Project Proposal: Predicting Future Loss Claims for Insurance Households

Group 8

Team Members: Nicholas Voltin, Mehmet Cingoz, and Lavenya Mohanasundaram

1. Project Overview

The goal of this project is to predict whether a household will experience a future insurance claim based on historical claims, application details, and subsequent loss experience. The project will use the following datasets:

- **Claim Data (Dataset 1)**: Historical claims at the household level, including claim amounts and fault status.
- **Predictor Dataset (Dataset 2)**: Household information from insurance applications submitted before January 1, 2017.
- **Subsequent Loss Experience (Dataset 3)**: Data on claims that occurred after the application date.

2. Data Sources

- **Dataset 1**: Contains claims data with at least one dollar paid out and prior to January 1, 2017.
- **Dataset 2**: Contains household-level data from the insurance application submitted before January 1, 2017.
- **Dataset 3**: Tracks claims after the application date to assess subsequent loss experiences.

3. Key Questions

- What factors predict whether a household will have a future claim?
- How do previous "At Fault" and "Not At Fault" claims affect the likelihood of future claims?
- Which features from historical claims and the insurance application are most significant for predicting future claims?

4. Objectives

- **Data Aggregation**: Aggregate claims data to count the number of "At Fault" and "Not At Fault" claims for the last 5 years.
- **Data Merging**: Combine the claims data with the application and subsequent loss experience datasets.
- **Modeling**: Build a predictive model to forecast whether a household will have a future claim.
- **Model Evaluation**: Evaluate the model's performance using accuracy, precision, recall, and F1-score.

5. Scope

- Preprocess and clean the data.
- Engineer features such as the number of past claims and types of claims.
- Use logistic regression to predict the future claim indicator (future_clm_ind).
- Provide insights for risk assessment and policy adjustments based on model results.

6. Expected Outcomes

The project will deliver a predictive model that identifies high-risk households and provides insights into the factors driving future claims. This will inform better risk management, customer targeting, and policy pricing.

This approach will enhance the company's ability to assess and predict future claims more accurately, leading to more tailored insurance policies and improved underwriting processes.