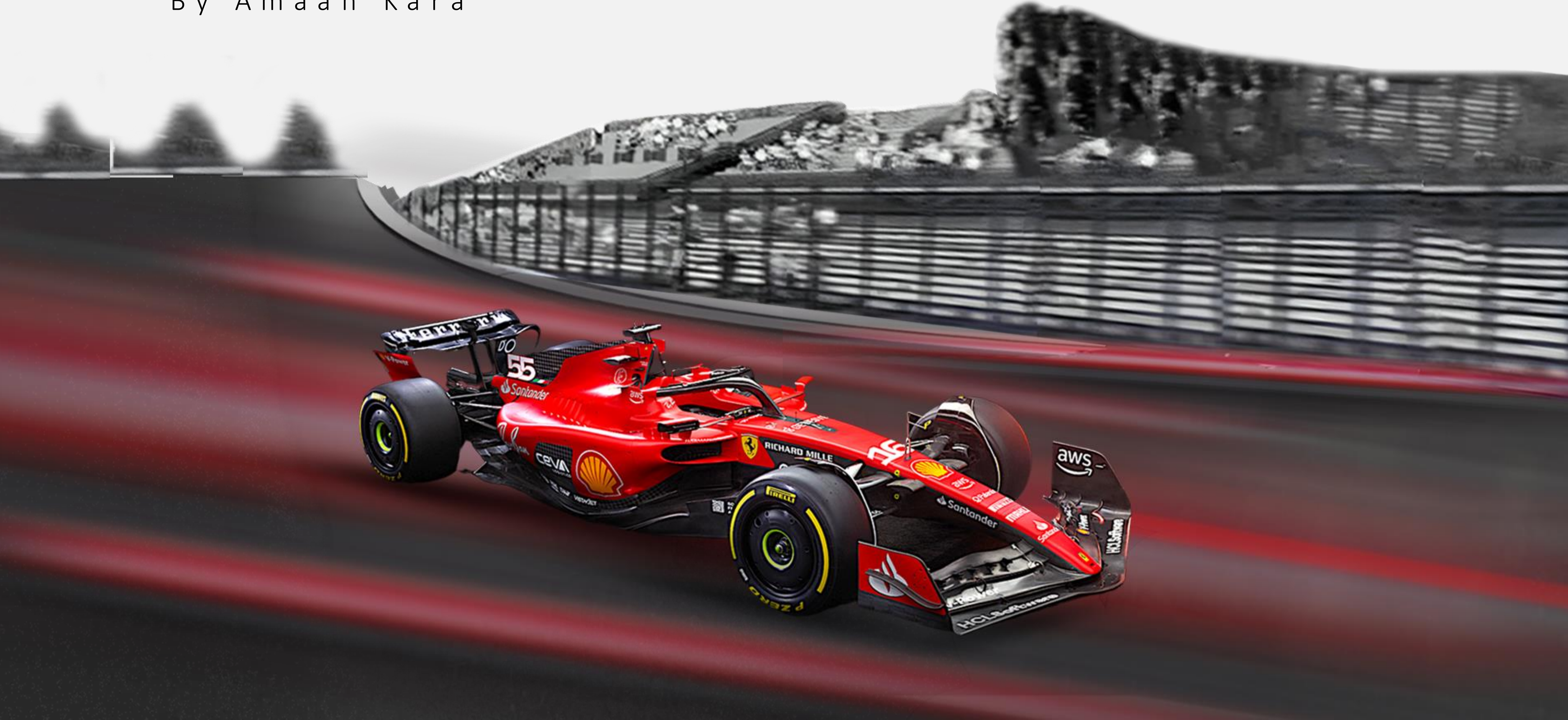
Predictive Analysis in Formula 1

By Amaan Kara



Introduction, Motivation & Research Questions

Motivation: To quantify the F1 industry's core adage: "Is it 90% Car, 10% Driver?"

Studies in Machine Learning and econometrics have applied Elo ranking systems and multi-level regression models to decompose performance, aiming to statistically separate the contribution of the driver's skill from the car's speed, a key aim we share in this project.

Prediction Goal: Predict whether a driver will score championship points (Top 10).

Research Questions:

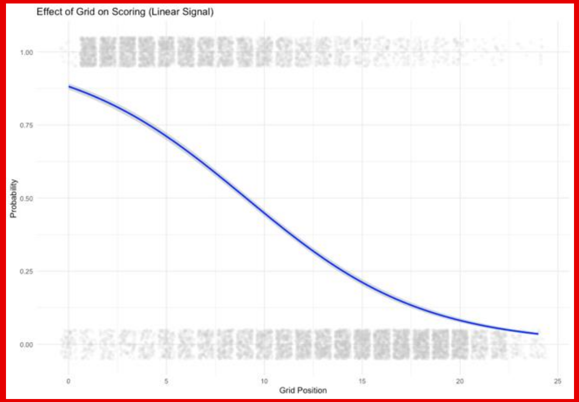
1. Can pre-race data accurately predict scoring probability?
2. Is F1 performance primarily **Linear** (Lasso) or defined by **Non-Linear Interactions** (Deep Learning)?
3. Do engineered contextual features (Momentum, Skill) add critical predictive value?

