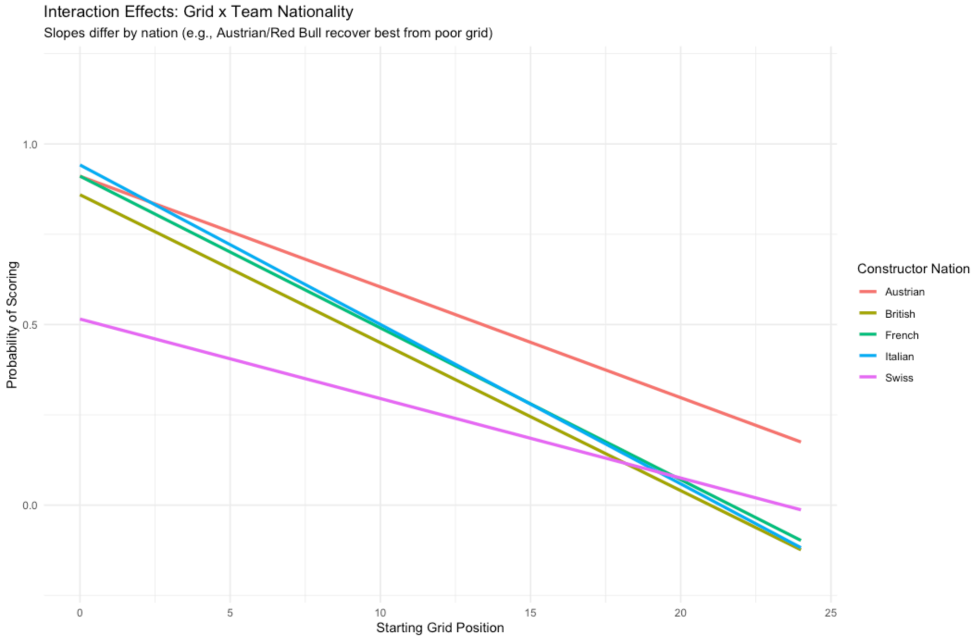

Figure 4: Interaction Effects

The final justification for deploying ensemble and deep learning models rests on the presence of interaction effects. While the Lasso GLM assumes a single average relationship between Grid Position and Probability of Scoring, this assumption fails if that relationship changes based on the team.

The final justification for deploying ensemble and deep learning models rests on the analysis of Interaction Effects. As illustrated in *Figure 4*, the slopes of the regression lines describing the grid penalty differ significantly across constructor nationalities (a strong proxy for team quality). For dominant teams (e.g., Austrian/Red Bull, German/Mercedes), the line is noticeably flatter, signifying that their superior car performance allows them to recover the penalty of starting in a midfield position (e.g., 8th) more effectively. Conversely, midfield teams (e.g., French or Japanese origin) show a steeper slope, meaning their probability of scoring drops rapidly as the grid position worsens. This empirical evidence, which shows that the effect of the primary predictor (grid) is conditional on the categorical variable (const_nat), validates the necessity of using non-linear models capable of learning these complex interactions.

Methodology

The dataset was partitioned into a 70% Training and 30% Independent Test set. Crucially, the split was arranged to ensure the ratio of point-scorers remained consistent across both sets.

Four distinct models were trained to represent the spectrum of the Bias-Variance trade-off:

1. **Lasso GLM (Linear):** A Logistic Regression with L1 regularization. This was chosen to handle multicollinearity (correlation > 0.8) between Grid Position and Qualifying Position.
2. **Random Forest (Bagging):** An ensemble of 100 decorrelated decision trees. A Grid Search was performed to optimize maximum depth (10, 20) to prevent overfitting.
3. **Gradient Boosting Machine (Boosting):** A sequential ensemble (GBM) designed to reduce bias by iteratively correcting the errors of previous trees.
4. **Deep Learning (Neural Network):** A feed-forward Multi-Layer Perceptron (MLP) trained using the H2O framework.
   - **Architecture:** 2 Hidden Layers (200 neurons each).
   - **Activation:** Rectifier with Dropout (to improve generalization).
   - **Class Handling:** balance_classes = TRUE was enabled to statistically oversample the minority class (point scorers).

Results and Discussion

The predictive performance of all four models was assessed on the independent test set (N ~ 3,000). The tight clustering of results around the 0.85 AUC mark is the study's primary scientific finding.

