**Multivariate Insights into Vehicle Attributes and Warranty Claims**

Group Name: Data Hunters

Sirisha Poolakunta Ramesh, Sri Charan Bodduna, Keerthi Malathkar

University of North Texas

November 09, 2024

**Table of Contents**

[Executive Summary 3](#_Toc1655909541)

[Case Background 3](#_Toc149503691)

[Data Preparation 4](#_Toc876721285)

[Data Understanding 4](#_Toc1870581210)

[Modeling 6](#_Toc32329193)

[Key Performance Indicator 7](#_Toc727377070)

[Individual attributes impact on increased warranty costs 7](#_Toc29905291)

[Specific Attribute Pairs that Amplify Warranty Claims 8](#_Toc1970251251)

[Key Takeaways from Modeling 9](#_Toc1566053758)

[Business Value Creation for Peterbilt 14](#_Toc960147028)

# Executive Summary

Peterbilt Motor Company manufactures heavy-duty trucks with customizable specifications for a range of customer needs. The company aims to understand how different vehicle attributes impact warranty costs, which is crucial for optimizing product offerings and improving profitability. Using two datasets—one detailing truck attributes and their associated option codes, and the other containing warranty claims—this analysis seeks to identify which individual attributes and attribute combinations are linked to higher warranty costs. By leveraging multivariate analysis, the goal is to uncover actionable insights that can help Peterbilt make informed decisions on future vehicle configurations to reduce warranty claims and enhance product quality.

# Case Background

Understanding warranty costs is vital for Peterbilt Motor Company as these costs have significant financial and operational implications. High warranty claims can erode profit margins, especially for a company that offers a wide range of customizable truck configurations. Warranty expenses also impact customer satisfaction; frequent or costly claims may result in customer dissatisfaction, damage to the brand’s reputation, and potential loss of business. By identifying which vehicle attributes or combinations of attributes contribute to higher warranty costs, Peterbilt can make more informed decisions on vehicle design, improve product reliability, and ultimately reduce the financial burden of warranty claims, enhancing both profitability and customer loyalty.

1. Which individual attributes or option codes are significantly associated with increased warranty costs?
2. Are there specific pairs of attributes that interact in a way that exacerbates warranty claims?
3. What modeling approaches can effectively quantify these relationships and offer predictive insights for future vehicle configurations?

# Data Preparation

## Data Understanding

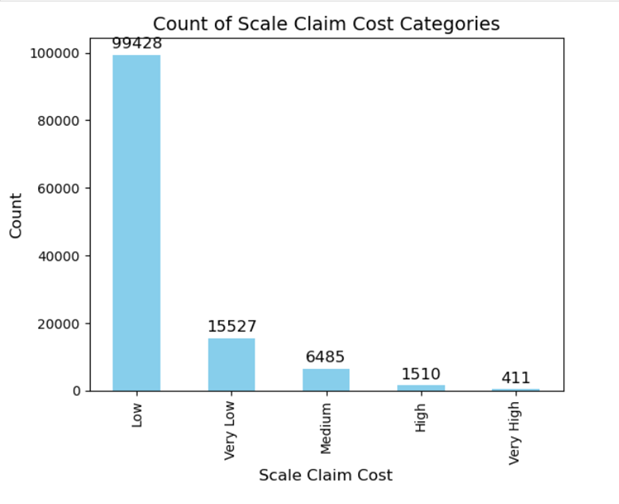
The first dataset contains eight truck attributes, each associated with an option code that represents specific configuration choices made for the trucks. These attributes include key specifications such as engine type, transmission, axle configuration, and other vehicle features tailored to meet the diverse needs of customers. The second dataset provides detailed warranty claims data for the same truck population, including the cost of claims and labor cost. The two datasets are linked through a common truck identifier, allowing us to analyze how different vehicle configurations relate to warranty costs.

In terms of data quality, we performed several important data cleaning steps to ensure consistency and accuracy. First, we merged the two datasets using a left join in Python, linking the truck attributes to their corresponding warranty claims based on the truck number. These steps were crucial in preparing the data for further analysis and ensuring that the results would be reliable and actionable.

