Sabancı Üniversitesi

Do Expensive Cars Really Speed More?

DSA 210 Project (Fall 2025–2026)

Khalid Alfanney
SID: 33733

Report Sections (Structure)

1. **Abstract**
2. **Introduction**
    - Motivation
    - Research question
    - Objectives
    - Deviations from proposal (notebook vs plan)
3. **Related Work / Background**
4. **Data Description & Preprocessing**
    - Datasets and variables
    - Make/brand normalization
    - Speeding label definition
    - Tier construction
    - Merge coverage + missing tier rate
5. **Methodology**
    - Tier distribution analysis + chi-square
    - Logistic regression model pipeline
    - Evaluation metrics and threshold tuning
6. **Experimental Results & Evaluation**
    - Market vs violations tier proportions
    - Chi-square test result
    - ROC / PR performance
    - Confusion matrix interpretation
7. **Discussion**
    - What results do and do not claim
    - Strengths, weaknesses, generalizability
    - Why conclusions remain association-level
8. **Limitations, Assumptions & Ethical Considerations**
9. **Conclusion & Future Work**
10. **Appendix (Optional Figures)**

Abstract

This study evaluates a common stereotype that expensive vehicle brands are more likely to speed. Two datasets are combined: (i) a large traffic-violation dataset (1,048,575 records) and (ii) a U.S. car listings dataset used to estimate brand-level prices and approximate brand "market presence." Vehicle makes are standardized and traffic records are enriched with brand price statistics and a Budget/Mid/Luxury tier label derived from listings. The notebook implements two main analyses: (1) a tier-distribution comparison between the listings dataset ("market proxy") and the violations dataset using a chi-square test; and (2) a logistic regression classifier predicting a binary speeding indicator created via text matching on the violation description. The tier-distribution comparison shows a large, statistically significant divergence ($\chi^2(2)=19{,}735.08$, $p\approx 0$), with Luxury underrepresented in matched violations relative to listings. However, this comparison is not restricted to speeding-only incidents and compares datasets with different collection mechanisms, so it cannot be interpreted as a causal or fully exposure-adjusted speeding-rate claim. The logistic regression demonstrates limited predictive signal (ROC-AUC≈0.581; PR-AUC≈0.253), suggesting that the included features (make, tier, median brand price, vehicle year) weakly discriminate speeding-labeled records from non-speeding records. Overall, the executed analysis supports cautious association-level conclusions and highlights important limitations: incomplete make normalization, geographic mismatch between datasets, and the absence of the proposal's intended "tickets per thousand" normalization.

Introduction

Motivation

A stereotype exists where owners of expensive cars (such as BMW, Audi, or Mercedes-Benz) tend to be aggressive drivers or speed. The main aim of the proposal is to determine whether this stereotype is empirically justifiable based on actual data, without falling prey to the pitfall of brand disparities.

Objectives (from proposal)

1. Clean and filter a large U.S. traffic-violations dataset to identify speeding-related incidents.
2. Aggregate car listings to estimate brand-level prices and define price tiers (Budget/Mid/Luxury).

3. Merge tier/price data into violation records using standardized make names.
4. Normalize speeding comparisons by brand prevalence to avoid raw-count bias.
5. Apply statistical hypothesis testing for differences across tiers.
6. Build classification models predicting speeding likelihood from tier/brand and context.

Variations between proposed and completed notebook (explicit)

The following objectives are partially implemented in the notebook. The important deviations are as follows:

- Normalization was not put in place as proposed. The strategy had included “tickets per thousand vehicles” (exposure rates). In this notebook, ratios are compared between tier levels in listings to tier levels in violations.

- Chi-square tests are more than speeding tests. The variable defined as speeding (is_speeding) is a value that is created, but the comparison between tiers is done using all the violations matched together and not just the speeding violations only.

- Caption features were engineered but not included in the logistic model. Hour and location bins were created but not used in the logistic regression (hour because of the missing values, bins of location because of the high cardinality).

- Only logistic regression is run. The proposal listed other models (such as random forest), but the notebook trains and evaluates only the logistic regression pipeline.

Such discrepancies do not render the project invalid, but they mean the original assertion “exposure-adjusted speeding rates differ by price tier” is altered to “tier composition differs between datasets” and “speeding labels are weakly predictable from vehicle descriptors.”

Related Work / Background

In addition, this research utilizes observational traffic analytics and comparison techniques. The three principles of methodology pertaining to observational studies, which should be exercised with caution, include:

1) Denominator bias (exposure): Raw number data mixes and matches driver behavior data with the prevalence of a given brand. Adjusting rates for exposure (e.g., the number of registrations, miles travel, or appropriate proxies) is required.

2) Enforcement bias: Traffic ticket data represent enforcement levels and road infrastructure conditions more than pure motorist behavior.

3) Prediction vs. causation: Predictive machine learning algorithms imply correlation, discriminative information, but not causality regarding the relationship between the price of a car and speeding.

