



Formula 1 Race Outcome Prediction Using Pre-Race Features

Team 1:

Rucha Bhale

Aiden Insall

The questions



Which pre-race factors are most predictive of a Formula 1 driver's race outcome?



Can regression models accurately predict driver finishing positions using only pre-race data, and which model performs best for this task?

Strategy



Data Collection

14 CSV files with data from Sourcing **10,494 races** from 1994 to 2024, providing a rich historical context for analysis.



Feature Engineering

Creating **6 critical features** from raw data, enhancing the predictive power of our models.



Model Development

Implementing 3 distinct regression models and an ensemble approach for comprehensive prediction.



Outcome Prediction

Generating accurate and reliable predictions for Formula 1 race outcomes and answer the 2 questions



Engineering Predictive Features

Grid Position

Starting position from qualifying determines initial race advantage

3 Recent Form (Last 3)

Short-term momentum captures current driver performance trends

5 Recent Form (Last 5)

Medium-term consistency reveals sustained competitive level

Circuit History

Driver's past performance at specific track shows track-specific skill

Team Performance

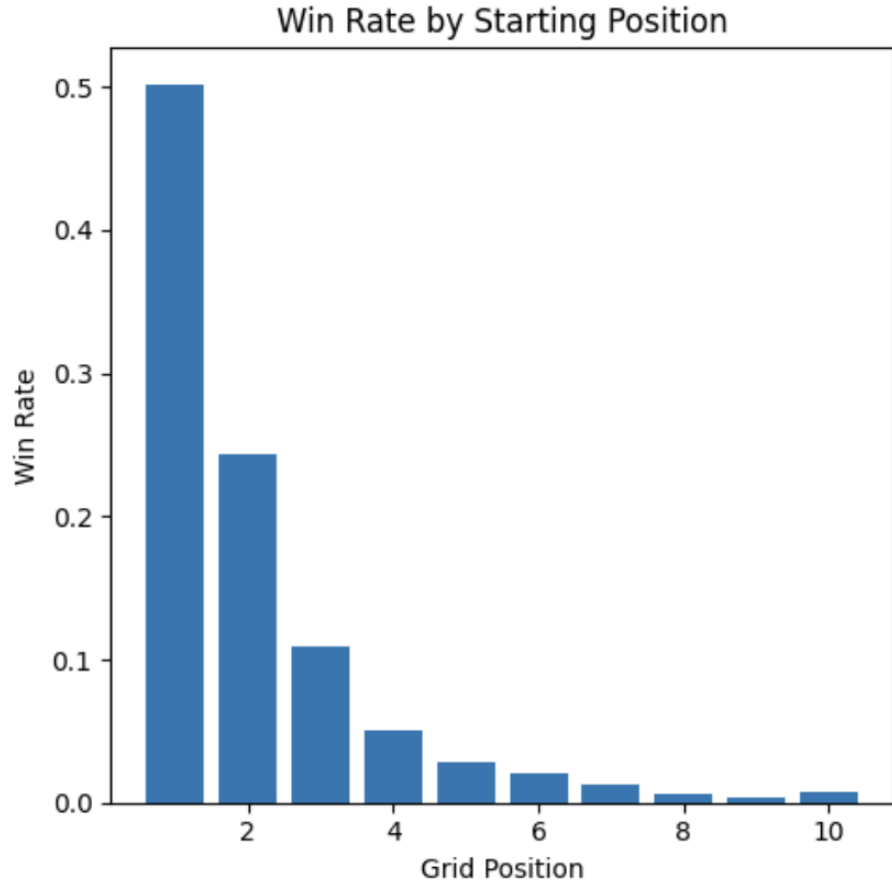
Constructor's average finish reflects car competitiveness and reliability

Championship Points

Points accumulated before race indicates overall season performance

Critical: We use `.shift(1)` to create a time barrier. Only data available BEFORE race start is used. This prevents the model from cheating by using race results to predict themselves, ensuring our predictions learn real patterns and remain trustworthy.

Which pre-race factor drives F1 success?

