CS 1090a Milestone 3

Evan Jiang, Lewis Tu, Paul Jeon

November 14, 2025

1 Finalized Research Question

Can we predict the success of an undercut pit stop strategy in Formula 1 based on real-time race conditions?

Background

An undercut is a pit stop maneuver where a driver pits before a rival ahead, taking advantage of fresh tires to gain time. This tactic has been especially important in the Hybrid Era (2014–present) due to reduced overtaking opportunities. Decisions depend on the gap to the car ahead, tire degradation, recent pace differential, circuit characteristics, and pit stop execution.

Predictive Goal

Our goal is to build a model that estimates the probability that an undercut attempt results in a position gain. The target label is `undercut_success` (1 for success, 0 for failure).

2 Data Description

The dataset comes from Kaggle's "Formula 1 World Championship (1950–2024)" built from the Ergast API and includes lap timing, pit stops, race metadata, results tables, and circuit characteristics. Key files include `lap_times.csv`, `pit_stops.csv`, `races.csv`, `results.csv`, and `circuits.csv`. The analysis focuses on 2014–2024 due to consistent hybrid regulations and improved timing fidelity.

3 Summary of the Data

The undercut dataset was created by identifying pit stops where a driver had a rival directly ahead, tracking the rival's response within five laps, computing features such as gaps, pace metrics, tire age, and pit stop durations, and labeling success based on post-pit positions. A 2-second gap threshold was applied to prune pit stops where undercuts are not realistic. The threshold removes non-strategic cases and leaves 761 valid attempts.

Important features include the pre-pit gap in milliseconds, pit lap numbers, three-lap pace averages for both drivers, pace differential, tire age for both drivers, pit stop durations, circuit and year identifiers, and the binary success label.

After filtering, gaps range between 0.2 and 2.0 seconds, tire age typically ranges from 1 to more than 40 laps, and pit stop durations fall within 18–35 seconds. Missing data is minimal.

4 Deeper Understanding of the Data

Before filtering, the dataset had roughly 94% failed undercuts and 6% successful ones due to cases with unrealistic gaps. After applying the 2-second filter, the overall success rate increased to roughly 10.1%, producing a more realistic strategic dataset. This filtering step improves dataset quality by aligning with what physics and race strategy allow.

The middle 50% of undercut attempts occur when the attacking driver is between 0.74 and 1.54 seconds behind after filtering. Tire age shows the middle 50% of attempts occur at 6–17 laps on tires, and tire age differentials often influence strategy more than absolute ages. Pace differential varies in a wide range from -2 to $+2$ seconds per lap, and many undercuts are attempted even when the attacker is slower, showing that teams often pit preemptively to avoid traffic.

Correlations with undercut success are weak: gap (-0.03), attacker tire age (-0.01), pace differential (-0.004), pit duration ($+0.14$). This suggests nonlinear models may later outperform linear ones.

From 2014 to 2024, undercut success rates vary between 5.2% and 13.6% without strong trends. Circuit effects are much more pronounced, ranging from about 25% at Monaco to about 34.5% at Circuit Gilles Villeneuve, indicating that pit lane length, tire degradation and overtaking difficulty play major roles.

5 Visualizations

The notebook includes the following visualizations: class distribution charts, histograms for gaps and tire age, a correlation heatmap, year-over-year success rate plots, circuit-specific success charts, ROC curves, precision-recall curves, and confusion matrices. Each visual is labeled and includes explanatory context.