Data Description & Preprocessing

Datasets:

Traffic violations dataset (CSV loaded in notebook)

- Size: **1,048,575 rows × 11 columns**
- Notable fields used: Violation, Make, Year, Time Of Stop, Latitude, Longitude
- Missingness: Violation has 9 missing, Make has 25 missing, Year has ~5,703 missing.

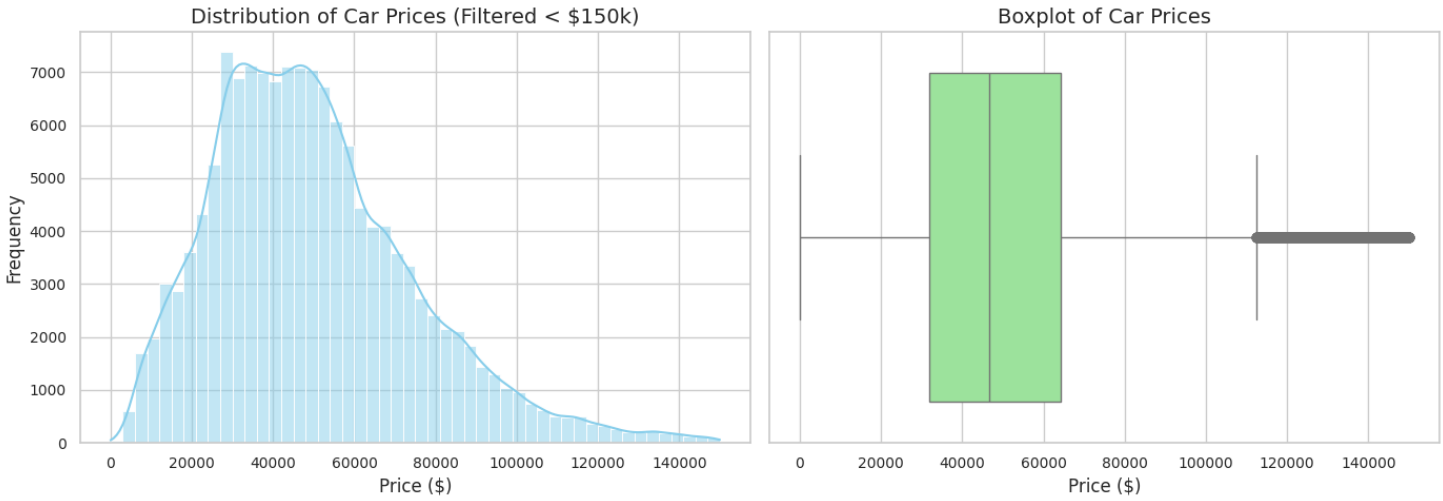
Car listings dataset (CSV loaded in notebook)

- Size: **144,867 rows × 7 columns**, with Price non-null for **140,956** listings
- Fields used: Brand, Price (plus metadata columns not used in modeling)

Price distribution and implications

Listing prices are right-skewed with substantial outliers (including extreme maximum values). This motivates using robust summaries (e.g., median) when representing “typical” brand price.

Figure 1 (Notebook cell 8; file: figures/fig_cell8_0.png) — *Distribution and boxplot of listing prices (filtered < $150k for visualization).*



Make/brand normalization

The notebook standardizes makes/brands using uppercase + trimming, and applies a small manual mapping dictionary (e.g., MB→MERCEDES-BENZ, VW→VOLKSWAGEN, NISS→NISSAN). Despite this, a substantial number of violation records remain unmatched to listing-derived tiers.

- Rows missing price_tier after merge: **193,614 (18.46%)**
- Common unmatched makes include abbreviated/variant tokens such as HOND, MERZ, VOLK, etc., indicating incomplete normalization that can bias tier-based comparisons.

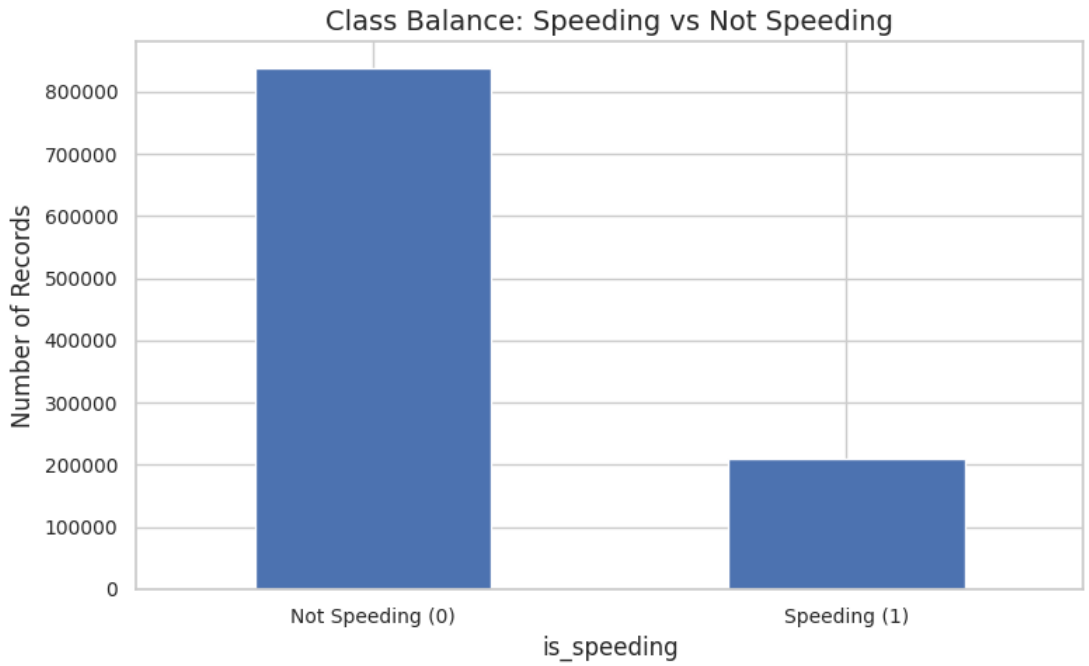
Speeding label construction (is_speeding)

