Predicting Opportunity of Subrogation with Real Claim Data

Model Citizens, 2025 Travelers UMC

Wenjie Gong, Cecilia Liu, Simeng Wu, Carol Zhou, Franklin Zhou

Department of Statistical Science, Duke University

Introduction

Business Context

Subrogation is a critical part of the claim lifecycle. When a third party is liable, recovery reduces net incurred loss, improves loss ratios, and enhances reserving accuracy—making subrogation a key financial and loss-mitigation lever.

The Core Challenge

Current subrogation identification relies heavily on adjuster judgment and manual file review. This process is slow, inconsistent, and often results in missed recovery opportunities across thousands of claims.

Subrogation Modeling Framework

Our Mission

Build a predictive model using 2020–2021 first-party physical damage claims to flag potential subrogation opportunities, identify key indicators, and provide recommendations for operational use.

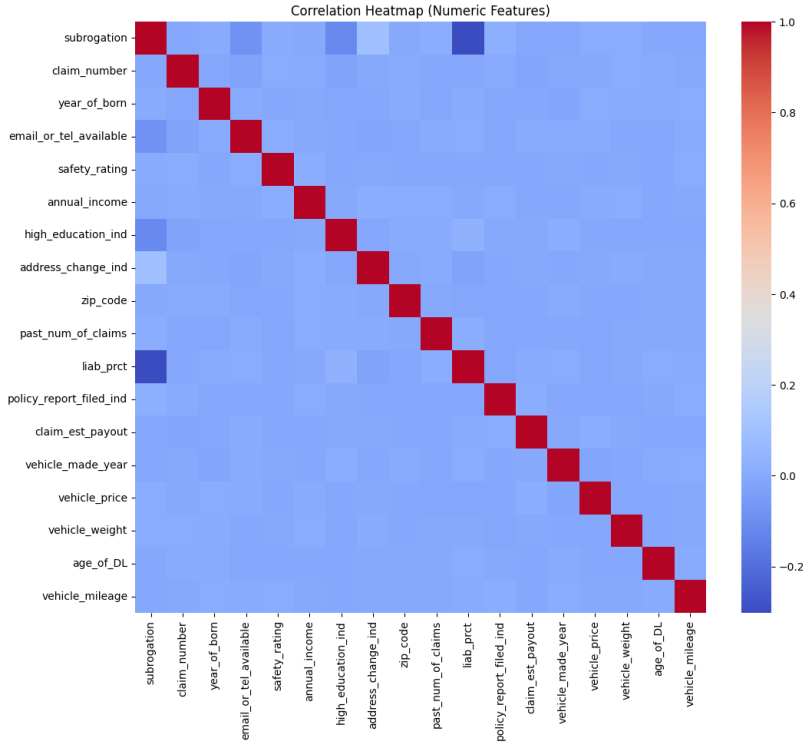
Modeling Objective

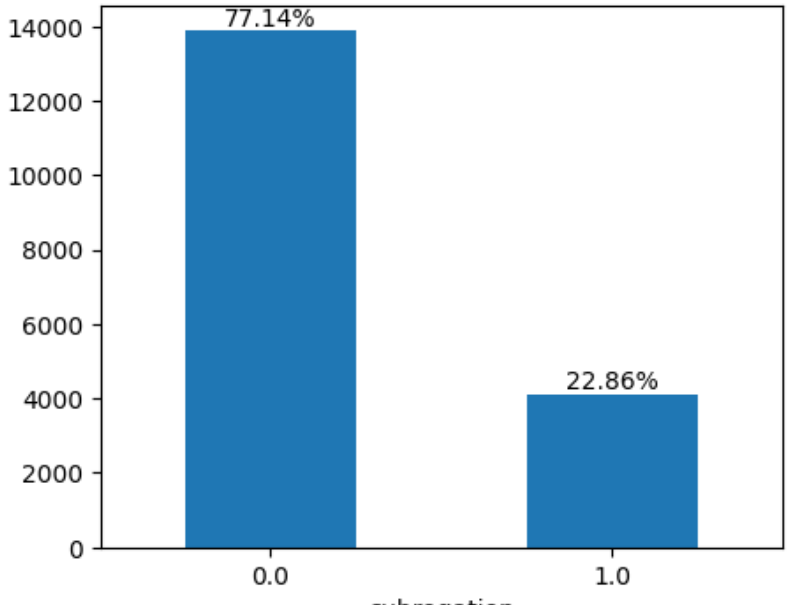
Predict a binary outcome *Subrogation Opportunity*
(1 = likely recovery, 0 = not likely).
Evaluation Metric: F1 score (balances precision and recall due to asymmetric business costs).

Business Value

- Improve recovery rates and reduce net incurred losses
- Help subrogation specialists prioritize high-value cases
- Reduce time spent on low-likelihood opportunities
- Support data-driven decision-making in the claims process

Data Overview





Dataset Summary

- Training data: 18,000 rows with subrogation indicator (0/1)
- Test data: 12,000 rows without the indicator
- Features include policyholder, driver, vehicle, accident details, and estimated payout

Data Quality Steps

- Removed 2 rows with missing subrogation indicator
- Removed 1 row with all values missing
- Test dataset contains no NAs
- Dropped `vehicle_made_year` (post-claim dates $\rightarrow$ impossible)
- Excluded `age_of_vehicle` (unreliable reporting)

Why We Used Multiple Preprocessing Pipelines?

Each model family has its own optimized pipeline, in accordance with their algorithm characteristics.

Model	**TabM**	**Linear Model**	**CatBoost**	**LightGBM**	**XGBoost**
Numerical	Normalize (quantile or z-score)	Normalize (z-score)	Works well with raw data	Works well with raw data	Works well with raw data
Categorical	One-Hot Encoding	One-Hot Encoding	Native handling	Native handling	One-Hot Encoding

Methods

- Placeholder
- Placeholder

XGBoost

- Placeholder
- Placeholder

LightGBM

- Placeholder
- Placeholder

CatBoost

- Placeholder
- Placeholder

TabM

- Placeholder
- Placeholder

Logistic Regression

- Placeholder
- Placeholder

Ensemble (XXBoost)

- Placeholder
- Placeholder

Results

- Placeholder

Conclusion

- Placeholder