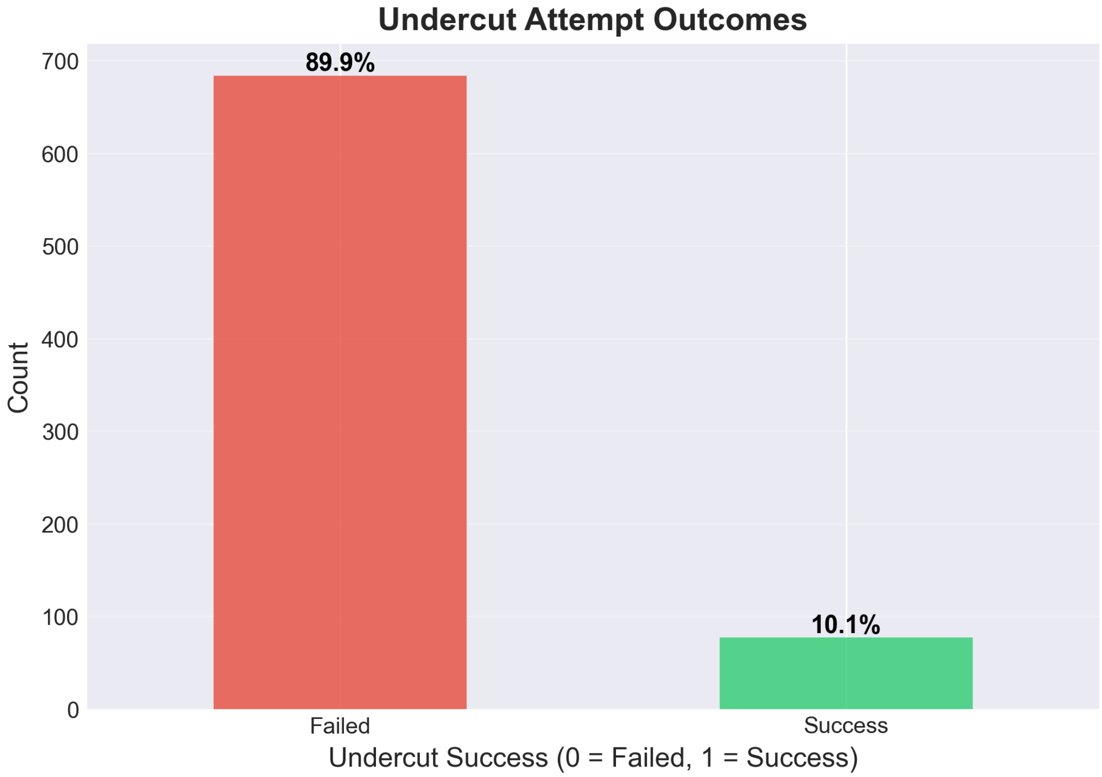

Figure 1: Outcome distribution for filtered undercut attempts.

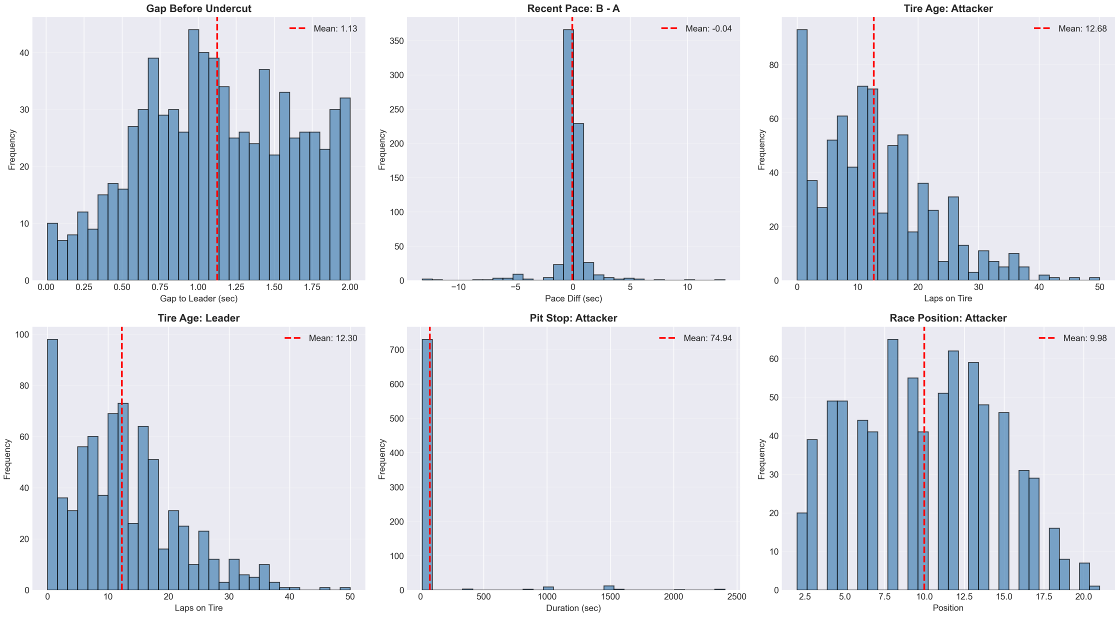

Figure 2: Feature distributions for gap, tire age, and pace differential.