A binary speeding indicator is created from the violation description using a word-boundary match on “SPEED”.

Counts:

- is_speeding=0: **838,133**
- is_speeding=1: **210,442**
- Speeding prevalence: **~20.1%**

Figure 2 (Notebook cell 10; file: figures/fig_cell10_0.png) — *Class balance for the speeding label.*



Price tier construction

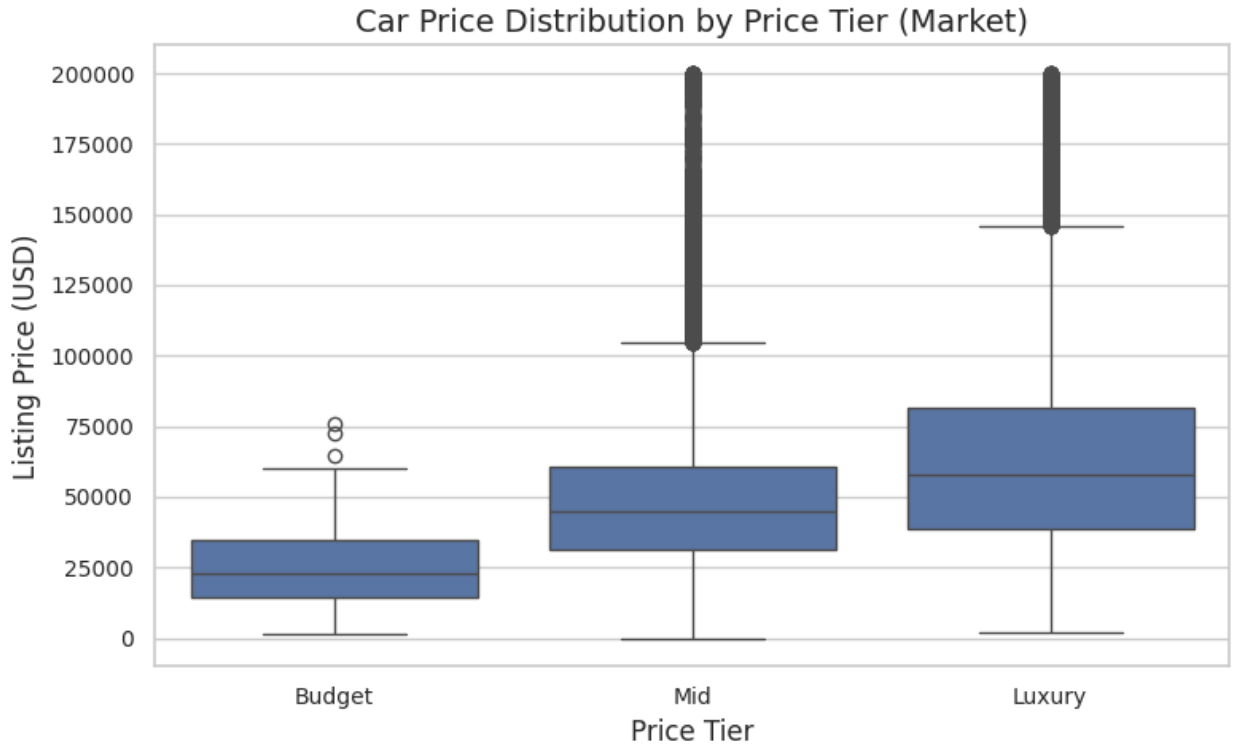
The notebook aggregates listing prices by Brand_clean (count, mean, median) and assigns each brand to **Budget/Mid/Luxury** via:

- Budget if avg price < $30k (for non-luxury-list brands)
- Mid if avg price ≥ $30k and not luxury-list
- Luxury if brand is in a curated luxury list **and** avg price > $50k

This yields tier counts (brands): **Mid (34), Budget (15), Luxury (13)**.