Overall Model Performance

Table 1: Final Performance

Model	**AUC Score**	**Rank**
Deep Learning	0.8576	1
Random Forest	0.8568	2
Lasso GLM	0.8523	3
GBM	0.8516	4

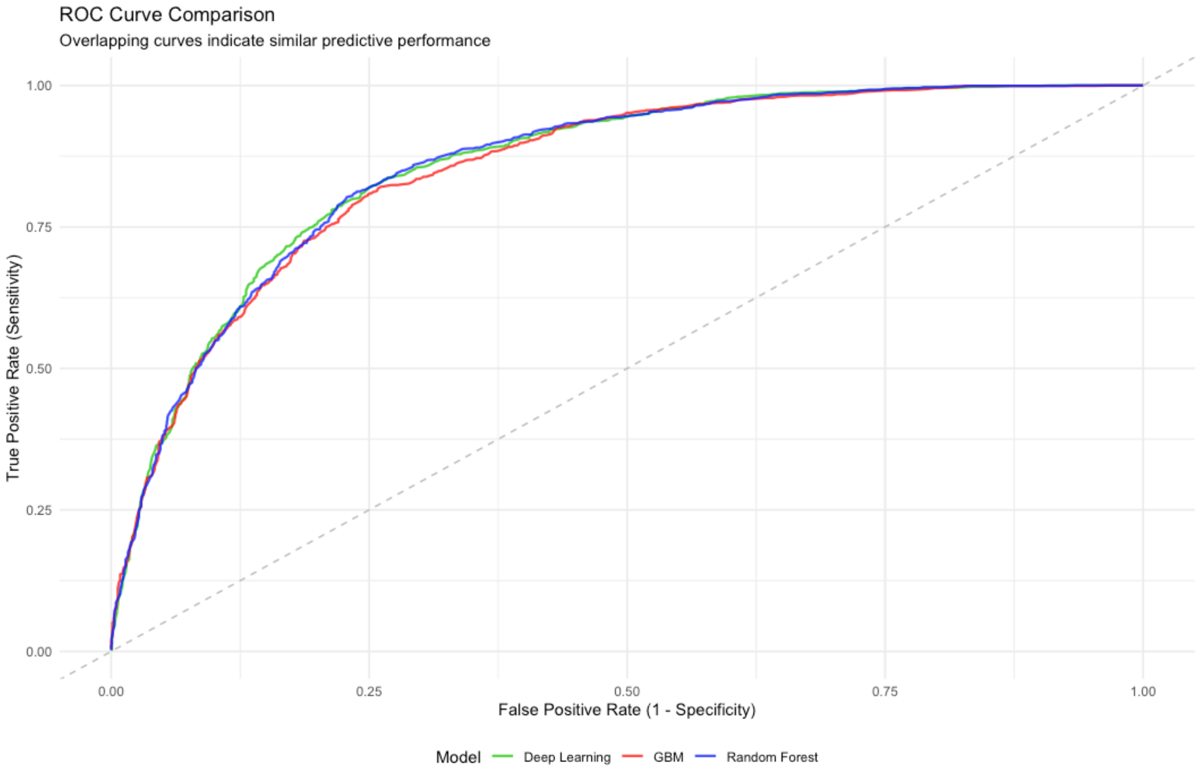

Figure 5: Combined ROC Curve

The Deep Learning Neural Network achieved the highest performance (AUC 0.8576), slightly edging out the Random Forest model. The most critical observation, visually confirmed by the overlapping curves in *Figure 5*, is the extremely narrow performance band: the difference between the winner (DL) and the third-place linear model (Lasso) is only 0.53%.

The Linearity of F1

The high performance of the Lasso GLM (AUC 0.8523), a simple linear model, proved that a large portion of the predictive signal in F1 is monotonic and easily accessible. While the Deep Learning model did achieve the highest AUC, the marginal gain (+0.0053) was minimal. This suggested that the cost and complexity of maintaining the Deep Learning model are not justified for such a small increase in performance. Therefore, in conclusion, F1 is primarily a linear problem: the car's speed (Grid Position) is the fundamental factor, which is easily captured by the simple linear model. The slight advantage held by the non-linear models suggests only a marginal non-linear component, likely due to the complex interaction effects and the slight predictive edge gained by the engineered features.

