FRAUDLENT CLAIM DETECTION

Case Study Submission

Submitted by:
Amit Kumar
Amit Kumar Roy

Problem Statement / Background

Background:

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process.

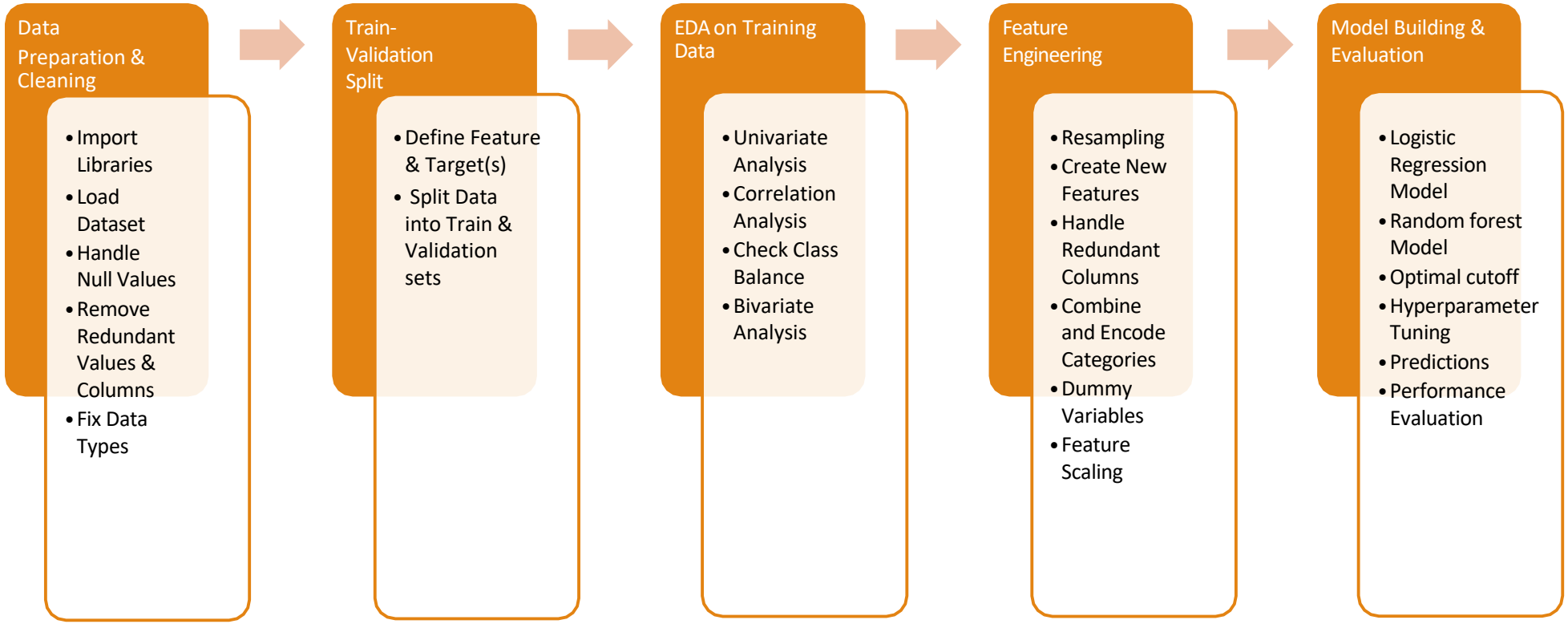
Business Objectives:

Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.

Goals:

1. Analyse historical claim data to detect patterns that indicate fraudulent claims.
2. Identify the features which are most predictive of fraudulent behaviour.
3. Predict the likelihood of fraud for an incoming claim, based on past data.
4. Draw the insights from the model that can help in improving the fraud detection process.

