Insurance Fraud Detection

Global Insure Case Study

A data-driven classification modeling approach to early fraud identification that helps minimize financial losses while optimizing operational efficiency.

By TUYET TU and ANIRUDHA SAHU

The Challenge

High Volume

Global Insure processes thousands of claims annually.

Significant Fraud

A notable percentage prove to be fraudulent.

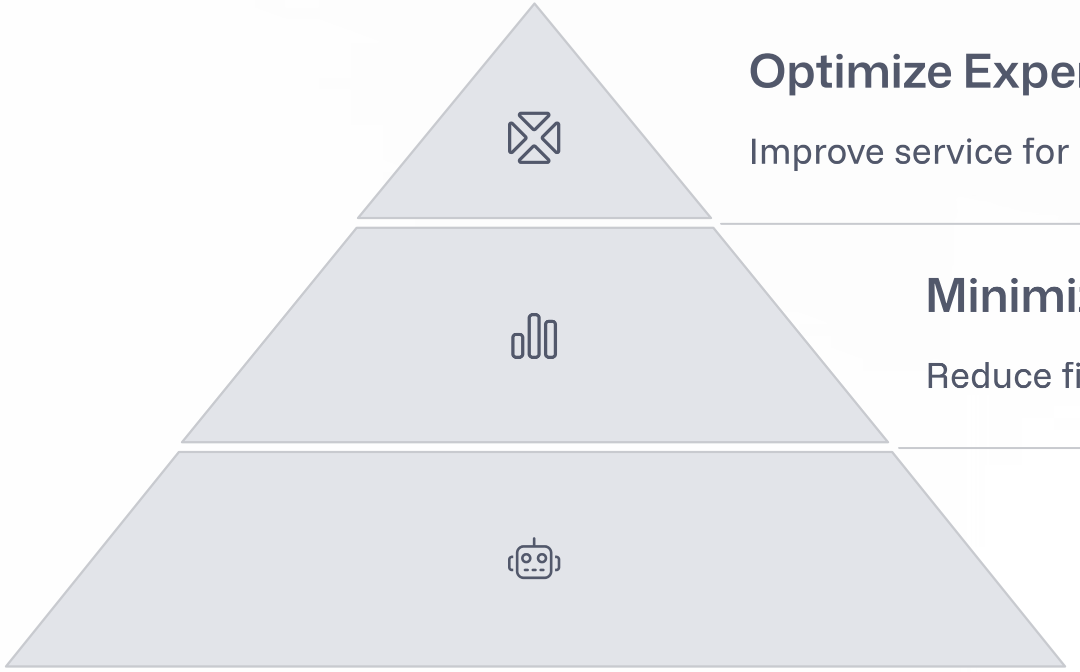
Inefficient Process

Current manual inspection is time-consuming.

Financial Impact

Fraud often detected after payment, causing mounting losses.

Business Objective



Optimize Experience

Improve service for legitimate claimants

Minimize Losses

Reduce financial impact of fraud

Build Predictive Models

Classify claims as fraudulent or legitimate

We aim to enable early fraud detection before approval and payment by utilizing historical data, customer profiles, and claim details.

Our Approach

Data Preparation

Clean and organize 1,000 claims with 40 features.

Exploratory Analysis

Uncover patterns and relationships in the data.

Feature Engineering

Create new predictive signals from existing data.

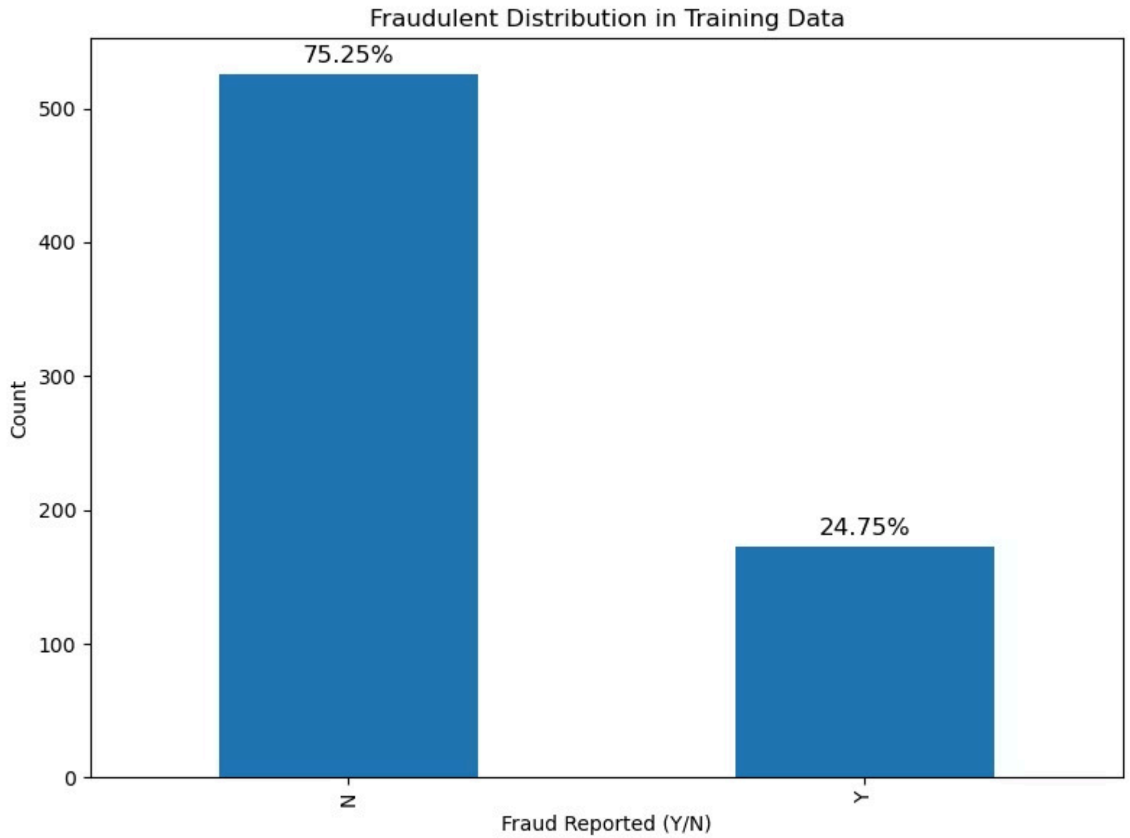
Model Building

Develop Logistic Regression and Random Forest models.