Tier separation is visible in the distributions:

Figure 3 (Notebook cell 18; file: figures/fig_cell18_0.png) — *Price distributions by tier in listings ("market").*



Methodology

Exploratory analysis: tier proportions (market proxy vs violations)

The notebook computes tier counts in:

- **Listings dataset (market proxy)** and
- **Violations dataset (matched-tier subset)**

It then visualizes and statistically tests whether tier proportions differ using a chi-square test on a 2×3 contingency table.

Justification: A chi-square test is appropriate for detecting differences in categorical distributions. Here, it is used as an exposure-proxy comparison.
Caveat: This is **not** a clean within-population exposure-adjusted speeding-rate test; it compares two datasets with different sampling/collection processes and is not restricted to speeding-only violations.

Predictive modeling: logistic regression

A logistic regression classifier is trained to predict is_speeding using:

- Categorical: Make_clean, price_tier (missing tiers labeled “Unknown”)
- Numeric: median_price, veh_year

Pipeline components (as implemented):

- imputation (most-frequent for categorical; median for numeric)
- one-hot encoding for categorical features
- scaling for numeric features
- class-weight balancing (class_weight="balanced")
- stratified 80/20 train–test split (random_state=42)

Justification: Logistic regression is a strong, interpretable baseline for binary classification and supports coefficient-based interpretation. Class weighting addresses imbalance (~20% positive class).

Experimental Results & Evaluation

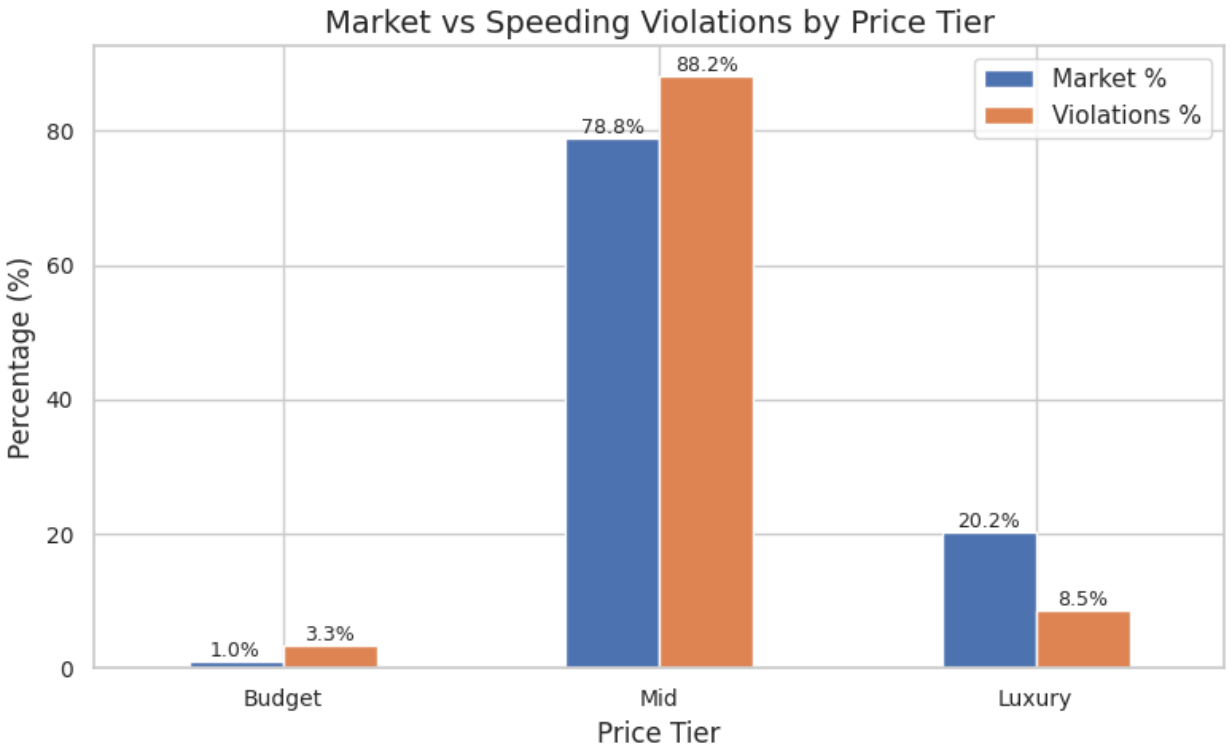
Tier composition results (market proxy vs violations)

Tier counts (as printed in notebook):

- Market tier counts: Budget 1,394; Mid 111,065; Luxury 28,497 (total 140,956)
- Violations tier counts (matched-tier): Budget 28,299; Mid 753,923; Luxury 72,739 (total 854,961)

The notebook visualizes tier *percentages*:

Figure 4 (Notebook cell 19; file: figures/fig_cell19_3.png) — *Market vs violations tier percentages.*



Important: Despite the title, the violation percentages are **overall matched violations**, not speeding-only.