Data Scope, Features and Justification

Source and Scope:

The dataset was found using Kaggle, however for consistency the data was filtered from 2000-2024, as rules in the 20th century were different to what is used now.

To allow for analysis, missing qualifying positions were imputed to 25 (back of the grid), rather than treating these as NAs



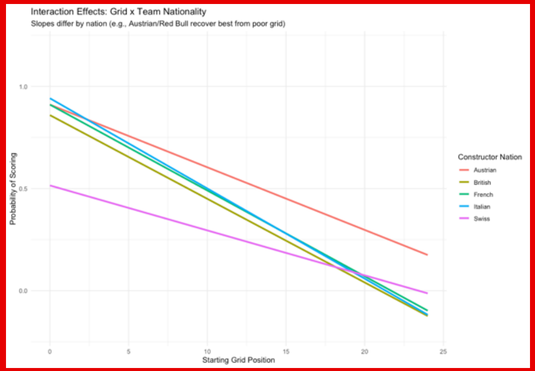
Feature Engineering:

Features engineered to capture the human element:

1. **Driver Form:** Rolling Avg. of points (Last 3 Races)
2. **Skill Isolation:** Qualifying position relative to the teammate
3. **Car Strength:** Cumulative constructor points.

Exploratory Data Analysis:

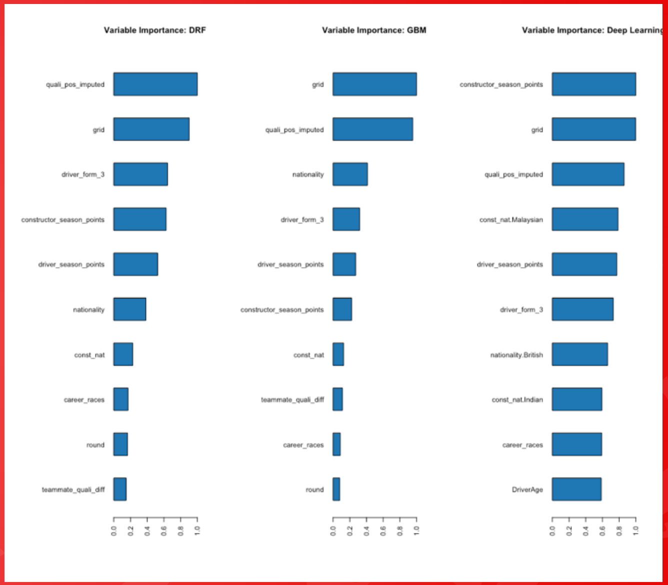
Initial EDA revealed both simple and complex patterns - Grid Position is Linear; Age and Team Interactions are Non-Linear.



The Final Verdict: Linearity Beats Complexity

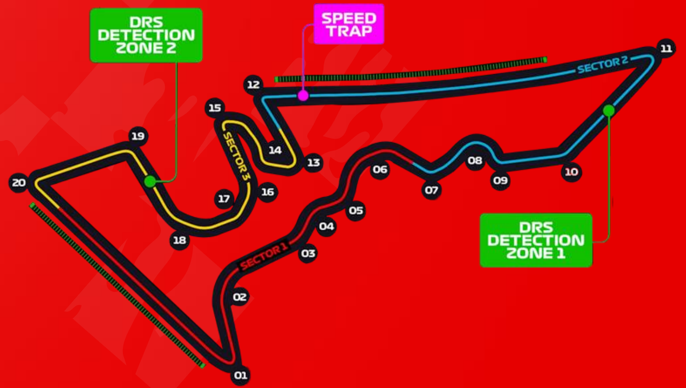
The Finding: Deep Learning achieved the highest AUC (0.8576), but the difference from the third-place Lasso GLM was only **0.53%**.

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



Success: Variable importance charts confirmed that **Driver Momentum (driver_form_3)** was consistently the #2 or #3 most important predictor.

Conclusion: Engineered features were critical for achieving peak performance.