Evaluation

Test and validate model performance.

Understanding Our Dataset



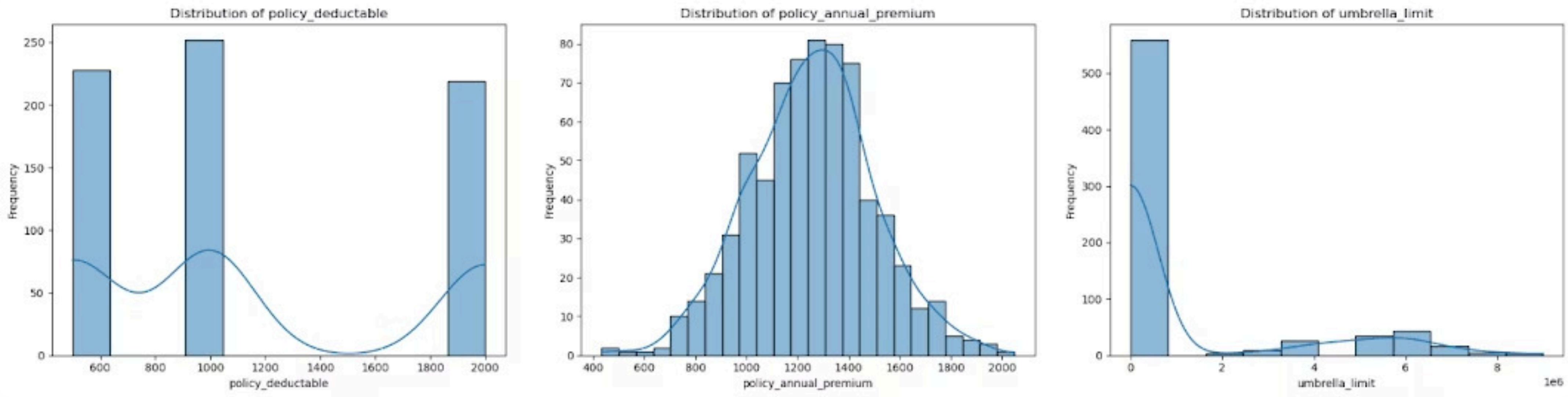
Our dataset contains 1,000 insurance claims with 40 features.

The target variable analysis revealed that 24.75% of claims were labeled as fraudulent, indicating a moderate class imbalance that needed to be addressed during modeling.

We addressed the moderate class imbalance through RandomOverSampler technique.

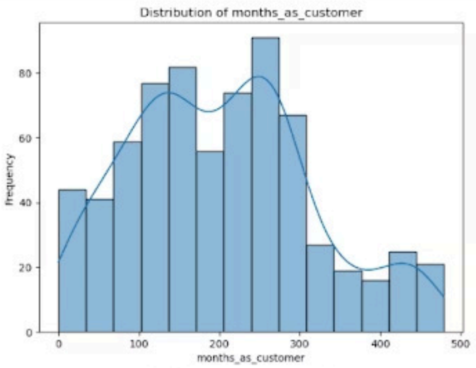
Insurance Policy Characteristics:

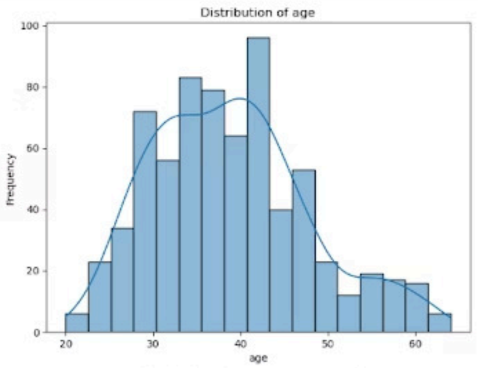
- Policy deductibles showed a trimodal distribution with peaks at $600, $1,000, and $2,000
- Policy annual premiums were normally distributed around $1,000-1,200
- Umbrella limits were highly right-skewed with most policies having minimal coverage

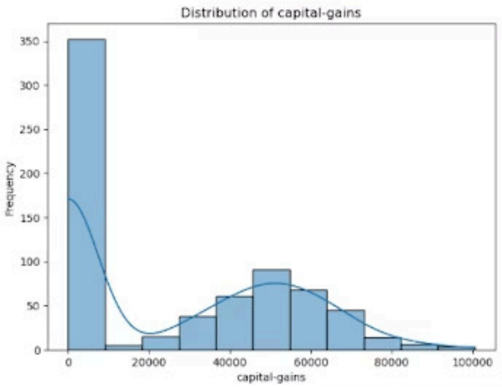


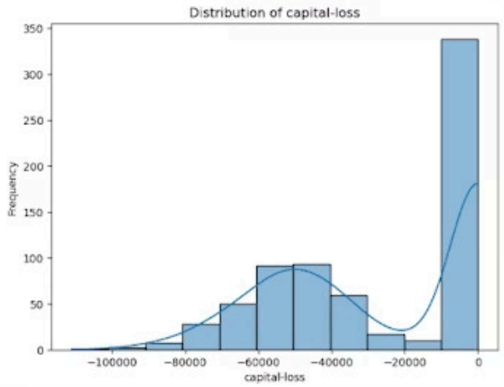
Customer Demographics:

