Milestone 2

Predicting Successful F1 Undercuts

1. Revised Problem Statement

Our project aims to model the success of an "undercut" strategy in Formula 1's modern hybrid era (2014-present). The undercut occurs when a trailing car (Car B) pits earlier than the car ahead of it (Car A), seeking to gain a time advantage from new tires.

Our objective is to identify and quantify the key factors that predict whether an undercut attempt will be successful. We have created a dataset of all valid undercut attempts and will use features like the time gap, relative pace, tire age, and pit stop durations to build a classification model that predicts a binary outcome: undercut_success.

2. Dataset Description

We accessed our data from the "Formula 1 World Championship (1950 - 2024)" dataset available on Kaggle (linked here). For this project, we are using several core CSV files:

- races.csv: Contains metadata for each race, including year, which is used for filtering.
- lap_times.csv: Provides lap-by-lap time, position, and millisecond data for each driver in each race.
- pit_stops.csv: Logs the lap, duration, and stop number for each pit stop for each driver in each race.
- results.csv: Provides each driver's starting grid position, final position, finish status id, and their constructor.
- status.csv: Defines the meaning of finish status ids.

3. Data Issues and Preprocessing Steps

These steps have been implemented in the attached .ipynb file.

Data Missingness

- The raw CSV files use the string \N to represent NULL values, which pandas does not recognize by default.
- When loading each CSV, we set na_values=r'\N'. This correctly converted all \N strings to NaN, allowing for proper handling with pandas.
- During our analysis, our logic to find the car ahead produced NaNs for the race leader (who has no one ahead of them). When identifying pit events, we dropped these rows with ~pit_events['a_driverId'].isna() as a car cannot perform an undercut attempt without a car in front of them.

Data Relevance

- The dataset spans 70+ years, but F1 regulations change constantly. A pit stop strategy from 1970 is irrelevant to one in 2025 due to major changes in refueling, tire technology, and car aerodynamics.
- We filtered the races table to include only the modern "Hybrid Era," defined as year >= 2014. This era involved completely new power units for cars, which directly affect how cars manage energy, tire wear, and cooling, all of which are essential elements of pit stop strategy.
- All other datasets were subsequently filtered to only include raceIds from this era.

Feature Engineering & Transformation

- Pace and Gap Calculation: We calculated cumulative race time for each driver per lap. We then used this to calculate the gap to the ahead car for every car on every lap. We also used a rolling 3-lap window to calculate a proxy for a driver's recent pace.
- Stint & Tire Age: We calculated laps since the last pit stop to serve as a proxy for tire age, a critical predictor.
- Identifying Undercut Attempts: We defined an undercut attempt as an event where a trailing car (Car B) pits, and the ahead car (Car A) pits within the next 5 laps.
- Defining Undercut Success: For each undercut attempt, we determined the success by comparing Car B's position to Car A's position on the lap after Car A completed its pit stop. If B emerged ahead, then we would define that as an undercut success.

Future Steps

- Data imbalance is a potential issue, as we found there to be significantly more unsuccessful undercut attempts compared to successful ones. We plan to test oversampling and undersampling techniques, along with class weight adjustments within our classifier models.
- Data scaling is another potential issue, as our final feature set contains variables on vastly different scales. We plan to apply StandardScaler before training our models to address this.