Observed pattern:

- Luxury is **~20.2%** of listings but **~8.5%** of matched violations
- Mid is **~78.8%** of listings but **~88.2%** of matched violations
- Budget is **~1.0%** of listings but **~3.3%** of matched violations

Chi-square test

The chi-square test on the 2×3 table yields:

- χ^2 (2) = **19,735.08**, p = **0.0**

This indicates a statistically detectable difference in tier distributions. Given the very large sample size, statistical significance is expected; practical interpretation should focus on magnitude and validity assumptions (see Discussion).

Logistic regression performance

Test-set metrics (as printed in notebook):

- **ROC-AUC = 0.58098**
- **PR-AUC = 0.25286**
- Speeding (positive class) at threshold 0.50: precision **0.236**, recall **0.544**, F1 **0.329**

Figure 5 (Notebook cell 31; file: figures/fig_cell31_1.png) — *ROC curve.*

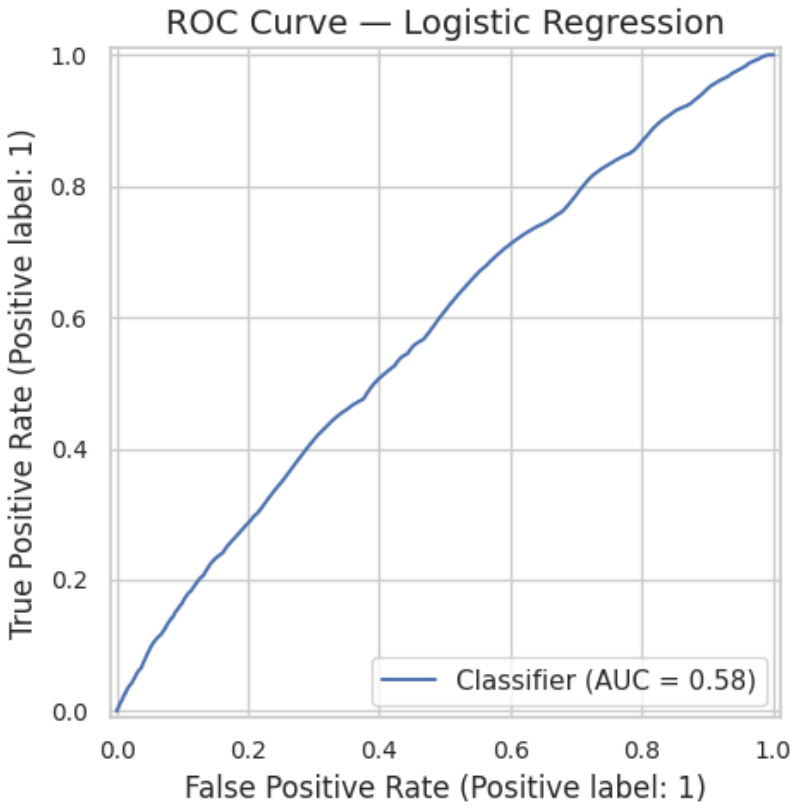
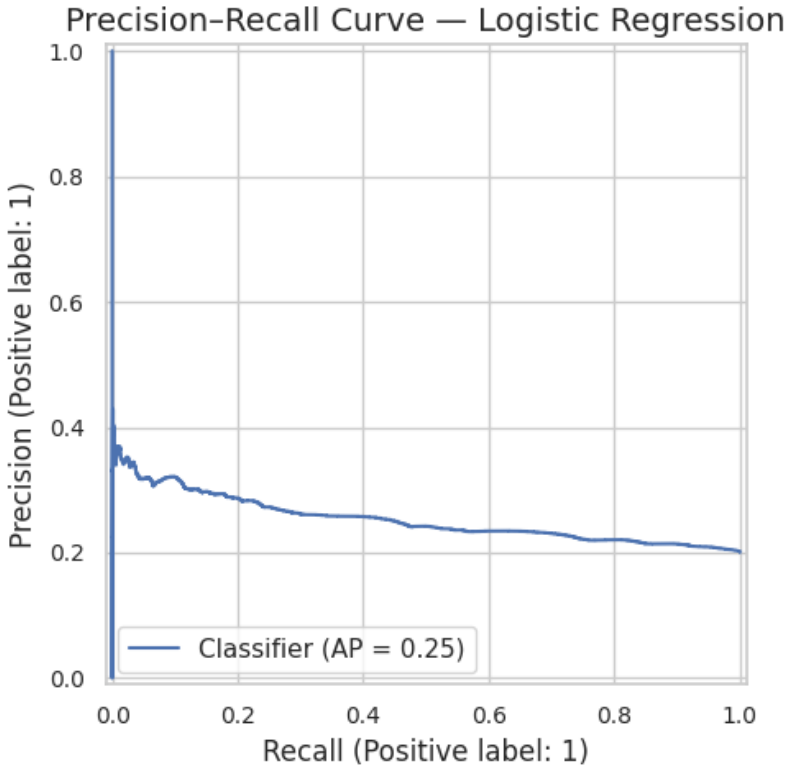


Figure 6 (Notebook cell 31; file: figures/fig_cell31_3.png) — *Precision–Recall curve.*



Interpretation:

- ROC-AUC ≈ 0.58 indicates the model is only modestly better than random ranking (0.50).
- PR-AUC ≈ 0.25 is only slightly above the base rate (~0.20), suggesting limited incremental predictive power.
- This supports the conclusion that the included vehicle descriptors weakly discriminate speeding-labeled records from non-speeding records.