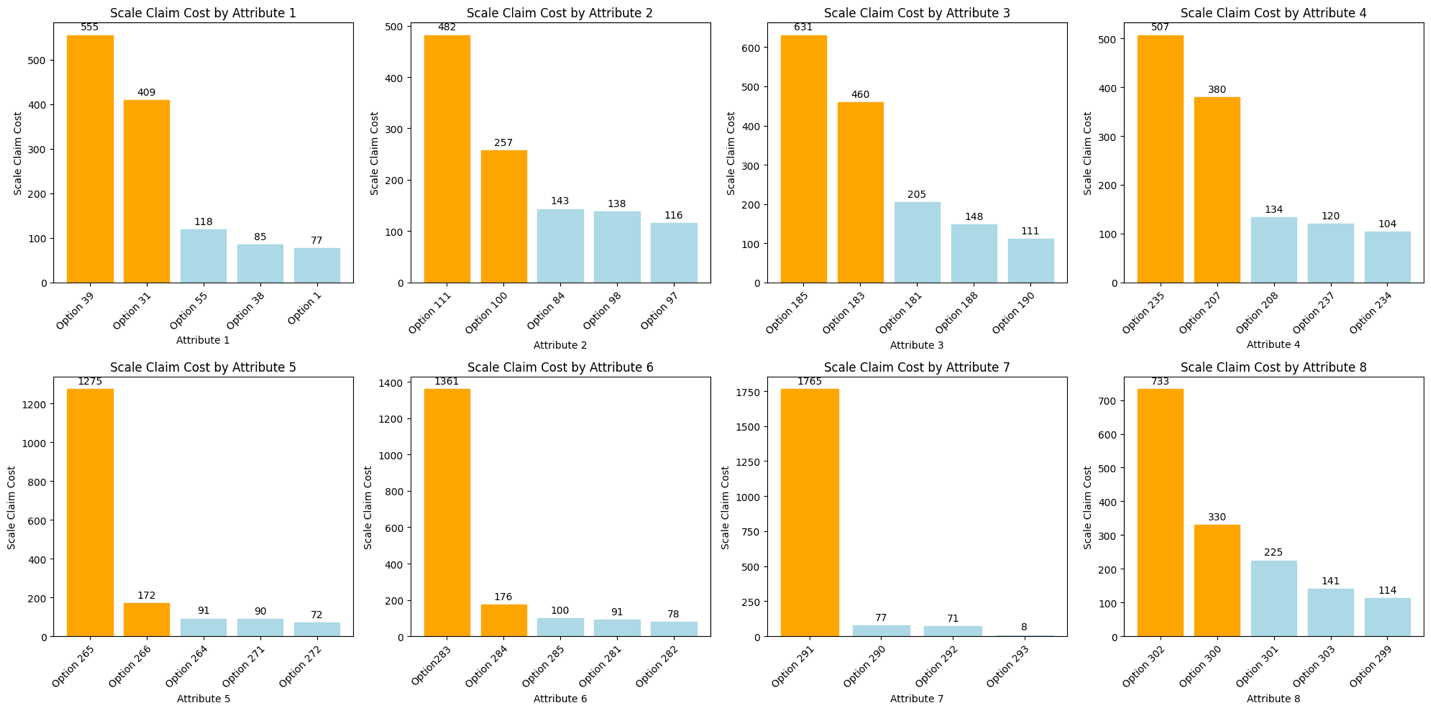
Next, we performed the binary classification on the target variable “Scale Claim Cost”. The target variable has a total of 5 distinct values such as “Very High”, “High”, “Medium”, “Low”, and “Very Low”. We have classified “High” and “Very High” as 1 and other values as 0. The final dataset consists of two categories in the target variable with High as 1 and Low as 0.   
  
This chart is a bar graph that shows the distribution of "Scale Claim Costs" across five categories:

Low: 99,428  
 Very Low: 15,527  
 Medium: 6,485  
 High: 1,510  
 Very High: 411

Interpretation:  
 Majority of Claims are Low Cost: Most scale claims fall into the "Low" cost category, with 99,428 claims. This suggests that most claims are relatively inexpensive to process or resolve. Decreasing Frequency with Higher Costs: As we move towards higher cost categories, the number of claims decreases significantly. There are considerably fewer claims in the "Very Low," "Medium," "High," and "Very High" categories compared to the "Low" category.