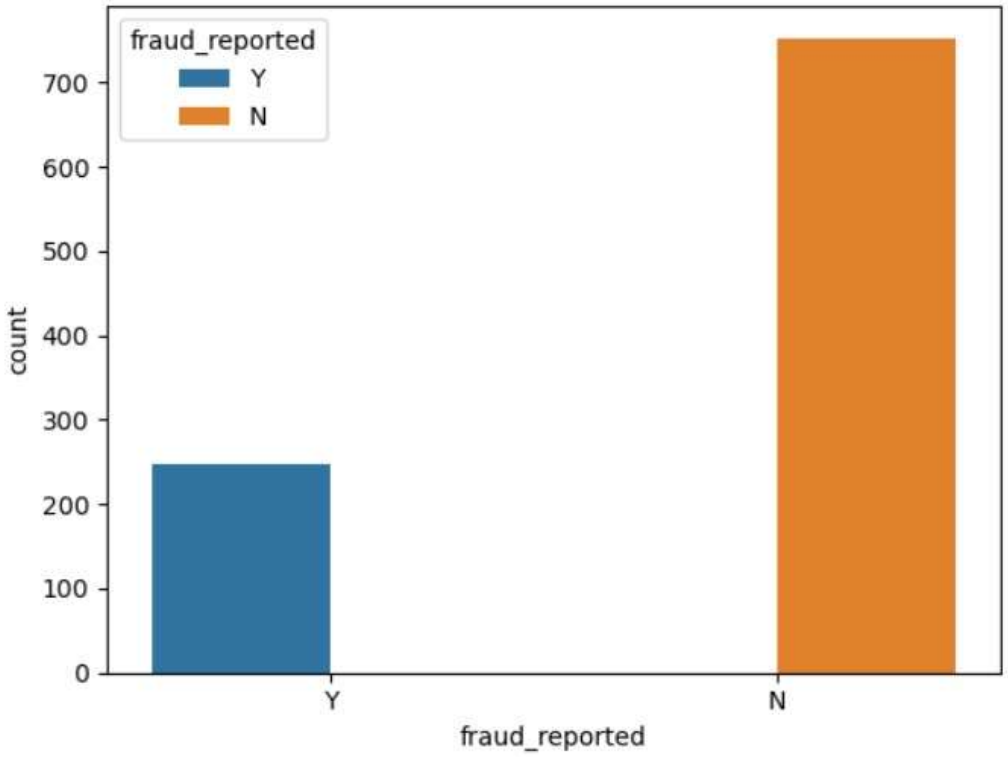
Process / Methodology



Class Distribution Overview

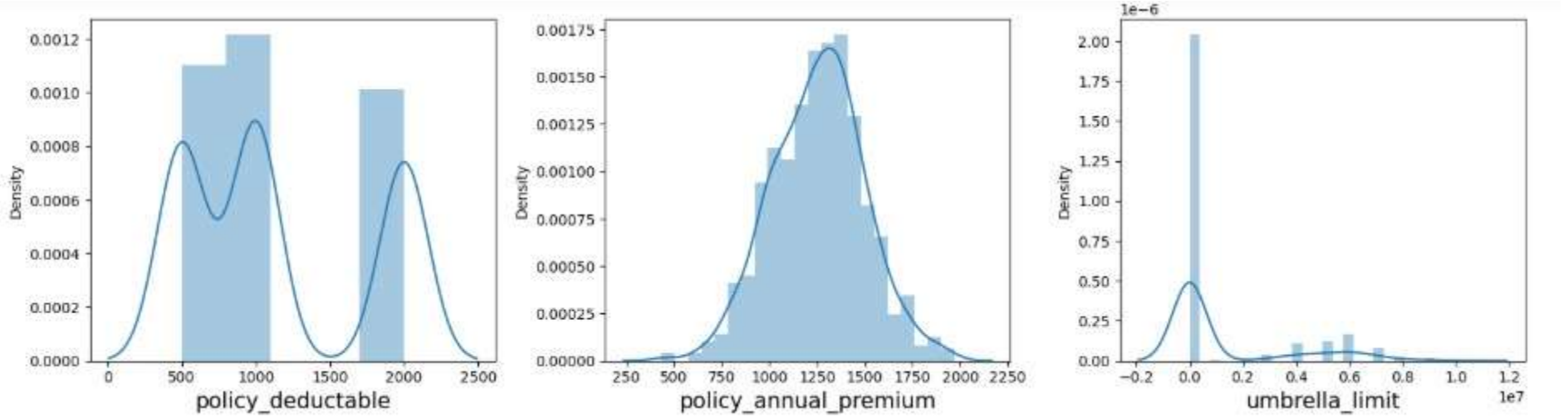
- **Chart**: Bar chart of fraud_reported vs. count
- **Observation:**

The dataset is imbalanced with ~77% legitimate and ~23% fraudulent claims. This indicates the need for stratified sampling or resampling techniques to address class imbalance during model training.



