

DSBA 6156 Project Proposal

Group 8: Jason Ackerman, Matthew Lawson, Davis Martin

Project Objectives

This project aims to use the NHTSA ODI consumer complaints to develop a method for identifying increasing safety concerns, prioritizing urgent complaints for review, and turning identified patterns into clear recommendations for investigation. We will focus on delivering real, practical value by combining multiple approaches rather than relying on a single model.

Dataset

We will use the ODI complaints dataset, which includes vehicle identifiers (make/model/year), component fields, incident/severity flags, dates, geography, and complaint descriptions. The full dataset totals over 2 million records and 49 features, with optional 5-year segments. To keep the data manageable and to avoid leakage, we will focus on the most recent complete segment (2020-2024) and reserve the current 2025-2026 data as a forward test set.

Proposed Methods

We will begin by cleaning and standardizing key predictors, removing duplicates, handling missing value and data consistency issues, and preparing descriptions for analysis. Feature engineering will be used to create two groups of new features: complaint predictors (vehicle attributes, component, time, location, description signals) and cohort predictors for specific make/model/year groups (number of complaints, changes over time, severity, component makeup). After this preprocessing, we will begin building our models. First, a severity ranking model will classify complaints as high or low risk to create a priority score to highlight urgent complaints. Second, a component model will predict the most likely component category causing issues, supporting faster assignment and trend tracking. Third, an early warning system will flag cohort and component groups that show unusual spikes in complaints and create a monthly watchlist ranked by volume and complaint duration. To support this, we will use unsupervised NLP to group complaint descriptions into topics and track which ones grow fastest over time and within which cohorts. To keep results realistic and unbiased, models will be tested on later time periods they haven't seen, performance metrics will be chosen based on class balance, and our error analysis will focus on whether the most serious complaints are flagged and whether complaints are assigned to the correct component category.

We also have the option to join the NHTSA recall notices dataset to strengthen business relevance and provide a source for validation. Recall timing and scope can be used to validate whether our modeled early warning signals and identified topics precede known recalls. It can also be used to aid forecasting in estimating the likelihood of future recalls of our cohorts.

Deliverables

Deliverables will include:

- a severity scoring method to rank complaints by urgency
- a component assignment model that adds complaints to the most likely issue category
- an early warning watchlist highlighting make/model/year and component groups with unusual spikes in complaints
- a defect topic library that organizes complaints into recurring issue types and tracks growth
- a final recommendations section that translates our findings into actionable investigation priorities, specific manufacturer/supplier actions, and ongoing monitoring.