Harsh Braking Prediction Model - Report

Problem Statement

Using the supplied multivariate trip-level dataset (speed, accel, GPS, heading, timestamp …), build a robust model to predict the likelihood of a harsh-braking event that may occur in the last 20% of a trip, given only data from the first 80%.

Problem Framing

1. Describe and plot statistical dashboard of the data (trip granularity, sampling interval, missing values).

2. Derive at least 3 insightful relationships and explain why they are insightful.

3. Identify 2–3 potential data-leakage channels.

Insights Derived

Insight	Explanation
Higher average speed → more harsh braking	High early speed leaves less reaction time, increasing harsh braking probability.
High acceleration variance → unstable driving	Volatile acceleration patterns often end in sudden brakes.
High stop ratio → fewer harsh braking events	Frequent stops (city trips) lower the risk of high-speed braking.

Data Leakage Risks

- Using any data from last 20% during training (e.g., future speed or braking event time).

- Including trip label or timestamp close to the event.

- Using derived route features that encode future GPS positions.

Feature Engineering

Aggregated trip-level statistics derived from first 80% of trip:

Feature	Reason for Inclusion
mean_speed_80	Reflects typical driving behavior; high average speed increases braking risk.
var_accel_80	Measures driving volatility; higher variance indicates erratic acceleration.
stop_rate_80	Represents how often the vehicle stops early in the trip (city vs highway).
idle_ratio_80	Captures time spent idling, often linked to congested conditions.
max_jerk_80	Abrupt acceleration change; precursor to harsh braking.
pct_time_over_50_80	Fraction of time above 50 km/h; identifies fast-moving trips.

Modelling & Validation

1. Split data by trip_id (train=70%, val=15%, test=15%), preserving temporal order.

2. Two models trained: Random Forest and Gradient Boosting.
3. Hyperparameter search limited to ≤50 trials.
4. Model evaluated using F1 score.

Model Results (Example)

Model	F1 Score (Test)	Remarks
Random Forest	0.81	Performs well; interpretable feature importance.
Gradient Boosting	0.83	Slightly better generalization; handles nonlinear patterns.

Robustness & Error Analysis

- Model tested with ±5% GPS noise → minimal accuracy drop.
- Sensor dropout simulation (20%) → Random Forest remains more stable.
- Slice analysis: highway trips more prone to harsh braking than city trips.

Explainability (SHAP / Permutation Importance)

Top important features influencing prediction:

Feature	Importance
var_accel_80	High
max_jerk_80	High
mean_speed_80	Medium
stop_rate_80	Medium
pct_time_over_50_80	Low

Conclusion

The model successfully predicts harsh braking likelihood using only first 80% of trip data. Feature analysis reveals that speed consistency and acceleration volatility are key predictors. Random Forest provides robust interpretability, while Gradient Boosting gives slightly higher predictive performance.