To understand the distribution of options for each attribute column, we have used the bar chart for each attribute. In this bar chart we have plotted a bar graph by calculating the value count based on target variable of “High”. Also applied the filter to show only top 5 options for each attribute based on value counts. From the chart below we observe that for each attribute bar 1 and bar 2 has the highest count. While “Attribute 7” has the highest count for option value 291 as first leading, “Attribute 8” with option 283 as the second leading, “Attribute 5” with option 265 as third leading, and “Attribute 2” with option 111 has the lowest count.



# Modeling

Understanding the relationships between Truck specifications and scale claim costs with the help of Machine Learning predictions gives better recommendations for the future production of Truck productions and configurations. In this project, we have utilized the data provided i.e., option code and claim costs. Our major focus in the prediction of claim costs is to classify high and low claim costs of the trucks in relation to truck’s configuration also known as Attributes in the dataset. So, we assumed that very high and high claim costs as 1 and medium, low, and very low categories as 0.

After removing the identification columns like Truck Number, Claim Number, and Style Number we found that there are large number of duplicate records for a given configuration of the truck. Moreover, we found that few trucks have multiple claim costs with different claims. For example, “Truck 1” has claim costs as low, medium, and very high with the same configurations. We have considered very high or high claim cost if there is a single record of very high or high claims for a particular truck respectively. After removing duplicates, we ended up with 6469 rows and 9 columns. Then, we performed one hot encoding on the categories present in each attribute which ramped up over input columns to 293 excluding target variables.

Subsequently, our modeling dataset consists of imbalanced target classes. with the ratio of 90% of non-high claim costs. We have introduced Synthetic Minority Over-sampling TEchnique (SMOTE) technique to equalize target classes. Then we performed training and testing subset splits in the ratio of 8:2. We used a training dataset to create models with different machine learning algorithms. As our dataset presents a huge number of categorical columns the best way to encounter this problem is to utilize ensemble techniques. These models have been evaluated using accuracy to draw better business insights and recommendations.

We have created a baseline model using decision trees which did not produce expected results hence we have utilized bagging and boosting ensemble techniques. At first, we used a random forest classifier which provided decent results compared to baseline model, we tried with boosting techniques as well. We implemented Gradient Boosting models like XGBoost, LightGBM and CatBoost which given almost similar results. These models have been hyper parameter tuned with the help of Grid Search cross validation technique.

# Key Performance Indicator

## Individual attributes impact on increased warranty costs

To identify which options within each attribute are most significantly associated with high and very high warranty claim costs, we performed a detailed analysis using the chi-square test for each option across the attributes. The results provide insights into the specific options that have a strong statistical relationship with increased warranty costs. Attributes 7, 6, and 2 have consistently shown significant associations with higher warranty costs, with extremely low p-values for certain options, indicating their strong impact. For instance, in **Attribute 1**, options such as "**Option 1**" (p-value of 8.48×10−26) and "**Option 60**" (p-value of 1.48×10−4) stand out as most strongly related to increased warranty claims, suggesting these specific configurations or selections in Attribute 1 are frequently linked to higher costs.

In **Attribute 2**, options such as "**Option 83**" (p-value of 8.48×10−26) and "**Option 158**" (p-value of 1.47×10−4) exhibit the strongest association with high claim costs. These options can be prioritized in any efforts to address cost mitigation, as they show a marked impact on warranty expenses. Similarly, in **Attribute 7**, "**Option 290**" (p-value of 8.48×10−26) is notably significant, highlighting this option as a potential contributor to elevated warranty costs. In **Attribute 6**, "**Option 281**" (p-value of 1.02×10−25) is also significantly associated with high costs, underscoring a consistent pattern across certain configurations in this attribute. **Attributes 4 and 8** also reveal noteworthy options such as "**Option 221**" (p-value of 2.44×10−22) and "**Option 294**" (p-value of 1.02×10−25), though they show slightly weaker associations compared to the previously mentioned attributes.