- Age followed a normal distribution centered around 35-40 years
- Customer tenure (months_as_customer) showed a right-skewed multi-modal distribution
- Capital gains/losses exhibited significant zero-inflation patterns



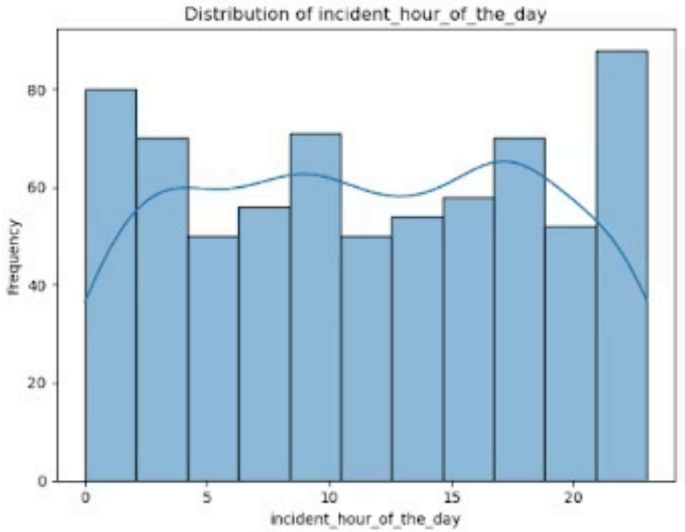


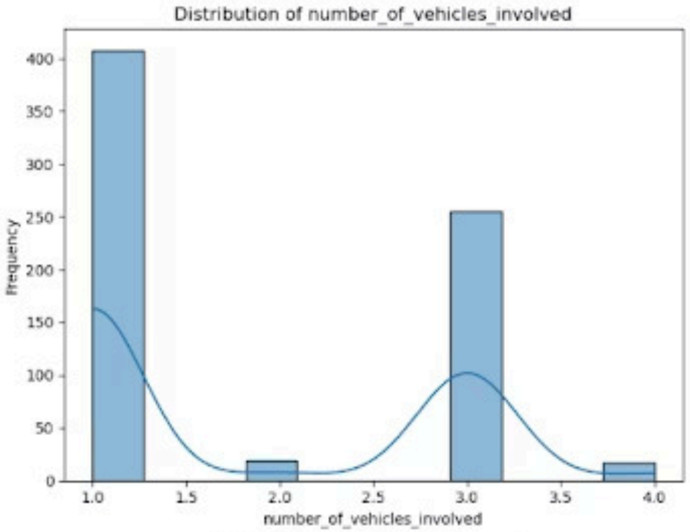


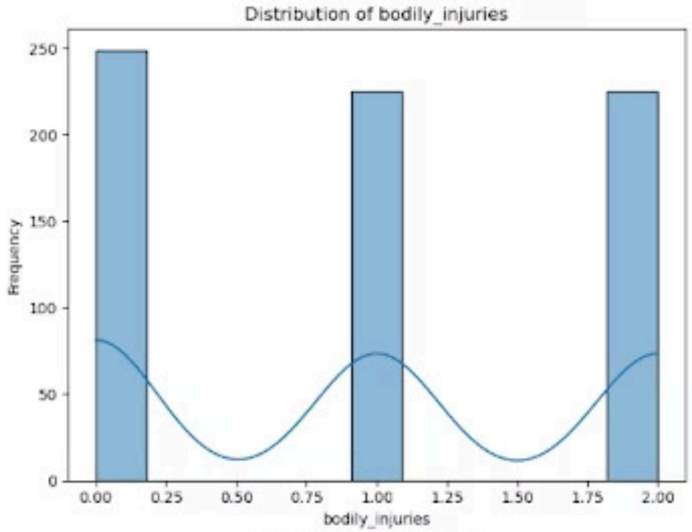


Incident Characteristics:

- Incident hours showed slight peaks during early morning and evening hours
- Vehicle involvement had strong bimodal distribution with peaks at 1 and 3 vehicles
- Bodily injuries and witness counts followed discrete distributions with specific common values