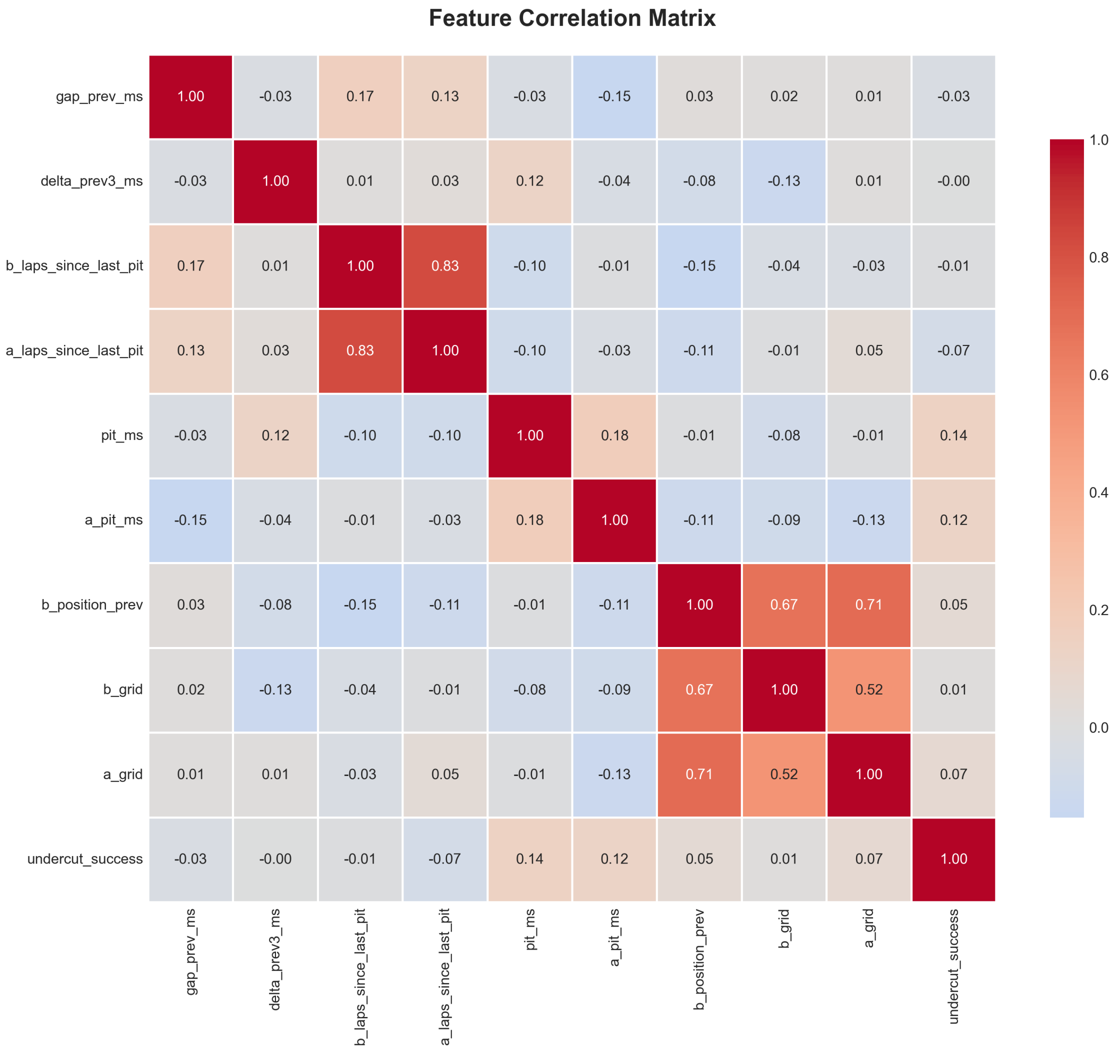

Figure 3: Feature correlation matrix.