Each of these significant options within the attributes contributes differently to warranty claims. Attributes 1, 2, 6, and 7, with their highly significant options, likely play a central role in driving increased warranty costs, while Attributes 4 and 8, though still significant, may have a relatively smaller impact. This analysis indicates that targeted measures aimed at addressing the issues associated with specific options in **Attributes 7, 6, and 2** may be most effective for cost control. By focusing on these configurations, organizations can address root causes and potentially mitigate the warranty costs associated with high-claim products or services.

|  |  |  |
| --- | --- | --- |
| **Rank** | **Attributes** | **Options** |
| 1 | 7 | 290 |
| 2 | 6 | 281 |
| 3 | 2 | 83,158 |
| 4 | 1 | 1,60 |

## Specific Attribute Pairs that Amplify Warranty Claims

The analysis of attribute pairs through chi-square testing reveals that specific combinations exhibit statistically significant associations with elevated warranty claims. These low p-values indicate that certain attribute pairs interact in a way that increases the likelihood of higher warranty costs, which may suggest underlying compatibility issues or stressors when specific attribute configurations are present together. Attribute pairs with particularly low p-values, such as **(Attribute 6, Attribute 7) with p = 6.71 × 10⁻²⁶** and **(Attribute 2, Attribute 7) with p = 7.85 × 10⁻²⁶**, demonstrate the strongest association with high warranty claims, signaling those combinations involving Attributes 6, 7, and 2 may require further scrutiny. These attributes likely exhibit synergistic effects when paired, compounding the risks that lead to increased warranty costs.

Other pairs, such as **(Attribute 1, Attribute 7)** and **(Attribute 1, Attribute 6)**, with p-values of **3.39 × 10⁻²¹** and **7.87 × 10⁻¹⁹** respectively, also show strong associations. The significant interaction between Attribute 1 and other high-impact attributes like Attributes 6 and 7 suggests that the configurations in Attribute 1 could be amplifying the effects of known warranty risk factors within Attributes 6 and 7. Similarly, **(Attribute 3, Attribute 7)** and **(Attribute 2, Attribute 5)** with p-values in the 10⁻¹⁵ to 10⁻²¹ range, emphasize that specific pairings may be leading to compatibility issues or performance-related problems when they occur together.

Overall, the data points toward **Attributes 2, 6, and 7** as central drivers of high warranty costs when combined with other attributes, especially with Attributes 1 and 3. These pairs could be prioritized in a root cause analysis to better understand the underlying mechanisms that drive high warranty claims in these combinations.

|  |  |
| --- | --- |
| **Rank** | **Attributes** |
| 1 | 7,6 |
| 2 | 7,2 |

## Key Takeaways from Modeling

Below are the results from different models. These models have been evaluated based on accuracy, precision and recall scores.

**Accuracy** is ameasure of overall correctness of a model, indicating the ratio of correctly predicted instances to the total instances.

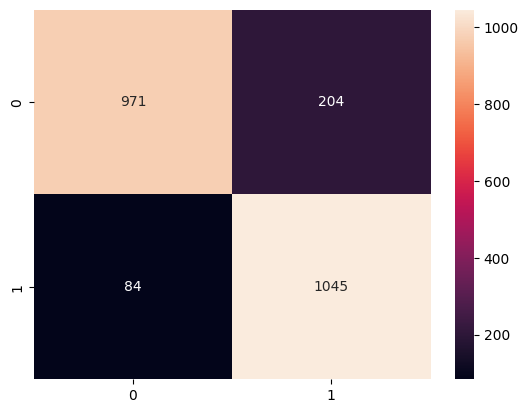
**Precision** indicates the ratio of correctly predicted positive observations to the total predicted positives, reflecting the quality of positive predictions.

**Recall** measures the ratio of correctly predicted positive observations to all actual positives, assessing the model's ability to identify all relevant instances.