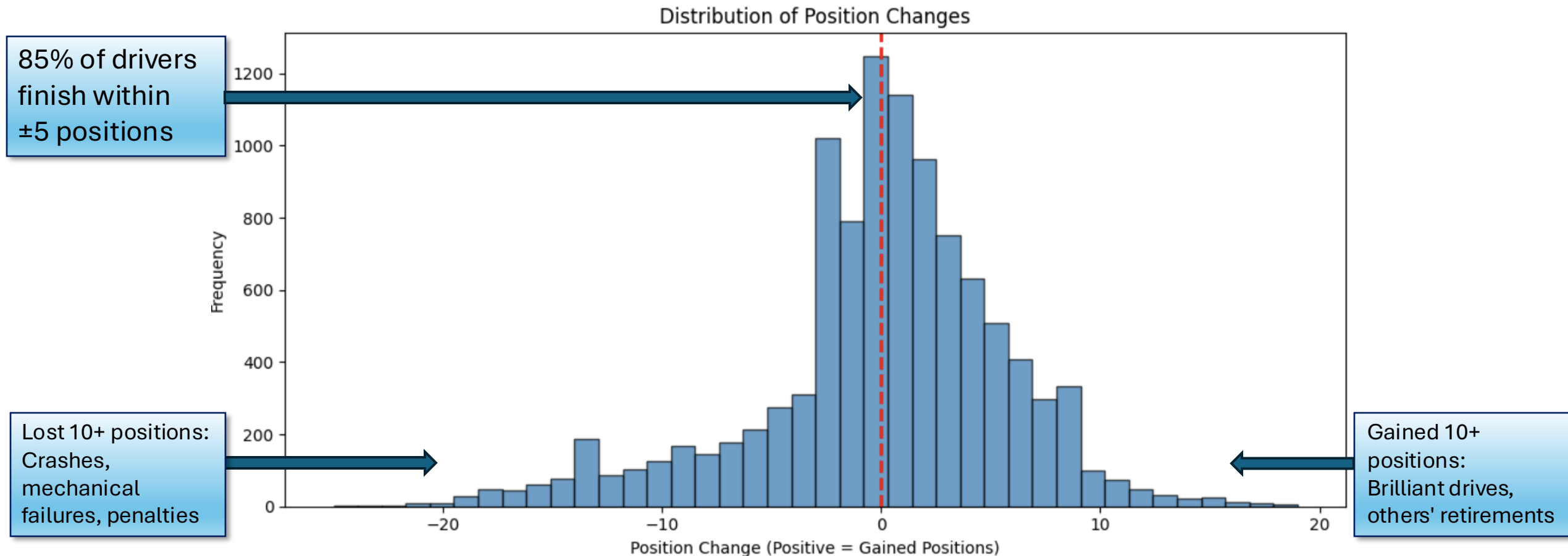


- Grid-finish position correlation: **0.58**
- Pole position win rate: **50.2%**
- P2 win rate drops to **24%**
- Strong linear relationship throughout field

*Starting position alone explains **36%** of race outcomes*



Deep Dive: But Races Aren't Predetermined

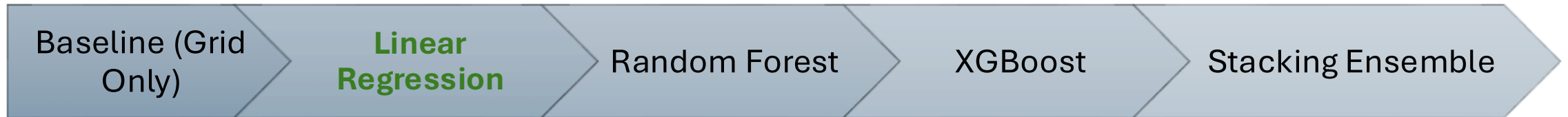


*This variance explains why even good models can't be perfect.
Our model predicts the center but it can't predict the tails*



Models Implemented

We tested several machine learning approaches to predict Formula 1 race outcomes, comparing their performance across various metrics. Our goal was to identify the most accurate and efficient model.