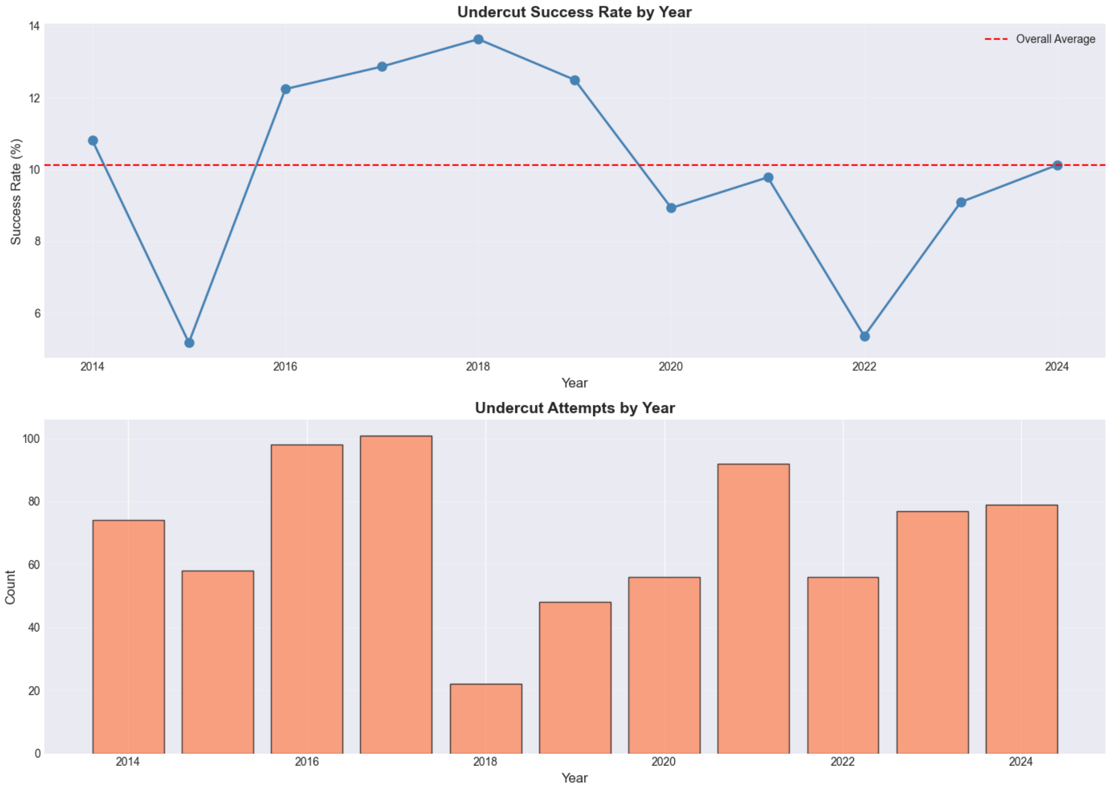

Figure 4: Temporal trends in undercut success rate (2014–2024).

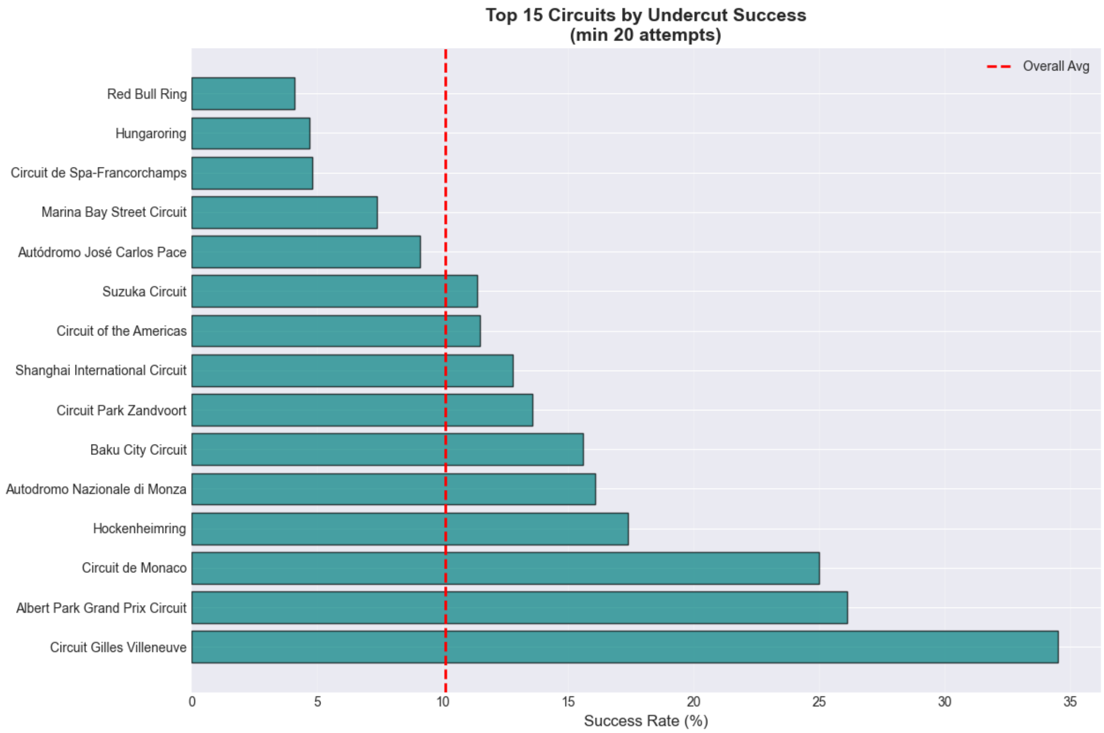

Figure 5: Top circuits by undercut success rate.