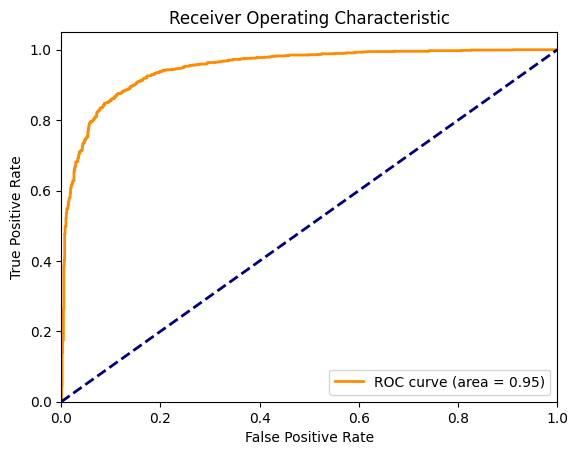
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** |
| Decision Tree | 74% | 67% | 73% |
| Random Forest | 80% | 85% | 75% |
| LightGBM | 87% | 92% | 82% |
| CatBoost | 87% | 91% | 83% |
| **XGBoost** | **88%** | **92%** | **83%** |

From the above results, we can clearly see that XGBoost model has given highest scores compared to other algorithms for the given problem.

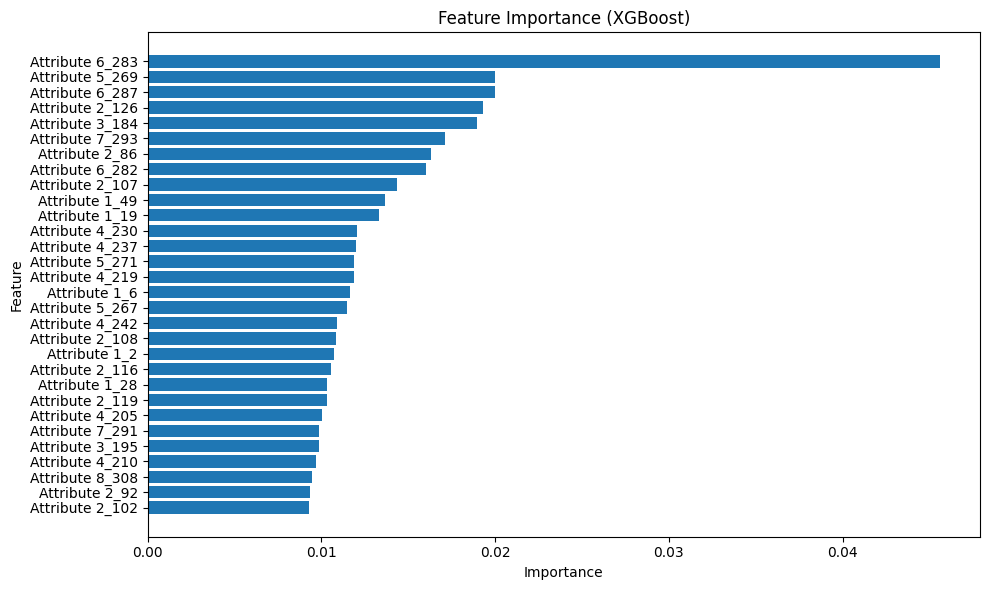
Here is the confusion matrix from the XGBoost results. With 88% accuracy, the model is able classify true values more accurately however, the False Positives are relatively higher than False Negatives.



Below image shows ROC-AUC Curve of XGBoost with Area Under Curve (AUC) value of 95%.