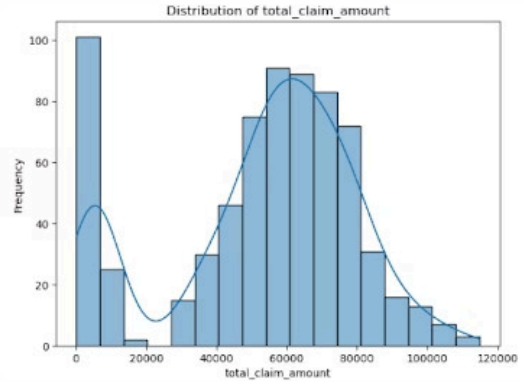


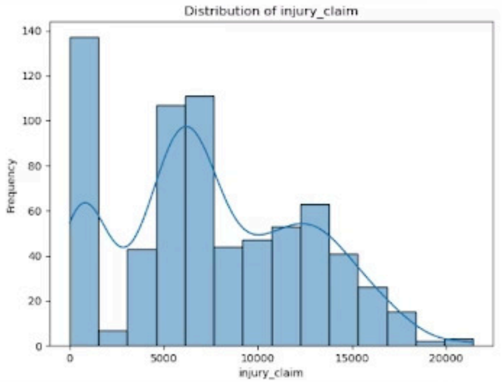


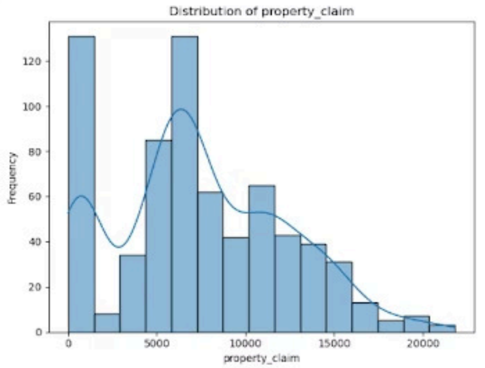


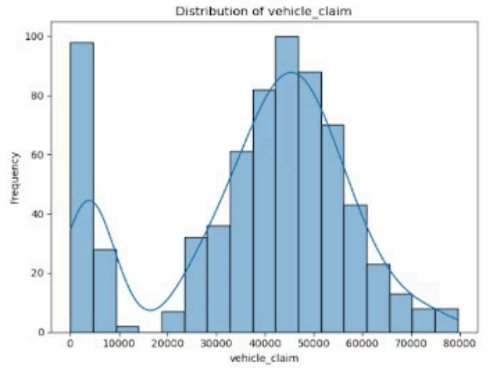
Claims Information:

- Total claim amounts showed bimodal distribution with small claims around $0 and larger claims at $60,000-80,000
- The claim components (injury, property, vehicle) all showed distinctive bimodal patterns
- Suspicious clustering at specific claim amounts suggested potential fraud patterns









Understanding Fraud Patterns

Demographic Insights

- Male claimants show higher fraud rates
- Master's degrees: 45% fraud rate
- High School: only 10-15% fraud rate

Claim Characteristics

- Vehicle theft claims have very low fraud rates
- Higher deductibles correlate with more fraud
- Fewer witnesses indicate higher fraud likelihood
- Later incident hours appear more suspicious

Numerical Features Analysis

Policy Deductible

Higher deductibles ($2,000 vs. $1,000) correlate with fraudulent claims.

Witness Count

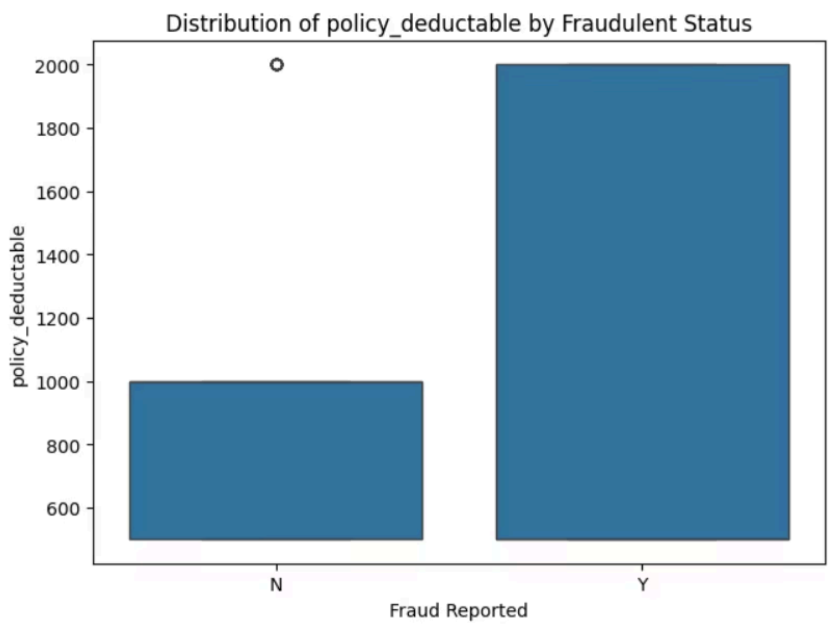
Fewer witnesses present in fraudulent claim scenarios.

Incident Timing

Later hours of the day show higher fraud probability.

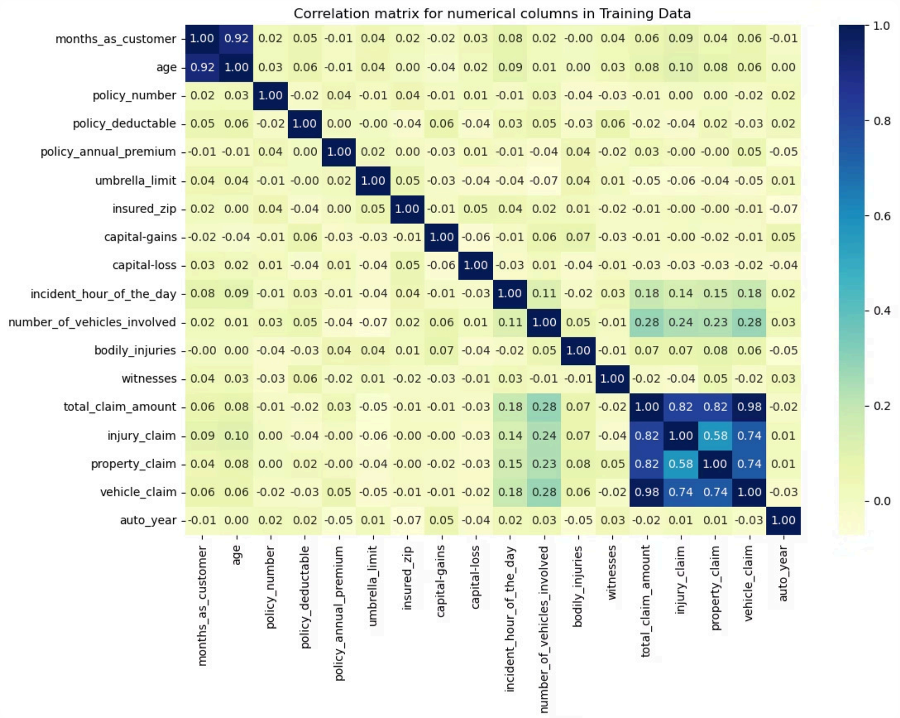
Injury Severity

More severe bodily injuries reported in fraudulent claims.