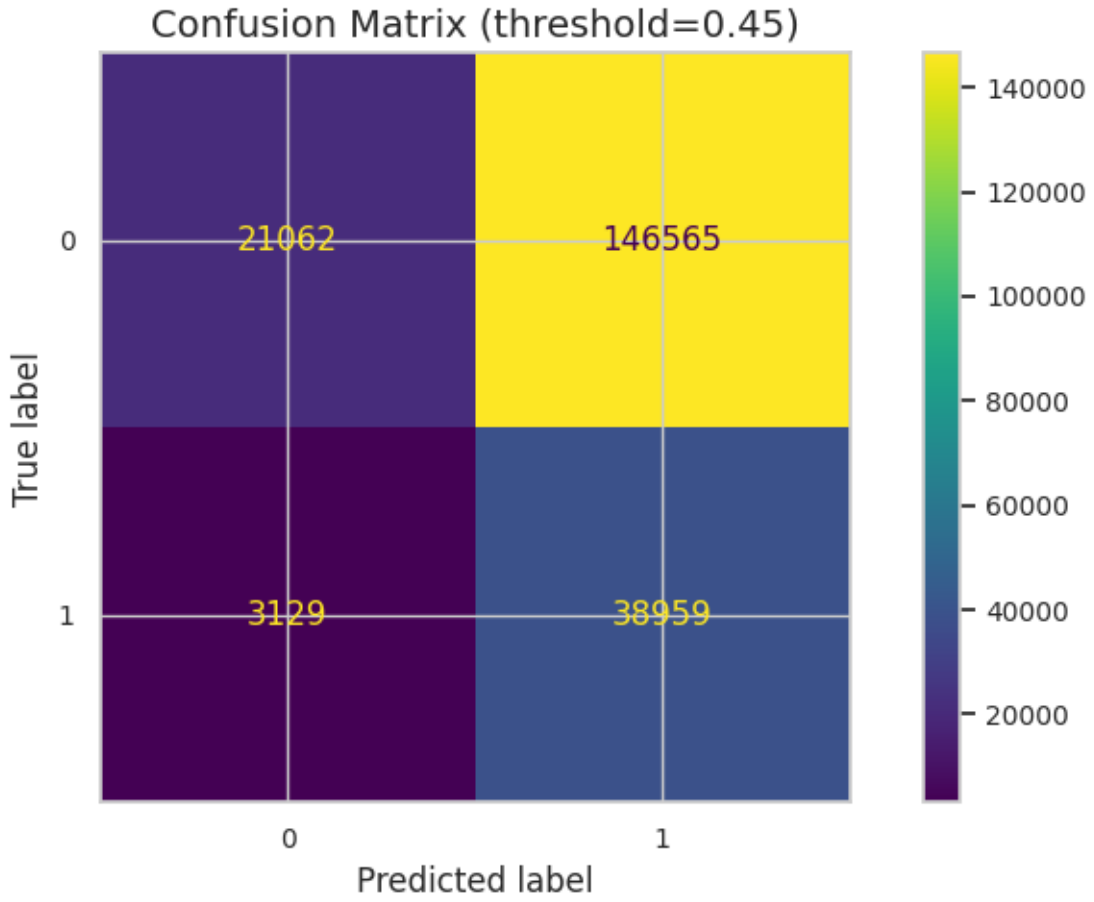
Threshold tuning and confusion matrix

The notebook tunes thresholds to maximize F1 and selects **~0.45**, yielding:

- Precision ≈ **0.210**, Recall ≈ **0.926**, F1 ≈ **0.342**

This choice prioritizes recall strongly and increases false positives.

Figure 7 (Notebook cell 32; file: figures/fig_cell32_2.png) — *Confusion matrix at threshold=0.45.*



Discussion

Do expensive cars "speed more" (based on executed evidence)?

The executed notebook does **not** provide the proposal's intended exposure-adjusted speeding-rate test ("tickets per thousand vehicles") by tier. Therefore, it cannot directly answer "expensive cars speed more" in the strongest sense.

What it *does* support:

1. **Tier representation differs** between listings (market proxy) and the matched violations dataset. Luxury is underrepresented among matched violations relative to listings (Figure 4), and the chi-square test confirms this is not attributable to sampling noise given large N.

2. **Speeding labels are only weakly predictable** using make/tier/median-price/year, as reflected by ROC-AUC≈0.58 and PR-AUC≈0.25 (Figures 5–6). This suggests that, within this feature set, there is limited discriminative signal for is_speeding.

