Analyzing NHTSA Customer Complaint Trends

December 12th, 2024

**Final Report: NHTSA Customer Complaints Data Analysis**

**Section 1:**

**Summary**

The dataset analyzed comprises over three decades (1990–2024) of customer complaints submitted to the National Highway Traffic Safety Administration (NHTSA). These complaints offer a detailed view of the issues affecting vehicle safety and reliability across manufacturers, states, and time. The project focuses on identifying patterns in complaints, investigating defect severity, and using advanced analytics to provide actionable insights for stakeholders, including manufacturers, policymakers, and consumers. With a focus on trends over time, geographic patterns, defect severity, time-to-failure, and sentiment analysis, the findings shed light on the critical areas needing attention. The analysis also leverages predictive modeling techniques to assess the likelihood of crash-related outcomes, offering a forward-looking perspective.

**Analysis Conducted:**

The analysis undertaken for this project covered a comprehensive range of investigative approaches to gain insights into the NHTSA customer complaints dataset. The analysis began with a trendanalysis to evaluate the yearly distribution of complaints and identify notable changes or anomalies over time. A geographic analysis was conducted to pinpoint complaint hotspots across U.S. states, revealing patterns in complaint density and frequency. Next, a manufacturer-level analysis compared the complaint counts and high-risk defect rates for major manufacturers, with a specific focus on their most problematic vehicle makes. Additionally, the relationship between vehicle age and complaints was analyzed, shedding light on how defect patterns change as vehicles age.

A critical part of the analysis involved categorizing complaints by defect severity into high, moderate, low, and unknown risk levels, coupled with sentiment analysis to assess customer perception based on complaint descriptions. To understand the lifecycle of defects, a time-to-failure analysis was conducted, exploring the time taken to report defects after vehicle purchase. Advanced predictive modeling was employed using Logistic Regression, Random Forest, and XGBoost to predict crash likelihood based on variables like defect severity, sentiment severity, and other vehicle attributes. Each model’s performance was evaluated using metrics such as accuracy, recall, F1-score, and area under the ROC curve (AUC), enabling a robust comparison of their predictive capabilities.

This multi-faceted analysis was supported by R-based visualizations, including trend lines, heatmaps, bar charts, histograms, and ROC curves, to effectively communicate the findings and their implications. Each step in the analysis contributed to a deeper understanding of the data, leading to actionable insights for improving vehicle safety and reliability.

**Key Findings:**

1. The number of complaints peaked in the early 2000s, followed by a gradual decline, potentially due to improved vehicle safety standards or changes in reporting behavior.
2. States like California, Texas, and Florida reported the highest complaint counts, indicating regional differences based on factors like population density and vehicle usage.
3. General Motors, Ford, and Chrysler had the highest complaint counts among manufacturers, with Ford leading in specific vehicle makes like the Ford F-Series.
4. High-severity defects such as engine and brake failures dominated the dataset, representing critical safety concerns.
5. Moderate and low-severity defects, including suspension and lighting issues, were also substantial but less frequent.
6. Sentiment analysis showed most complaints were described with a "Moderately Negative" tone, even for high-severity issues, indicating that complaint tone does not always reflect the defect’s severity.
7. High-severity defects often occurred early in a vehicle’s lifecycle, highlighting potential quality control problems.
8. Manufacturers like Ford and GM exhibited longer average defect reporting periods, while others showed quicker defect identification.
9. Random Forest was the best-performing predictive model with an accuracy of 81.45% and an F1-score of 84.35%, while Logistic Regression achieved high recall (92.31%), making it effective at identifying true positives.
10. Complaints were most frequent for vehicles aged 1–10 years, with fewer complaints for older vehicles.
11. General Motors had the highest percentage of high-risk defects, followed by Ford and Chrysler, indicating areas for quality improvement.
12. The top vehicle makes with the most complaints included the Ford F-Series, Chevrolet Silverado, and Dodge Ram.

**Summary of Findings:**

The analysis of the NHTSA customer complaints dataset reveals critical insights into patterns and trends related to vehicle safety and reliability. High-severity defects, such as engine and brake failures, dominate the dataset and often occur early in a vehicle’s lifecycle, signaling potential quality control issues. Manufacturers like General Motors, Ford, and Chrysler not only report the highest number of complaints but also exhibit a significant proportion of high-risk defects, indicating areas requiring immediate attention. While states like California, Texas, and Florida show the highest complaint volumes, sentiment analysis highlights that customer tone does not always align with the severity of defects, as even critical issues are often reported with a "Moderately Negative" tone. Predictive modeling efforts show promising results, with Random Forest emerging as the best-performing model for predicting crash likelihood, achieving an F1-score of 84.35%. These findings collectively emphasize the need for targeted interventions by manufacturers and policymakers to enhance vehicle safety and improve customer satisfaction.

**Other Analysis:**

**Manufacturer-Specific Complaint Analysis**: Analysts have categorized complaints data to evaluate the performance of different automakers. By weighting the number of complaints based on total vehicle sales, they calculated metrics such as "Complaints per 100,000 Vehicles Sold." This approach allows for a relative comparison among manufacturers, identifying brands with disproportionately high or low complaint rates.