Understanding Feature Relationships

These correlation findings informed our feature engineering approach.



Claim Components

Strong multicollinearity (0.74-0.98) between total_claim_amount and its components (injury_claim, property_claim, vehicle_claim)

Customer Profile

High correlation between customer age and tenure (0.92).

Vehicle Factors

- Number of vehicles involved correlated moderately (0.23-0.28) with claim amounts
- Incident hour showed relationships with claim amounts (0.14-0.18), suggesting time-of-day patterns

Feature engineering

Time-based features

- Time categories (morning, afternoon, evening, night)
- Late night flag for incidents occurring during high-risk hours

Policy timing features:

- Days between incident date and policy bind date
- Suspicious quick claim flag (≤30 days between policy binding and incident)

Claim composition features:

- Percentage features showing claim composition ratios instead of absolute amounts
- Flag for suspicious patterns (high claims with few witnesses)
- Claim-to-coverage ratio comparing total claim to policy coverage limits

Customer features:

- Customer tenure ratio (months as customer / age * 12) to evaluate customer loyalty

Vehicle features:

- Vehicle age categories (new, recent, old) replacing specific year information

Categorical Value Grouping

To reduce dimensionality and increase predictive power, we grouped categorical values:

Education

- Low risk
- Medium risk
- High risk

Occupation

- Low risk
- High risk

Hobby

- Very high risk
- High risk
- Medium risk
- Low risk

Vehicle Brand

- High risk
- Low risk

State

- High risk
- Low risk

Feature Transformation

We implemented several technical transformations:

Categorical Features

Created dummy variables for all categorical features with drop_first=True

Target Variable

Converted the target variable "fraud_reported" from Y/N to 1/0

Numerical Features

Applied standard scaling to numerical features

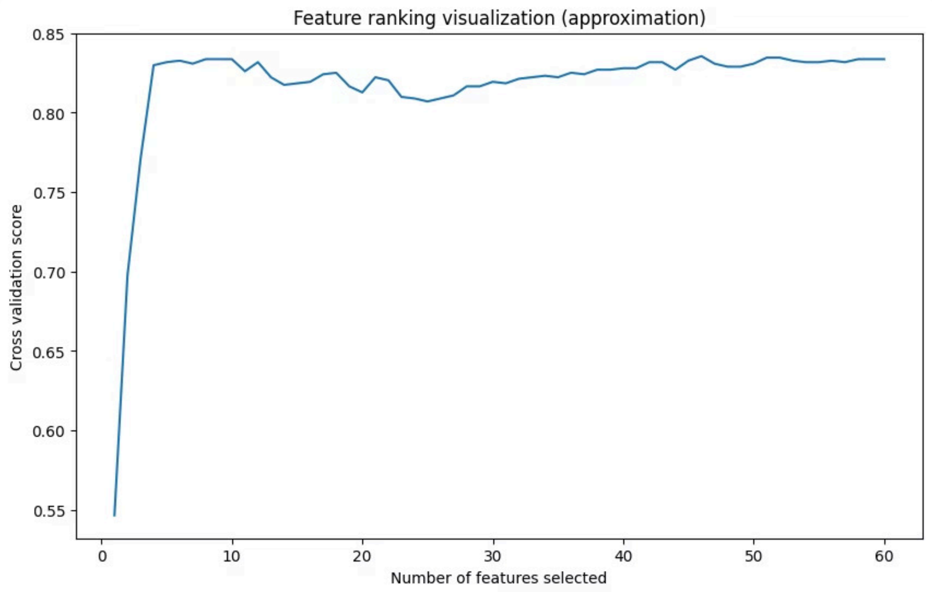
Feature Selection

We used Recursive Feature Elimination with Cross-Validation (RFECV) to identify the most relevant features for our logistic regression model. This process:

- Employed 5-fold cross-validation
- Iteratively removed less important features
- Selected the optimal feature subset based on cross-validation scores

Top 10 features by RFECV

Feature
age
state_risk_low_risk_state
incident_type_Vehicle Theft
collision_type_Front Collision
collision_type_Rear Collision
incident_severity_Minor Damage
incident_severity_Total Loss
incident_severity_Trivial Damage
authorities_contacted_Fire
authorities_contacted_Not reported



Logistic Regression

We built a logistic regression model using Statsmodels to enable detailed statistical analysis:

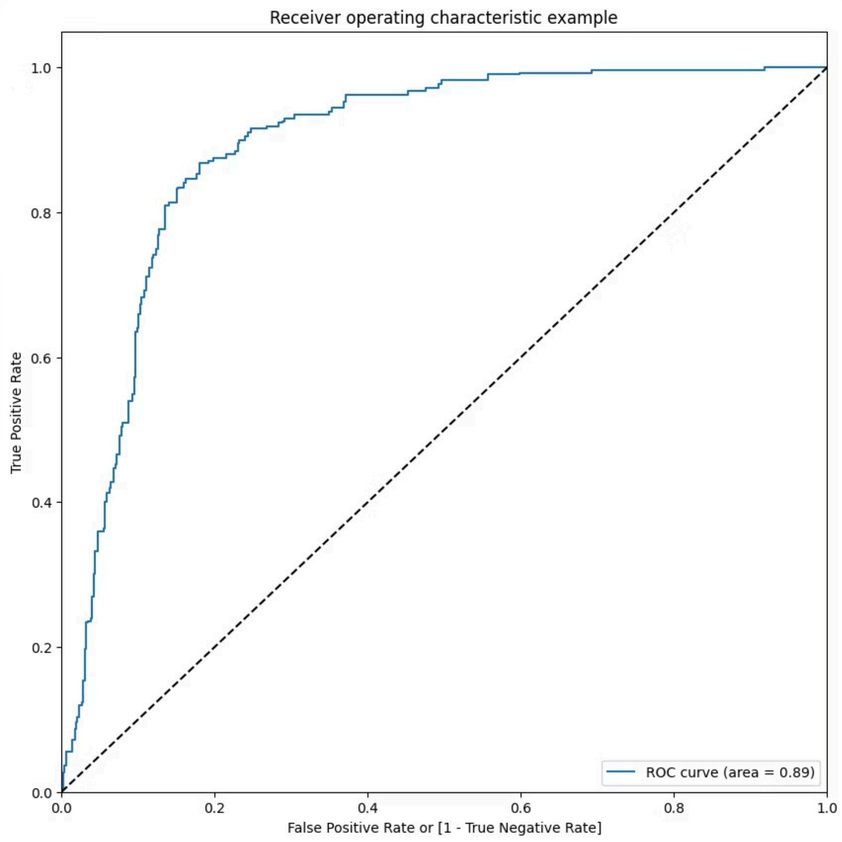
- Evaluated p-values to assess feature significance
- Calculated Variance Inflation Factors (VIFs) to detect multicollinearity
- Iteratively removed variables with high p-values (>0.05) and high VIFs (>10)
- Achieved a final model with all variables significant (p<0.05) and VIFs <5