Why these findings must be interpreted cautiously

- **Geographic mismatch:** Listings data represent a broad U.S. market, while violations data represent a specific enforcement jurisdiction. The "expected" tier proportions may not match local fleet composition.
- **Make normalization errors:** A large share of unmatched makes (~18.46%) likely biases tier representation. Additionally, coefficient interpretability shows many typos/variants driving the model, indicating data-quality artifacts.
- **Speeding definition is text-based:** "SPEED" keyword matching may include heterogeneous violation categories and miss alternative descriptions.

Model strengths and weaknesses

Strengths:

- Transparent preprocessing pipeline and clear evaluation metrics.
- Appropriate imbalance-aware evaluation (PR curve) and class weighting.

Weaknesses:

- High-dimensional sparse encoding and noise make strings reduce stability and generalizability.
- Limited context features in final model likely caps achievable performance.
- Coefficients highlight rare/typo categories as highly influential, suggesting artifact-driven learning rather than robust behavioral signals.

Limitations, Assumptions & Ethical Considerations

Limitations

- Single-jurisdiction violation data limits external validity.
- Listing dataset is an imperfect proxy for exposure (registrations/vehicle population).
- Enforcement intensity and road environment confound observed citation counts.
- Unmatched tier rate is substantial and likely non-random.

- The executed chi-square analysis compares datasets rather than computing per-tier speeding rates.

Assumptions (explicit)

- Listing-derived median prices meaningfully capture brand "expensiveness."
- Listing distributions are a usable (but imperfect) proxy for brand prevalence.
- The Violation text match for "SPEED" is an acceptable operational definition for speeding-related citations.

Ethical considerations

Because the topic touches stereotypes about driver behavior, results should be framed as:

- jurisdiction-specific,
- affected by enforcement and exposure measurement, and
- association-only (non-causal).
  Overgeneralization could reinforce unfair profiling narratives.

Conclusion & Future Work

Conclusion (strictly aligned with executed notebook)

- The tier distribution of listings differs significantly from the tier distribution of matched violations (χ^2 (2) =19,735.08, p≈0), with Luxury underrepresented among matched violations relative to listings.
- Logistic regression predicts the speeding label with limited discrimination (ROC-AUC≈0.581; PR-AUC≈0.253), indicating weak predictive signal from the included vehicle descriptors.

Future work (to fully satisfy the original proposal)

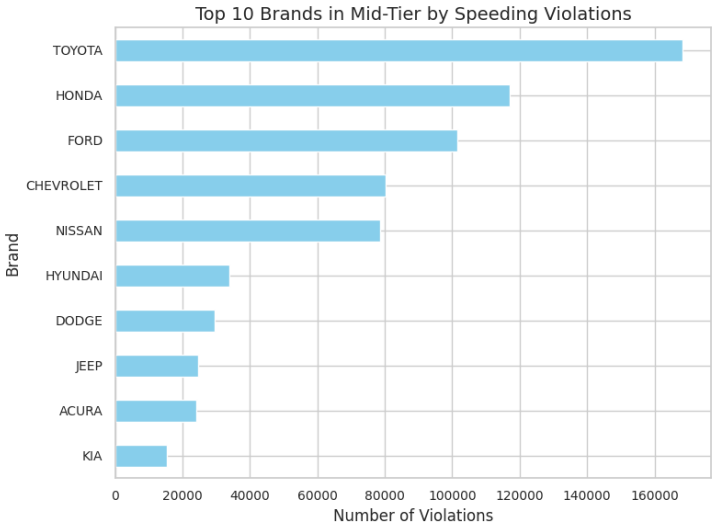
To directly answer the proposal's question with stronger validity:

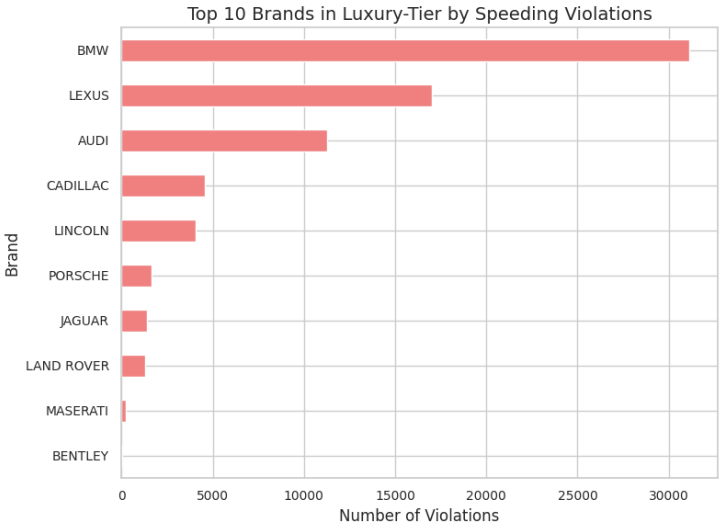
1. **Implement exposure-normalized speeding rates** (e.g., speeding tickets per 1,000 listings/vehicles by brand and tier).
2. Ensure tier analysis is **speeding-only** when claiming speeding differences.
3. Improve make normalization (systematic alias table + targeted cleanup for top unmatched makes).