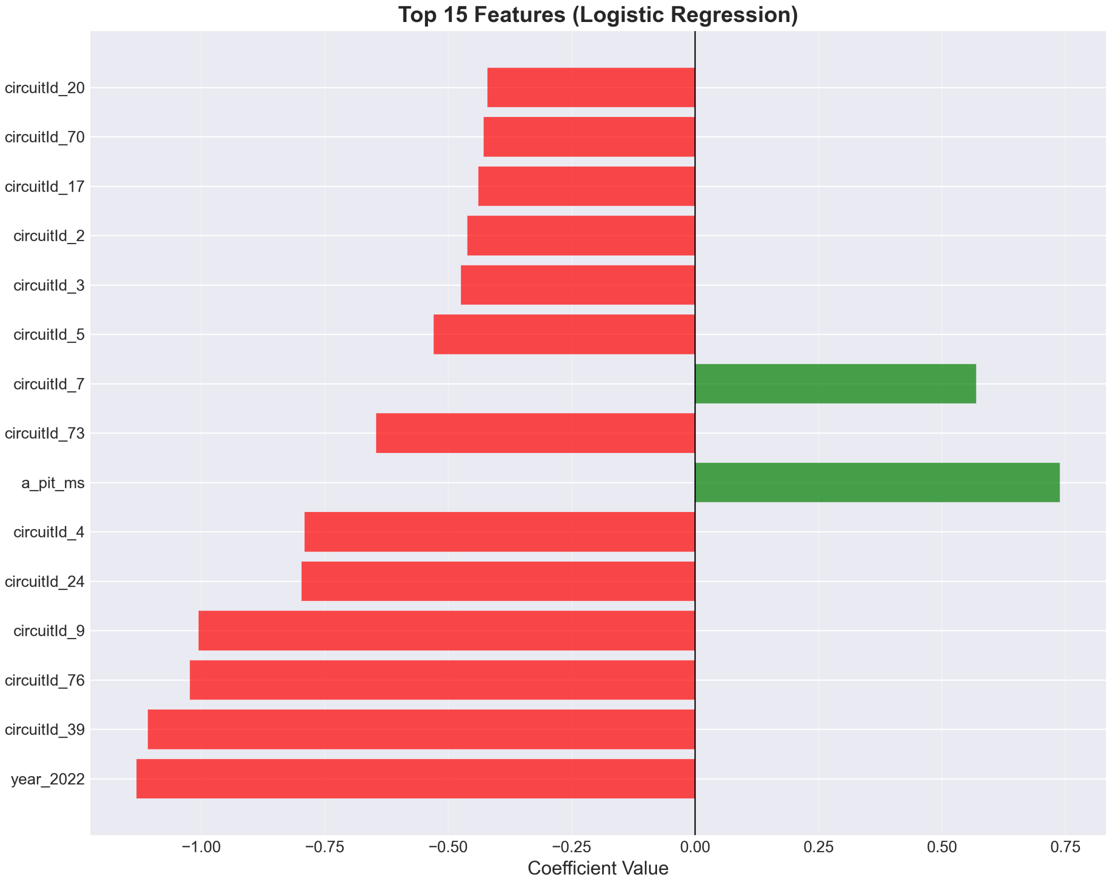

Figure 6: Standardized logistic regression feature importance.

