STA 141C
Xiao Hui Tai

Predicting Insurance Claims Using Vehicle Data

Group Members: Cameron Zaidi, Nico, Johnson, Mariia

Introduction:

In the realm of auto insurance, accurately predicting the likelihood and cost of claims can significantly enhance operational efficiencies and customer satisfaction. Our project aims to utilize machine learning to analyze vehicle data and predict insurance claim outcomes. This endeavor will not only provide insights for insurers but also aid in risk assessment and policy pricing.

Research Questions and Objectives:

The primary goal of our project is to determine how various car-related factors influence insurance claims. We aim to address the following key questions:
- Risk Factors: What vehicle characteristics (e.g., make, model, age, mileage) most significantly predict the likelihood and magnitude of an insurance claim?
- Modeling Claim Severity: Can we predict the cost associated with a claim based on the specifics of the car and the circumstances of the incident?
- Policyholder Impact: How do modifications and usage patterns affect the probability and cost of claims?

Data Description:

Our dataset comprises records from a comprehensive database of insurance claims linked to detailed car data. This includes:
- Vehicle attributes like make, model, year, engine size, color, and safety features.
- Historical claims data detailing claim amounts, types of claims (e.g., collision, theft, liability), and outcomes.
- Policyholder data including age, driving history, and location.

Methodology:

We will employ several machine learning techniques to predict the outcomes of insurance claims:
- Random Forests: For their robustness and ability to handle complex interactions between features. We will employ the random forest and balanced random forest models.
- Gradient Boosting Machines (GBMs): To improve predictive accuracy through sequential learning due to the imbalance of the dataset.
- Deep Neural Networks: To capture non-linear relationships and interactions at scale.

- Cross-validation: We will deploy the validation and error testing methods learned in class to evaluate the accuracy of our models.
- Bootstrapping: In order to generate more units from the minority class we will be boostrapping and sampling from that data.
These methods will be evaluated for their accuracy, precision, and recall to ensure the most reliable predictions.

Expected Challenges:

- Data Quality and Completeness: Ensuring the dataset is clean and comprehensive enough to train effective models. The data is also highly imbalanced towards the no claim status which will initially limit the accuracy of our models.
- Balancing Bias and Variance: To avoid overfitting while still capturing essential trends and patterns in the data.
- Interpretability: Creating models that not only predict accurately but also provide insights into the underlying factors influencing claim outcomes.

Practical Significance:

This project will equip insurance companies with predictive insights that can help in tailoring insurance premiums more accurately, managing risks better, and ultimately enhancing profitability. By understanding the drivers of insurance claims, insurers can also guide policyholders towards safer driving behaviors and vehicle choices. It will also give us real world business insights and skills that we can use to build our professional proficiency.

Project Responsibilities:

- Data Preprocessing: Nico
- Model Development and Training: Cameron
- Evaluation and Refinement: Johnson
- Final Analysis and Reporting: Mariia

Conclusion:

By leveraging machine learning to predict insurance claims based on car data, our project not only addresses a crucial industry need but also paves the way for more data-driven decision-making in the insurance sector. The insights gained will teach us how premiums are calculated, and how claims are managed.