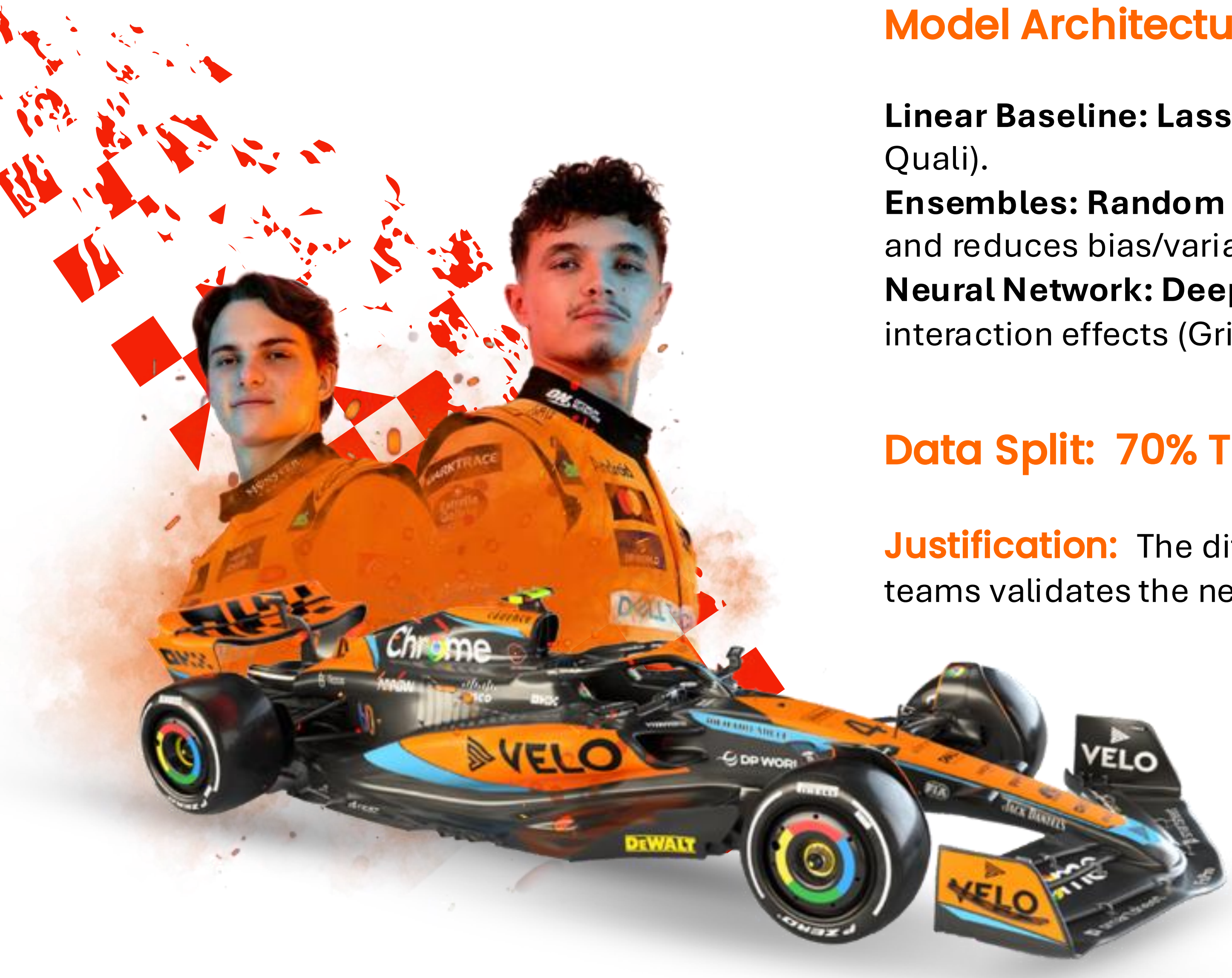
Methodology

Model Architectures

Linear Baseline: Lasso GLM - Handles multicollinearity (Grid vs. Quali).
Ensembles: Random Forest & GBM - Captures non-linearity (Age) and reduces bias/variance.
Neural Network: Deep Learning (H2O) - Targets high-order interaction effects (Grid x Team).

Data Split: 70% Training/30% Testing

Justification: The differing grid penalty slopes across teams validates the need for non-linear models.



Conclusion

Recommendation

Lasso GLM is the recommended model for business deployment. It offers the best balance of **Interpretability** and **Efficiency**, achieving near-optimal AUC with minimal cost.

F1 is Primarily a Linear Problem: The strong car signal dominates. The complex non-linear models were ultimately redundant.

Limitations

No model exceeded 86% AUC due to **Irreducible Error**. The models cannot predict unrecorded chaos variables (Weather, Mechanical Failure).

Future projects should focus on integrating external data:
1. Live Weather API to predict chaos.
2. External Penalties (grid position changes) for full context.