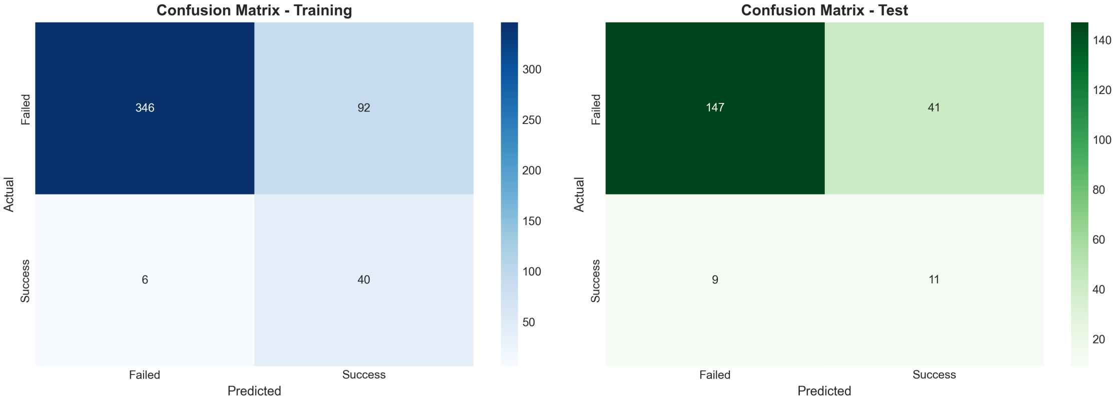

Figure 7: Confusion matrices for the baseline logistic regression model.

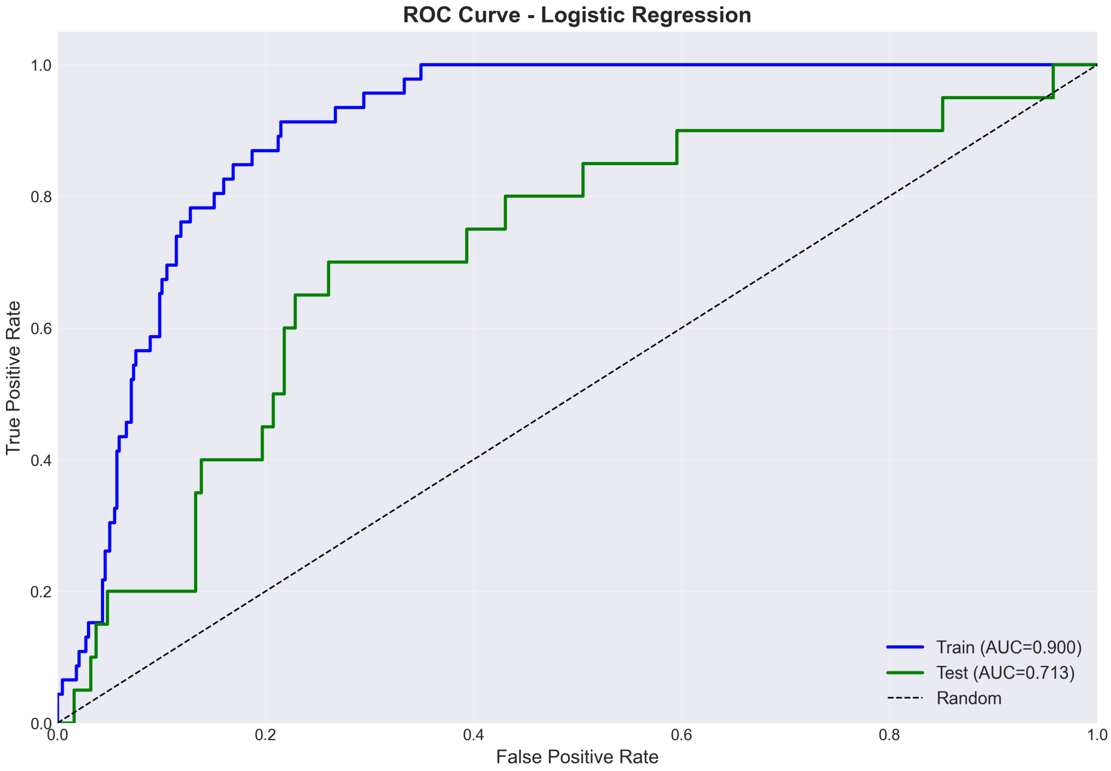

Figure 8: ROC curve and AUC for the logistic regression model.

