Predicting F1 Undercut Success in the Hybrid Era (2014–2024)

Evan Jiang Lewis Tu Paul Jeon

Harvard University
CS 1090a: Data Science I

Milestone 5 – Final Presentation

The Undercut Strategy & F1 World Championship Dataset

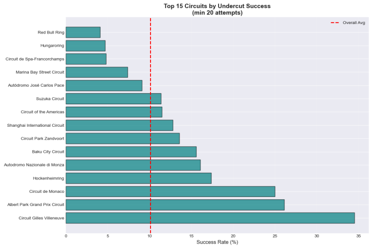
- An *undercut* occurs when a trailing driver pits **before** the car ahead.
- Idea: fresh tyres $\rightarrow$ faster laps $\rightarrow$ gain enough time before the rival pits.
- When the leader pits, outcome is binary: leader stays or attacker jumps ahead.
- Teams must decide this in **real-time**.

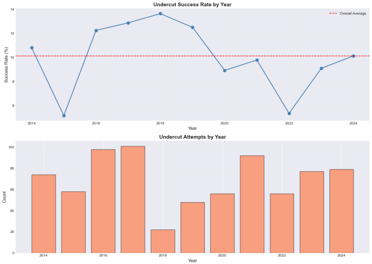
Goal: *Predict undercut success using gap size, tyre age, circuit & season context, etc.*

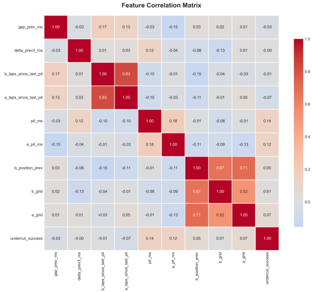
- **Source:** Ergast Developer API via Kaggle, using official FIA timing data (2014-2024).
- **Scale:** 248k lap times, 8.36k pit stops, 228 races, and 32 circuits.
- **Undercut attempts:** attacker ≤ 2 sec behind. **761** legit attempts, with only 10% success.

Exploratory Data Analysis: Three Key Patterns

- **1. Circuit dominance** — huge variation across tracks Montreal (34.5%), Monaco (25%) vs global 10.1%.
- **2. Pit stop performance** — pit time differential is the strongest linear predictor (ρ 0.12–0.14).
- **3. Temporal variation** — year-to-year success ranges 5–14%.







Feature Engineering

Race dynamics

- Gap to car ahead (ms)
- Tyre age differential
- Recent pace differential
- Pit stop duration differential

Contextual

- Circuit baseline success rate
- Season / year effects
- Starting grid positions

All feature engineering details are in the notebook.

Baseline: Logistic Regression with Circuit Dummies

Setup

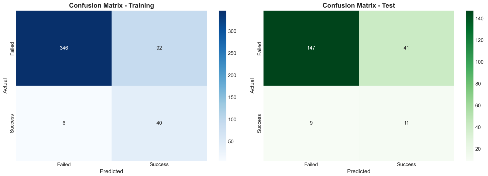
- Logistic regression + class weighting.
- 48 Inputs: race features + 30+ circuit one-hot dummies.

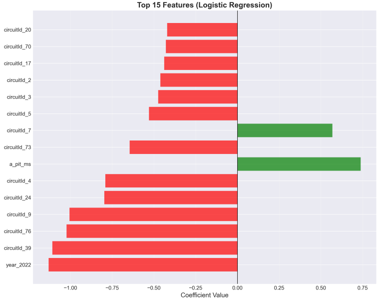
Performance (test)

- AUC-ROC: **0.713**, F1: 0.306

Limitation

- Many correlated circuit dummies.
- Hard to interpret & generalize for rare circuits.





Final Model: Hierarchical Circuit-Level Effects

Key idea

- Circuits dominate → model with **partial pooling**.

Stage 1: Empirical Bayes baselines

- Estimate circuit-specific success rates.
- Apply shrinkage toward global mean.

Stage 2: Logistic regression

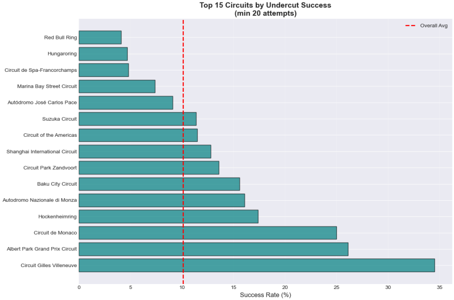
- Only 15 features (vs 48): 7 dynamics + 1 circuit baseline + 7 year dummies.

Model:

$$\log \frac{p}{1-p} = \beta_0 + \mathbf{x}^\top \beta + \beta_{\text{circ}} \cdot \text{baseline}$$

Training details

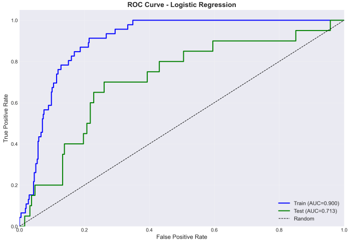
- 532 samples, stratified split.
- L2 regularization ($C = 1.0$).
- Balanced class weights.

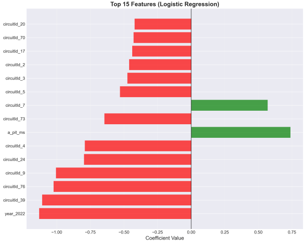


Model Comparison: Baseline vs Hierarchical

Metric	Baseline	Hierarchical
AUC-ROC	0.713	0.683
F1	0.306	0.289
Features	48	15

- Hierarchical model: **massive reduction** in features.
- Circuit baseline is most important feature.
- Gap, pit times, tyre age become cleaner to interpret.





Key Insights & Practical Value

- **Circuit is key** — Montreal/Monaco $3\times$ easier than average.
- **Execution matters** — pit stop differential strongly affects success.
- **Hierarchical modeling fits grouped F1 data** — much more interpretable.

Practical use: *Teams can input gap, pace, pit times, tyre ages, circuit $\rightarrow$ get real-time undercut probability.*

Limitations & Future Work

Limitations

- Linear model cannot capture complex interactions.
- Missing driver skill, tyre compound, weather.
- Class imbalance challenges remain.
- Only $\leq 2s$ gap attempts included.

Future Work

- Use ensemble methods such as XGBoost or Random Forest.
- Add driver/team random effects.
- Include tyre compounds, weather, safety cars.
- Build real-time strategy API for race engineers.