Feature Importance

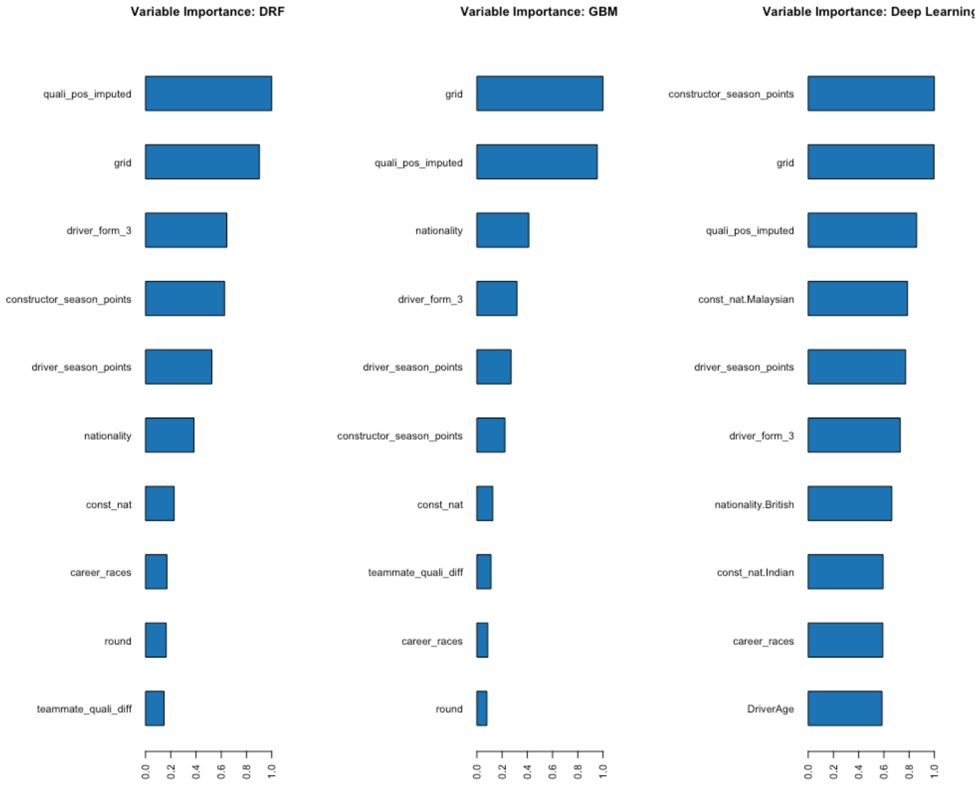

Figure 6: Feature Importance

Analysis of the Variable Importance plots *(Figure 6)* across the ensemble models provided further insight into the nature of F1 performance. Consistent with the overall findings, Starting Grid Position was the overwhelming dominant predictor for all models, confirming the superior importance of car performance. However, despite the strength of the car factor, Validation of Engineering (Research Question 3) was achieved: recent momentum (as measured by the driver_form_3 feature) consistently appeared in the top 5 features across all architectures. This validates the effort of Feature Engineering, confirming that the human element of momentum is a statistically significant, required component for successful prediction.

Conclusion

This project successfully developed and critically evaluated a predictive framework for Formula 1 championship points, achieving a robust Test AUC of 0.8576. By successfully engineering context-aware features (Driver Form, Teammate Skill) and comparing four distinct machine learning architectures, the primary objective of quantifying the predictability of the sport was met. The models confirmed that the dominant predictive signal originates from the vehicle's inherent performance, which is easily captured by the starting Grid Position.

The most significant finding of this study relates to methodological choice: although the Deep Learning Neural Network was the statistical winner, its performance gain (approximately 0.5%) over the simple Lasso GLM was negligible. This demonstrates that for tabular sports data heavily dominated by one powerful, linear predictor (the car), increasing model complexity yields diminishing returns. While the GBM and Deep Learning architectures succeeded in capturing minor non-linear and interaction effects, the interpretability and efficiency of the Lasso GLM would make it the preferred model for practical deployment.

Ultimately, the analysis proved that context matters (Research Question 3). The successful implementation and high ranking of our custom features, like driver_form_3, confirm that the human and temporal elements of the sport are statistically significant components required to separate elite drivers in similar machinery. Despite the robust predictive results, the remaining irreducible error (the 14% that couldn't be predicted) highlights the need for future work to incorporate chaotic variables, such as external data sources like live weather conditions and potentially NLP analysis of driver confidence.

Appendix

The code can be found in the **Github repository link:**
https://github.com/amaanukc/Final-Project-MACT6100

Link to dataset: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020