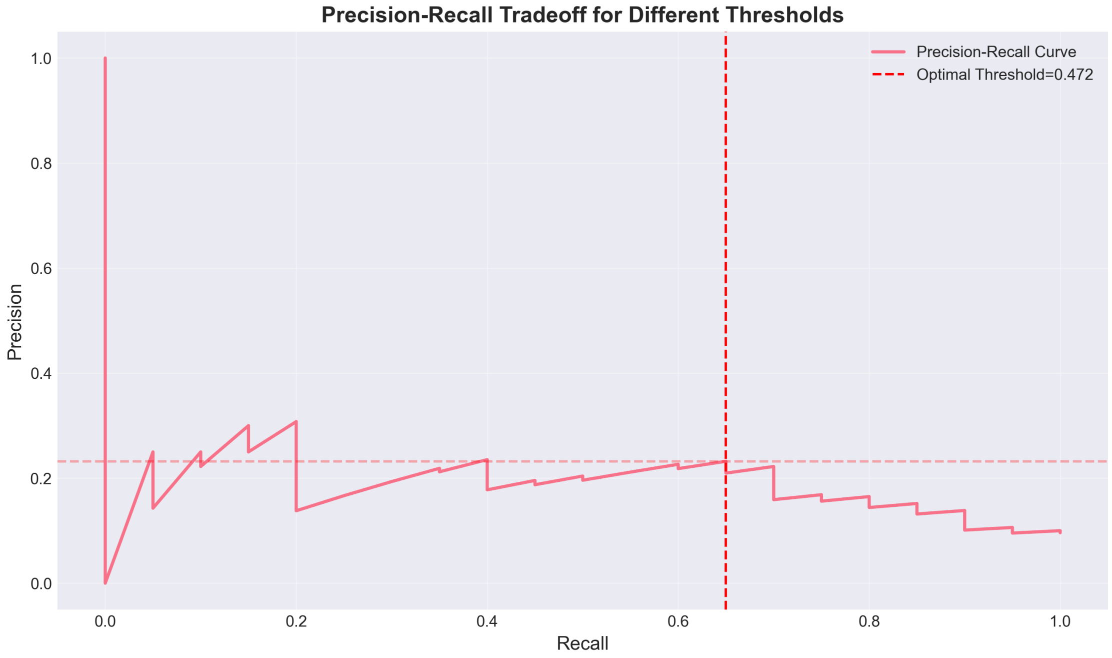

Figure 9: Precision–recall tradeoff curves across thresholds.

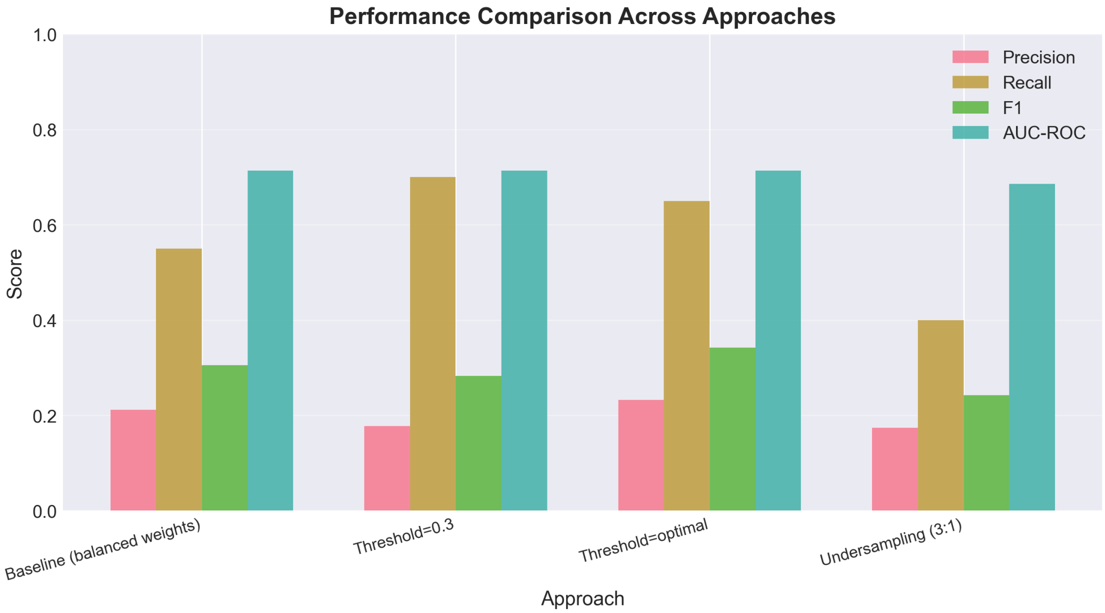

Figure 10: Comparison of model performance under different imbalance strategies.