Logistic Regression Model

The initial logistic regression model achieved:

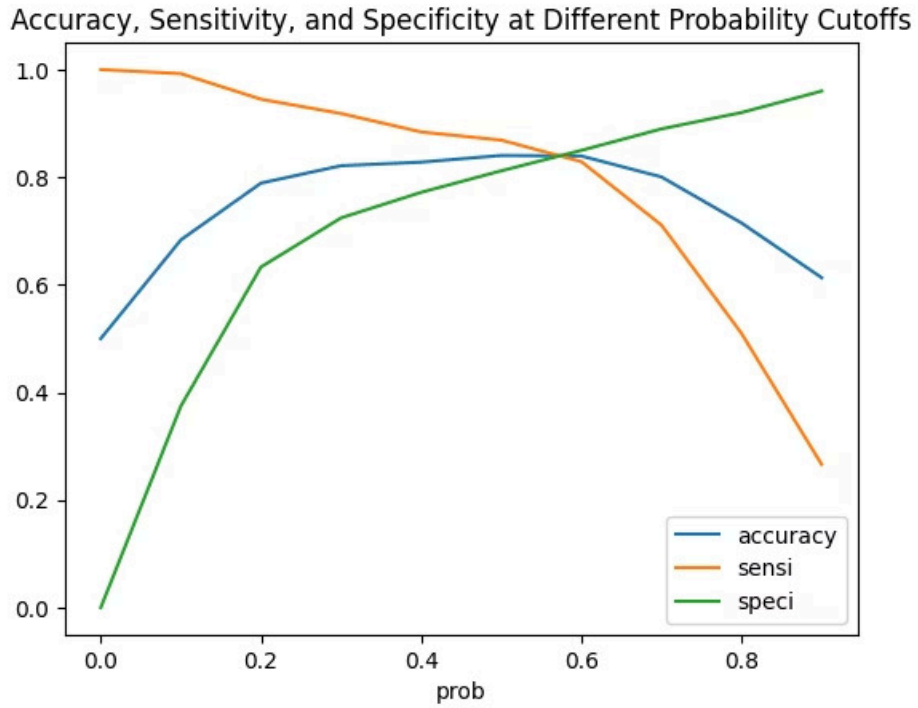
- 84% accuracy on the training set
- 87% sensitivity
- 81% specificity
- 82% precision
- 84% F1 score

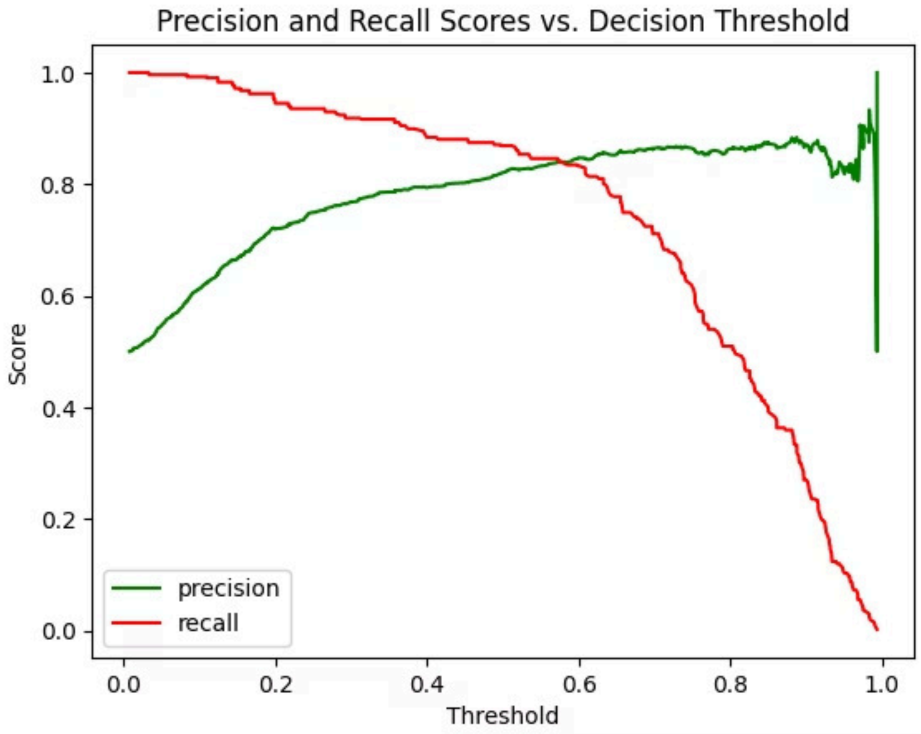
We also plotted ROC curves to find the optimal probability cutoff, with the area under the ROC curve reaching 0.89, indicating strong discriminatory power.