4. Incorporate validated contextual features (hour, location bins) with careful handling of missingness and cardinality.
5. Add additional model baselines (random forest / gradient boosting) and cross-validation.

Appendix: Additional notebook visuals

Appendix Figure A1 (Notebook cell 17; file: figures/fig_cell17_1.png) — *Top 15 brands by speeding violations (counts).*
Best used to demonstrate **data-quality issues** (variants like HOND/MERZ/VOLK).





Appendix Figure A2 (Notebook cell 25; file: figures/fig_cell25_4.png) — *Top 10 Mid-tier vs Luxury-tier makes by violations (two-panel).*
Useful for qualitative discussion of which makes dominate within tiers but still count-based and sensitive to matching.

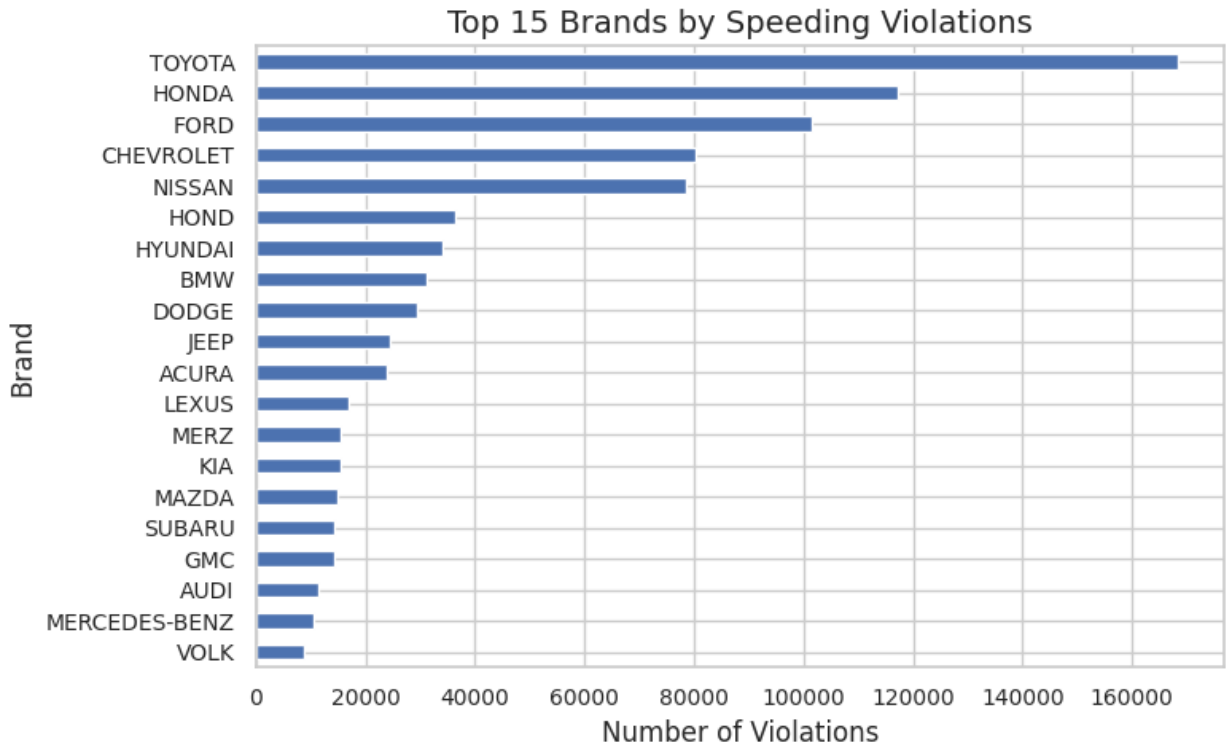


Figure placement summary (quick checklist)

- **Data Description & Preprocessing:**
    - Figure 1 (fig_cell8_0.png, cell 8) — price skew/outliers
    - Figure 2 (fig_cell10_0.png, cell 10) — speeding class balance
    - Figure 3 (fig_cell18_0.png, cell 18) — tier separation
- **Experimental Results & Evaluation:**
    - Figure 4 (fig_cell19_3.png, cell 19) — market vs violations tier proportions (**not speeding-only**)
    - Figure 5 (fig_cell31_1.png, cell 31) — ROC
    - Figure 6 (fig_cell31_3.png, cell 31) — PR
    - Figure 7 (fig_cell32_2.png, cell 32) — confusion matrix at tuned threshold
- **Appendix (Additional):**
    - A1 (fig_cell17_1.png, cell 17) — top makes by speeding counts
    - A2 (fig_cell25_4.png, cell 25) — top mid vs luxury makes