6 Noteworthy Findings

The 2-second gap filter is crucial for obtaining realistic data and removes many non-strategic pit stops. Gap is the most important individual feature but is insufficient alone; tire age differential and circuit characteristics matter significantly. Circuits exhibit large variability, with approximately three-fold differences in success rates between top and bottom circuits. Pit stop duration shows positive correlation ($+0.14$), indicating faster pit stops for the attacker increase success probability. Roughly one-third of attempts occur when the attacker is slower, reflecting preemptive strategy.

7 Baseline Model

Mathematical Framework

For binary undercut success, we model:

$$P(y_i = 1 \mid X_i) = \sigma(X_i^\top \beta) = \frac{1}{1 + e^{-X_i^\top \beta}},$$

with log-odds

$$\log\left(\frac{p}{1-p}\right) = X_i^\top \beta.$$

Class imbalance is handled via class weights:

$$w_{\text{class}} = \frac{n}{2n_{\text{class}}}.$$

The weighted log-likelihood is

$$L(\beta) = \sum_i w_i \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right].$$

Training Setup

Training uses a 70/30 stratified split, standardization, logistic regression, and class weighting. The model includes features such as gap, pace differential, tire ages, pit stop durations, tire age differential, gap per lap, and one-hot encoded circuit and year features. After encoding, the model uses roughly 40 inputs.

Performance

On the test set, the model achieved accuracy of 0.76 (not meaningful due to imbalance), precision of 0.21, recall of 0.55, F1 score of 0.31, and AUC–ROC of 0.71, which is the most informative metric since it is threshold-independent and robust to imbalance. The model performs better than random chance (AUC–ROC > 0.50) and demonstrates predictive capability. Circuit characteristics dominate feature importance, with circuit IDs comprising most of the top features, validating the EDA finding that circuit is the dominant factor. Pit stop times (a_pit_ms) are the strongest non-circuit predictor, and year 2022 shows a strong negative effect, consistent with temporal variation observed in EDA.

8 Addressing Class Imbalance

Threshold tuning replaces the default 0.5 threshold with a value that maximizes F1, typically around 0.47, increasing recall to 0.65 but reducing precision to 0.23. A lower threshold of 0.3 increases recall further to 0.70 but reduces precision to 0.18. Random undersampling reduces the majority class to a 3:1 ratio relative to the minority, achieving precision of 0.17 and recall of 0.40, though this approach was detrimental to AUC–ROC, precision, and recall compared to the baseline.

Comparing methods shows that the baseline (balanced weights) yields precision of 0.21 and recall of 0.55 with F1 of 0.31; threshold tuning at 0.3 increases recall to 0.70 but reduces precision to 0.18 with F1 of 0.28; the optimal F1 threshold (0.47) yields precision of 0.23 and recall of 0.65 with F1 of 0.34; random undersampling yields precision of 0.17 and recall of 0.40 with F1 of 0.24. The optimal threshold approach provides the most balanced performance, but selection depends on whether false positives or false negatives are more costly.

9 Conclusion

We built a clean undercut attempts dataset for 2014–2024, filtered for realistic undercut attempts (2-second gap threshold), performed detailed EDA, engineered features such as tire age differential and gap per lap, and constructed a baseline logistic regression model with class imbalance treatment. The final dataset contains 761 legitimate undercut attempts after filtering. The EDA strongly influenced modeling choices such as the gap filter, imbalance handling via class weights and threshold adjustments, and the use of circuit and tire age differential features.

The baseline model achieved an AUC–ROC of 0.71 on the test set, demonstrating predictive capability beyond random chance. Circuit characteristics dominate feature importance, with top-performing circuits (Circuit Gilles Villeneuve at 34.5%, Albert Park at 26.1%, Monaco at 25.0%) achieving 2.5–3.5× higher success rates than the overall 10.1% average. Pit stop times show the strongest linear correlation with success (pit_ms: 0.14, a_pit_ms: 0.12), and weak linear relationships (all correlations < 0.15) suggest non-linear interactions dominate.

This model can support real-time strategic decisions, risk assessments, circuit-dependent analysis, and tire management optimization. Future improvements include SMOTE-based augmentation, XGBoost or LightGBM with imbalance-aware loss functions, cost-sensitive learning, interaction terms (circuit × pit_time, gap × tire_age_diff), weather and temperature features, and nonlinear deep learning architectures. Given the dominance of circuit effects, hierarchical models that explicitly account for circuit-specific characteristics should be explored, potentially using circuit-specific intercepts or random effects to capture the substantial variation in success rates across tracks.

For full code and visualizations, refer to the Jupyter notebook `m3-coding.ipynb`.