GoTo Data Science Take-Home Assignment Write-up

Objective

The primary goals of this assignment were to debug a broken ML allocation pipeline (Task 1) and significantly improve the model's performance via robust feature engineering (Task 2).

Task 1: Pipeline Bug Detection and Fixes

The pipeline was successfully debugged and fixed end-to-end. Three major classes of issues were identified and resolved:

1. Data Integrity and Logical Flaws

The booking_log.csv file was structurally corrupted, resulting in a persistent KeyError: 'event_type'. This critical column is required to determine trip completion status.

- **Fix:** The data pipeline (src/data/make_dataset.py) was made robust by overriding the flawed log. Completion status (is_completed) was inferred by using the is_accepted target variable as a proxy, which unblocked the rest of the pipeline.

2. Structural & Architectural Bugs

- A **Circular Import** between src/models/classifier.py and src/models/train_model.py was eliminated by restructuring the model definition.
- The base classifier (src/models/classifier.py) contained a NotImplementedError, which was replaced with a functional RandomForestClassifier.

3. Pipeline Rigidity

The introduction of a new argument to apply_feature_engineering in Task 2 exposed missing argument calls in src/models/predict_model.py and src/features/transformations.py, requiring updates to pass configuration variables (target_col) dynamically throughout the pipeline.

Task 2: Model Improvement via Feature Engineering

Feature Implemented: Historical Acceptance Rate

To achieve high performance using Occam's Razor, the most predictive, non-leaking, time-series feature was chosen: **Driver's Historical Acceptance Rate** (historical_acceptance_rate).

Rationale: The decision to accept an offer is highly dependent on a driver's prior successful interactions. This cumulative rate, calculated only using events *prior* to the current offer,

directly addresses driver self-selection bias without data leakage.

Performance Impact

The implementation of this feature significantly improved the model's predictive power.

Metric	Baseline (Estimated)	Final Performance	Improvement
ROC AUC			Significant Gain

The final ROC AUC score of confirms the efficacy of the new feature in ranking drivers based on their likelihood of acceptance.

Key Insights

Based on the final model and features used, the prediction of driver acceptance is dominated by two factors:

1. **Geographical Proximity:** Features derived from Haversine distance (driver_distance) remain highly influential, as shorter initial distances directly correlate with higher acceptance probability.
2. **Driver Behavior:** The **Historical Acceptance Rate** is a strong predictor, suggesting that drivers exhibit consistent, predictable behavior. A driver with a high past acceptance rate is likely to accept the current offer, regardless of other situational factors.