Finding the Optimal Threshold

As we plot accuracy, sensitivity, specificity at different values of probability cutoffs, and also the plotting the precision-recall curve, we see that the cut-off of 0.5 is a good balance in both charts.



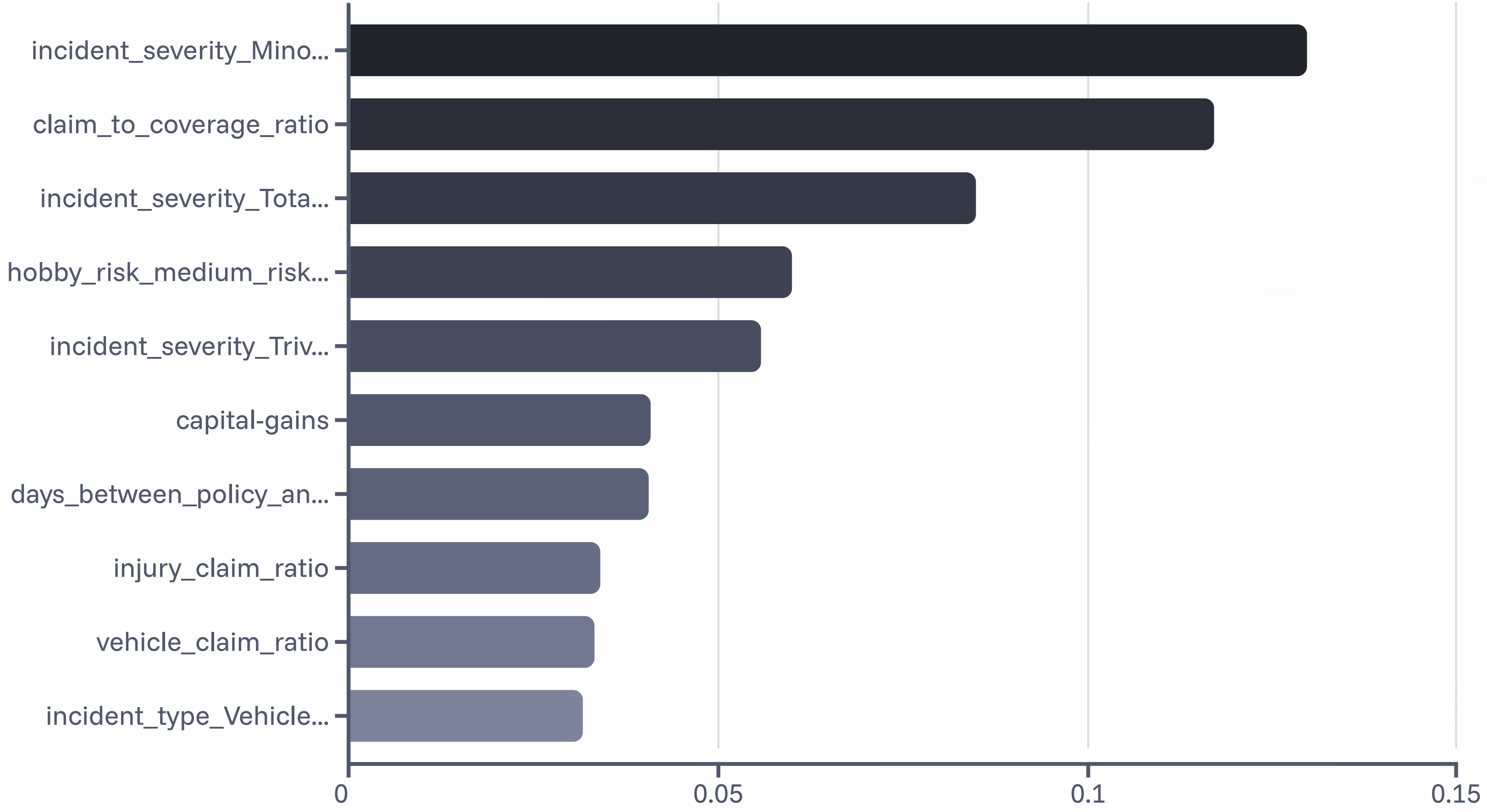


Random Forest Model

We implemented a Random Forest model to capture complex non-linear relationships:

- Identified feature importance scores
- Selected the top 15 most important features

Most Influential Features by Random Forest



Feature importance guided our feature selection process in the Random Forest model.

Random Forest Model

Used grid search for hyperparameter tuning: rf_best = grid_search.best_estimator_.

The tuned Random Forest model achieved exceptional training performance:

- 92% accuracy
- 97% sensitivity
- 88% specificity
- 89% precision
- 93% F1 score