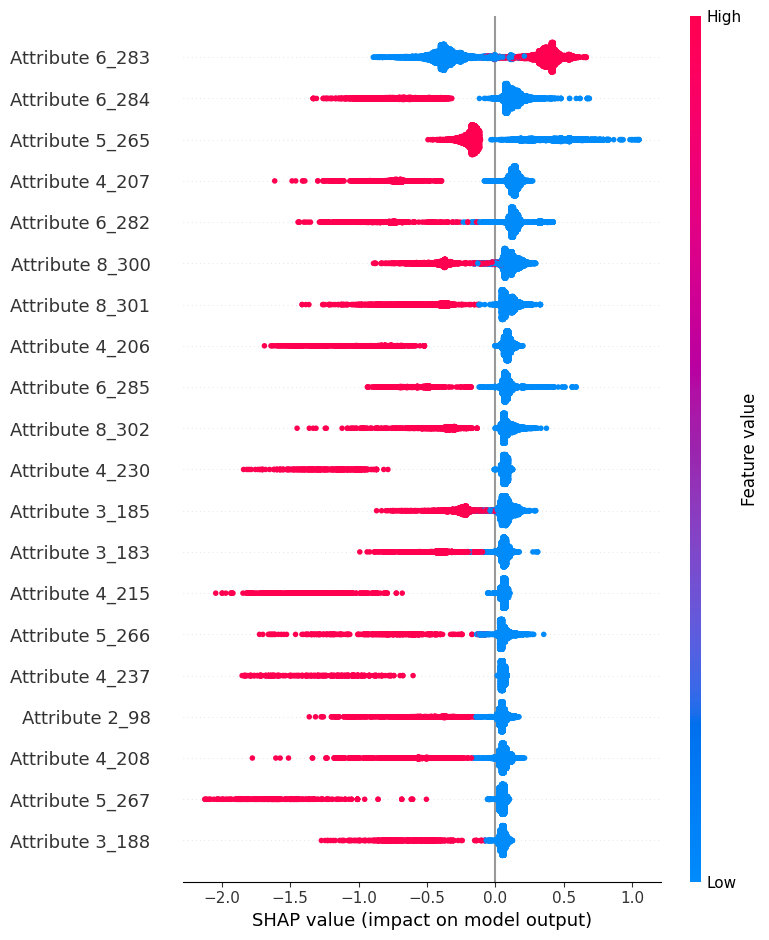
We also investigate the feature importance provided by the model showing highly important variables in the XGBoost model. It says that “Attribute 6” has a significantly highest importance in classifying claim cost either high or low.



From the above results of outperforming model i.e., XGBoost, the classification of the claim high and low are classified accurately for 88% of the times. The truck configuration options from attributes like “Attribute 6”, “Attribute 5” and “Attribute 2” are very important in the classification between high and low claims. Especially, “Option code 283” in “Attribute 6” is very crucial.

From the chart below, we can understand the relationship between claim costs and attributes. The chart shows top feature attributes with their option codes. To interpret this chart, consider red color as 1 (very high and high claim cost) and 0 (medium, low and very low claim cost).

Based on the XGBoost model, the feature “Attribute 6\_283” tells that when option code 283 is configured in the truck then the claims go high. Similarly, the feature Attribute 6\_284 tells that when there is no option code 284 the claim cost goes low.



# Business Value Creation for Peterbilt

1. **Targeted Quality Improvements:** Attributes with very low p-values (such as Attributes 6 and 7, or Attributes 2 and 7) indicate significant interactions that likely drive higher warranty costs. By focusing on these attribute pairs, the company can address specific quality issues in the design or manufacturing of these components, leading to reduced claims and lower costs.
2. **Optimized Product Configurations:** Some attribute pairs, such as (Attribute 1, Attribute 6) and (Attribute 2, Attribute 5), also show strong associations with higher claim rates. For future products, the company could limit or offer alternative configurations for these specific attribute combinations, which would reduce the likelihood of costly warranty issues in the field.
3. **Enhanced Predictive Maintenance Strategies:** For attribute pairs that have shown strong associations with higher warranty claims, predictive maintenance programs could be tailored. By identifying vehicles or customers with these specific attribute combinations, the company can proactively perform maintenance to prevent potential warranty issues, thereby improving customer satisfaction and retention.
4. **Product Development Insights:** Certain attributes and attribute combinations repeatedly show strong interactions leading to high claims. These findings suggest areas for innovation in product design, were new materials, alternative technologies, or improved engineering could be applied to mitigate these cost-driving factors.
5. **Informed Customer Recommendations:** Insights from these attribute combinations could also be shared with customers in fleet or bulk sales. The company could advise customers on product configurations that are less likely to incur high warranty costs, enhancing customer trust and helping to position the company as a proactive partner in managing long-term vehicle reliability.