Model	Test MAE
Grid Only	3.86
Linear Reg 🏆	3.32
XGBoost	3.41
Stacking	3.35



Feature Importance

1

Grid Position

35% Impact

2

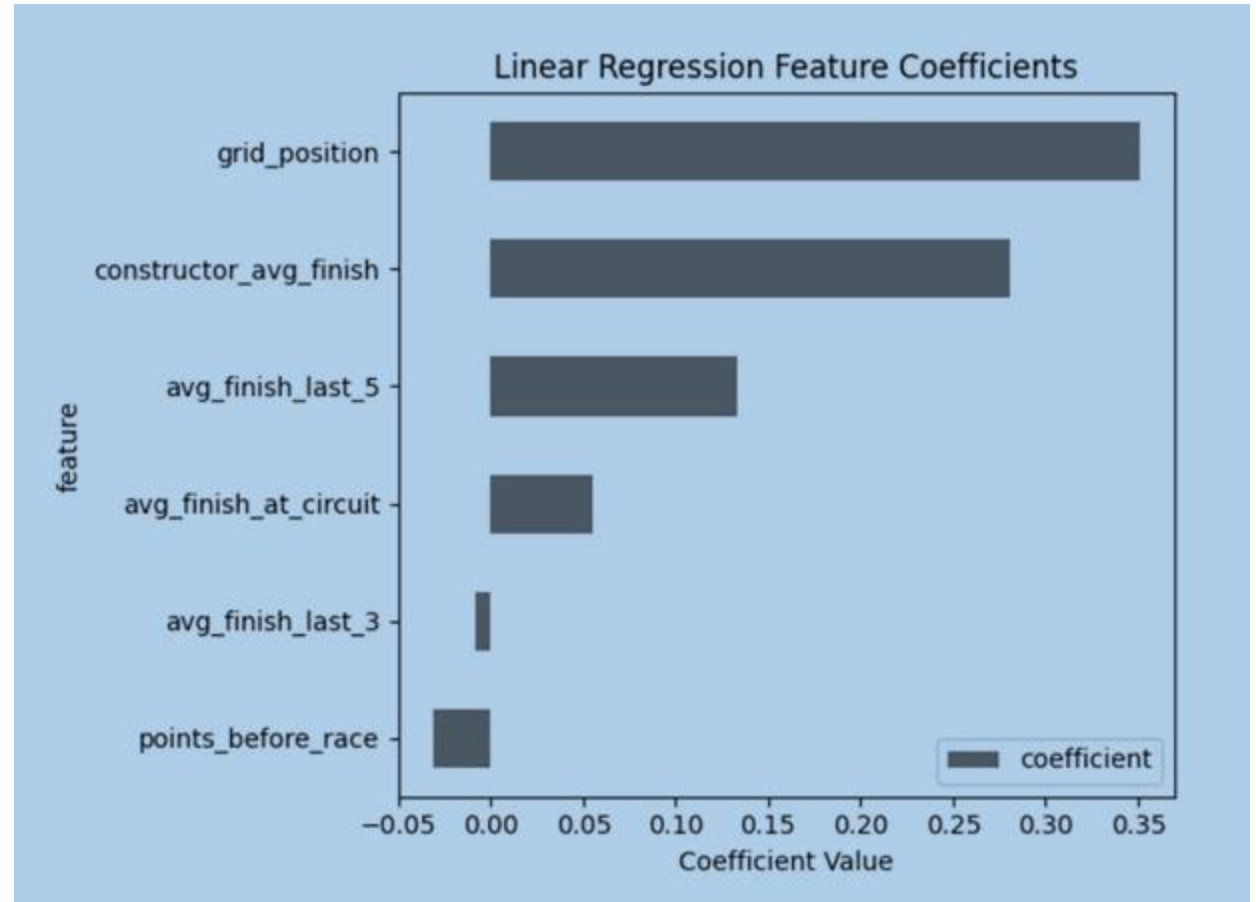
Constructor Average

28%

3

Past 5 race rolling average

13%



Real world Test

Predictions for a race from 2024

Driver	Grid	Predicted	Actual	Error
Stroll	P12	10.8	P10	0.8
Alonso	P6	9.5	P9	0.5
Pérez	P5	8.8	P2	6.8

Overall Performance: **2.93** positions average error



The 45-55 split

Our model explains 45% of race outcomes (R²)

What's predictable from the pre-race data :

- **Qualifying pace**
- **Car quality**
- **Driver/team form**

What's unpredictable:

- Crashes & incidents
- Weather changes
- Strategy calls
- Mechanical failures



This balance makes the F1 exciting !



Key Takeaways

1

1. Grid position + Constructor Average = 63% of what matters

2

2. Linear Regression best model (3.32 MAE)

3

3. F1 is balanced: half predictable, half chaos





Thank you