Formula One Race Outcome Prediction

Exploratory Data Analysis & Feature Engineering

MSDS 422: Practical Machine Learning

Authors: Sara Alsiyat · Qifan Yang · Boqi Niu

Professor: Dr. Irene Tsapara

Date: February 2026

Agenda

01 **Introduction & Problem Statement**
Why predict F1 top-10 finishes?

02 **Data Integration Strategy**
11 tables merged into one unified dataset

03 **Exploratory Data Analysis**
Key patterns and insights from 74 years of F1

04 **Feature Engineering**
35+ leakage-safe features across 12 categories

05 **Data Leakage Prevention**
How we ensured model integrity

06 **Baseline Results & Next Steps**
Model performance and future work

Problem Statement & Research Objectives

The Problem

F1 race outcomes depend on driver skill, constructor performance, track characteristics, and race conditions.

Finishing in the top 10 determines whether a driver earns championship points, a critical threshold for competitive success.

Can machine learning predict top-10 finishes using only pre-race information?

Research Objectives

- Build and evaluate ML models for top-10 finish prediction using historical F1 data
- Examine how qualifying position, constructor, driver history, and circuit affect outcomes
- Explore circuit-level variation in race unpredictability
- Compare models for accuracy, interpretability, and decision-support value

Dataset & Integration Strategy

1950–2024	26,000+	11	80+
Time Span	Driver–Race Records	Tables Merged	Features Created

Tables Integrated

Results	Races	Drivers	Constructors
Circuits	Qualifying	Status	Pit Stops
Lap Times	Driver Standings	Constructor Standings	

Target Variable: top10_finish — binary indicator (1 = driver finished in top 10, 0 = otherwise). Derived from positionOrder ≤ 10.

Key EDA Insights

Grid Position Dominance

Starting position is the strongest predictor of top-10 finishes. Pole position converts to top-10 at >95% rate; grid 15+ drops below 30%.

Constructor Effect: 88%

Van Kesteren & Bergkamp (2023) found ~88% of F1 result variance is attributable to the constructor, not the driver.

DNF Rate Decline

Did Not Finish rates have dropped from ~40% in the 1950s to under 10% in recent seasons, reflecting reliability improvements.

Circuit Variation

Some circuits show much higher position volatility, suggesting more unpredictable races — important for modeling uncertainty.

Feature Engineering: 35+ Features Across 12 Categories

Temporal

Driver age, race month, season stage

Performance History

Rolling 5-race top-10 rate, DNF rate, avg position

Constructor

Team rolling stats, season top-10 rate

Circuit

DNF rate, volatility, driver-circuit history

Driver–Team Interaction

Races together, joint success rate

Qualifying

Grid position, front-row start, top-5/top-10 flags

Competitive Context

Field size, grid percentile

Momentum

Top-10 streak, points last 3, position trend

Pit Stop

Avg stops per race, constructor pit efficiency

Championship

Standings position, points gap to leader

Era

Regulation era encoding (1950s to ground effect)

Categorical

Label-encoded driver, constructor, circuit, country

Data Leakage: The Issue & The Fix

What Went Wrong

Initial models scored 1.000 ROC AUC — a clear sign of data leakage.

Post-race features were included as model inputs:

- positionOrder — directly defines the target
- points — only awarded to top-10 finishers
- milliseconds, laps — post-race outcome data
- Championship standings included current race

How We Fixed It

Strict Leakage Guard
Explicit blocklist of 12+ post-race columns removed from feature matrix.

Temporal shift(1) on All Rolling Features
Every expanding/rolling window excludes the current race.

Championship Standings Fix
Shifted to previous round's standings so model only sees pre-race info.

Leakage Audit Cell
Automated check flags any feature with |correlation| > 0.85 to the target.

Baseline Model Performance

Temporal 3-way split: Train (< 2018) | Validation (2018–2021) | Test Holdout (2022–2024)

Validation Set (2018–2021)

Model	ROC AUC	PR AUC	F1
Logistic Regression	0.822	0.794	0.727
Random Forest	0.840	0.802	0.781
Gradient Boosting	**0.827**	**0.801**	**0.755**

Test Holdout (2022–2024)

Model	ROC AUC	PR AUC	F1
Logistic Regression	0.857	0.849	0.775
Random Forest	**0.869**	**0.854**	**0.808**
Gradient Boosting	0.858	0.833	0.796

Key Takeaways

- Random Forest achieves best overall performance (0.869 ROC AUC on holdout)
- Test scores slightly higher than validation — no overfitting detected
- All models use only pre-race features after leakage fix — results are realistic

Implementation Strategy & Pipeline

1 — 2 — 3 — 4 — 5 — 6

1. Data Collection
Kaggle F1
11 CSV tables

2. Data Integration
Merge into
driver–race rows

3. EDA
Patterns,
missing data

4. Feature Engineering
35+ features
shift(1) safe

5. Leakage Audit
Block post-race
columns

6. Baseline Models
LR, RF, GB
3-way split

Key Design Principles

Reproducible pipeline: From EDA → feature engineering → modeling → evaluation

Consistent model comparison: Same split, same preprocessing, same metrics across all models

Pipeline-based preprocessing: All transformations applied in a structured, reusable way

Validation-first mindset: Model choices based on validation results; test set held out for final confirmation

Next Steps & Future Work

1

Hyperparameter Tuning

Grid search / Bayesian optimization on Random Forest and Gradient Boosting using the validation set.

2

Advanced Models

XGBoost, LightGBM, and TabNet (Urdhwareshe, 2025) for stronger gradient boosting baselines.

3

SHAP Interpretability

Feature-level explanations to understand which factors drive individual predictions.

4

Weather & Safety Car Data

Incorporate real-time race variables as identified by Jafri (2024) for improved accuracy.

5

Model Deployment

Build a prediction interface for pre-race top-10 probability estimates.

Thank you...

Sara Alsiyat · Qifan Yang · Boqi Niu

MSDS 422 · Dr. Irene Tsapara · Northwestern University