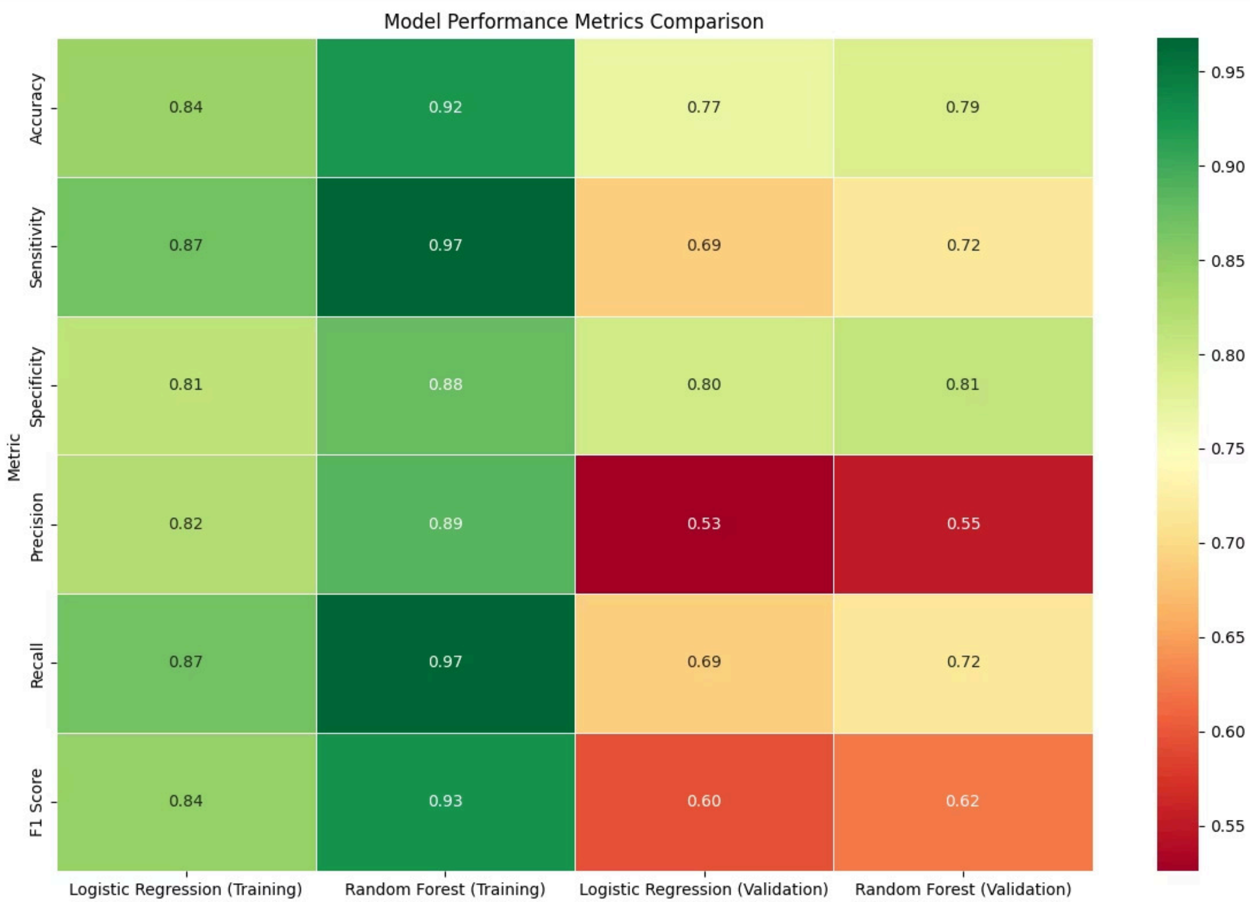
Model Performance on Validation data

Metric	Logistic Regression	Random Forest
Accuracy	77%	79%
Sensitivity	69%	72%
Specificity	80%	81%
Precision	53%	55%
F1 Score	60%	62%

The Random Forest model shows slightly better performance across all metrics on validation data.

Validation Performance

When evaluating both models on the validation set, we observed a significant performance drop:



Our fraud detection models achieved reasonable accuracy (77-79%) on validation data, significantly improving over random classification. However, several performance concerns emerged:

1. **Overfitting**: Both models showed substantial drops between training and validation data, with Random Forest exhibiting more extreme overfitting (92% training accuracy vs. 79% validation).
2. **Precision challenges**: The low precision (53-55%) means nearly half of claims flagged as fraudulent were actually legitimate, risking customer dissatisfaction without careful review.
3. **Recall-precision tradeoff**: While achieving reasonable sensitivity (69-72%), this came at the cost of precision, highlighting the challenge in fraud detection—balancing false positives and negatives.
4. **Model comparison**: Random Forest slightly outperformed Logistic Regression in most metrics, suggesting benefits from capturing non-linear relationships, despite greater overfitting.

Key Questions Addressed

How can we analyze historical claim data to detect patterns that indicate fraudulent claims?

Our approach combined exploratory data analysis, feature engineering, and machine learning. The most effective techniques were bivariate analysis comparing fraud/non-fraud characteristics, creating derived features (like claim-to-coverage ratios and policy timing flags), and applying both linear and non-linear models. This multi-faceted approach revealed patterns that would be difficult to detect through manual review alone.

Key Questions Addressed

Which features are the most predictive of fraudulent behavior?

The strongest fraud predictors were:

1. Claims filed shortly after policy initiation (within 30 days)
2. Claim amounts approaching coverage limits
3. High claim amounts with few witnesses
4. Specific demographic factors (education, occupation)
5. Vehicle characteristics (certain makes and older vehicles)
6. Late-night incident timing

Key Questions Addressed

Based on past data, can we predict the likelihood of fraud for an incoming claim?

Yes, with reasonable accuracy. Our models achieved 77-79% accuracy on validation data, with the Random Forest model correctly identifying 72% of fraudulent claims. While precision remains a challenge (55%), the models provide sufficiently reliable probability scores to prioritize claims for investigation, enabling early fraud detection before payment processing.

Key Questions Addressed

What insights can be drawn from the model that can help in improving the fraud detection process?

Key actionable insights include:

1. Implement tiered risk classification (low/medium/high) rather than binary decisions
2. Enhance verification for claims filed shortly after policy initiation
3. Apply risk-based verification protocols based on demographic and geographic factors
4. Strengthen witness documentation requirements for high-value claims
5. Incorporate vehicle characteristics into risk assessment procedures

These insights can transform Global Insure's fraud detection process by enabling earlier identification, more efficient resource allocation, and reduced impact on legitimate claims.

Thank You