<https://www.edmunds.com/industry-center/data/nhtsa-complaints-activity-report.html>

**Text Mining for Trend Analysis**: Researchers have applied text mining techniques to the NHTSA vehicle complaint database to identify latent trends in consumer complaints. By analyzing free-response complaint narratives, they extracted patterns related to vehicle defects and associated safety implications, providing a deeper understanding of common issues reported by consumers.

<https://www.academia.edu/74514935/Vehicle_Consumer_Complaint_Reports_Involving_Severe_Incidents_Mining_Large_Contingency_Tables>

**Comparison to Other Analysis:**

When comparing our project to other analyses of the NHTSA complaints dataset, several distinctions emerge in terms of scope and depth. The Manufacturer-Specific Complaint Analysis highlighted in external research uses a weighted metric like "Complaints per 100,000 Vehicles Sold" to normalize complaint counts relative to total vehicle sales. While this approach offers a fair comparison among manufacturers, our project extends beyond mere complaint counts by incorporating high-risk defect percentages and predictive modeling to assess crash likelihood. This additional layer of analysis provides actionable insights into safety-critical issues, enabling manufacturers to prioritize defect resolution based on severity rather than frequency alone.

Similarly, Text Mining for Trend Analysis employs advanced text analysis techniques to extract latent patterns from free-response complaint narratives. While our project includes sentiment analysis of complaint narratives, the focus was on classifying sentiment severity and correlating it with defect severity rather than uncovering hidden trends. Our approach bridges subjective customer feedback with objective defect categorization, offering a unique perspective that complements the text mining studies. By combining sentiment scores with structured defect categories and incorporating time-to-failure analysis, our project provides a more integrated understanding of customer experiences and safety implications.

In summary, while other analyses emphasize normalized complaint metrics and latent trends, our project advances the field by blending defect severity, sentiment analysis, and predictive modeling to uncover practical and safety-oriented insights, making it highly relevant for manufacturers and policymakers.

**Section 2: Implications**

The comprehensive analysis of the NHTSA customer complaints dataset yields significant implications for manufacturers, policymakers, and consumers. Manufacturers must prioritize addressing high-severity defects, such as engine and brake failures, which dominate the dataset and often occur early in the vehicle lifecycle. These findings point to potential quality control gaps that require immediate intervention, particularly for major manufacturers like General Motors, Ford, and Chrysler, who exhibit higher rates of high-risk defects. Addressing these defects proactively can reduce safety risks, improve brand reputation, and enhance customer satisfaction.

The geographic analysis reveals that states like California, Texas, and Florida experience the highest complaint volumes, suggesting regional patterns that could guide targeted safety campaigns and regulatory focus. Policymakers can leverage these insights to allocate resources more effectively and enforce stricter safety standards in areas with high complaint densities. The time-to-failure analysis further underscores the need for early defect detection and resolution, as many severe issues surface within the first few years of vehicle ownership.

Sentiment analysis shows that customer tone does not always align with the objective severity of complaints, highlighting the importance of integrating sentiment insights with structured defect data for a more holistic understanding. This insight enables manufacturers to improve customer service processes by addressing perceived frustrations and fostering better communication with consumers.

Predictive modeling offers transformative potential by accurately predicting crash likelihood based on defect severity, sentiment scores, and other vehicle attributes. Random Forest, as the best-performing model, demonstrates the feasibility of deploying data-driven tools to prioritize safety-critical defects and allocate resources efficiently. For policymakers, these models can guide defect investigations and recall decisions, ensuring timely action on high-risk issues.

Overall, the findings emphasize the importance of a multi-faceted approach to vehicle safety, combining defect severity analysis, sentiment insights, geographic patterns, and predictive modeling. This integrated perspective empowers stakeholders to enhance vehicle safety standards, reduce crashes, and build consumer trust in the automotive industry.

**Section 3: R Code Snippets and Outputs**

Data Load and preprocessing.

**A computer screen shot of a computer error

Description automatically generated**

**A computer code with text

Description automatically generated with medium confidence**

A graph with red dots and blue lines

Description automatically generated

**A screenshot of a computer code

Description automatically generated**

A map of the united states with numbers and text

Description automatically generated

**A computer code with text

Description automatically generated with medium confidence**A graph of a number of days to report

Description automatically generatedA graph of different types of numbers

Description automatically generated

**A computer code with red blue and white text

Description automatically generated**

A graph with blue and red dots

Description automatically generated

**A screenshot of a computer code

Description automatically generated**

A graph of a number of companies

Description automatically generated

A graph of a company's high-risk defect percentage

Description automatically generated

**A computer screen shot of text

Description automatically generated**

A graph of company's company's company's company's company's company's company's company's company's company's company's company's

Description automatically generated

A graph with text overlay

Description automatically generated

**A screenshot of a computer code

Description automatically generated**

**A computer code with text

Description automatically generated with medium confidence**

**A screenshot of a computer code

Description automatically generated**

**A computer code with text

Description automatically generated with medium confidence**

A graph of a number of blue rectangular bars

Description automatically generated

A graph of blue squares

Description automatically generated

A graph of different colored squares

Description automatically generated

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a computer code

Description automatically generated**

**A computer screen shot of a program

Description automatically generated**

A graph of a logistic curve

Description automatically generated

**A close up of text

Description automatically generated**

**A computer code with numbers and symbols

Description automatically generated**