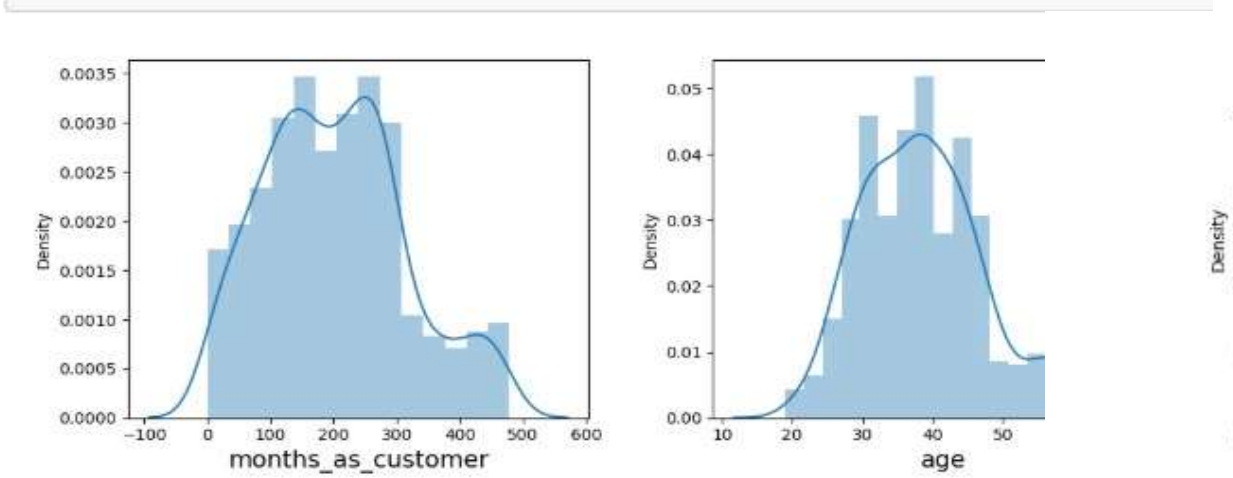
Insurance **Policy Characteristics**

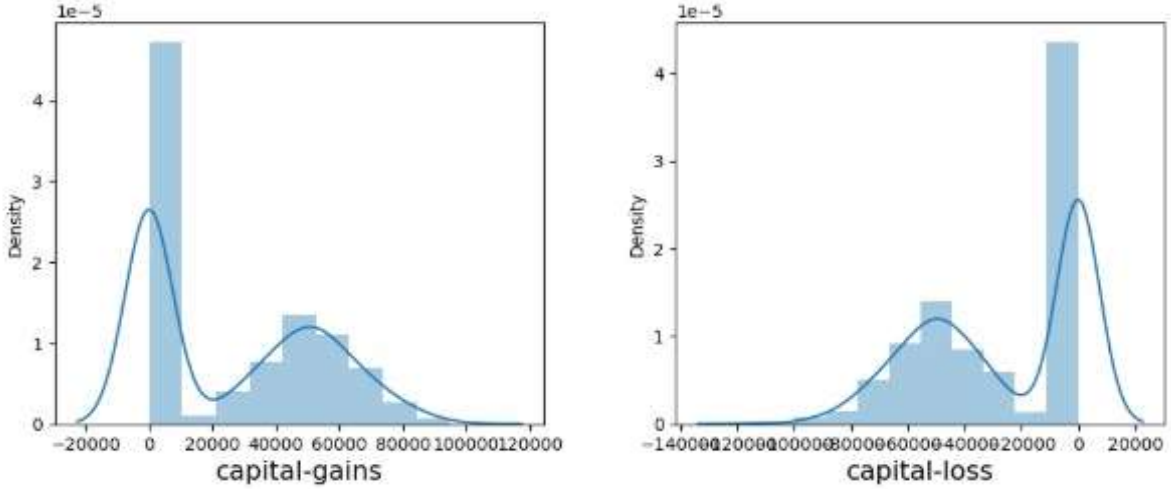
- Policy deductibles showed a trimodal distribution
- Policy annual premiums were normally distributed around $1,000-1,200
- Umbrella limits were highly right-skewed with most policies having minimal coverage



Customer Demographics

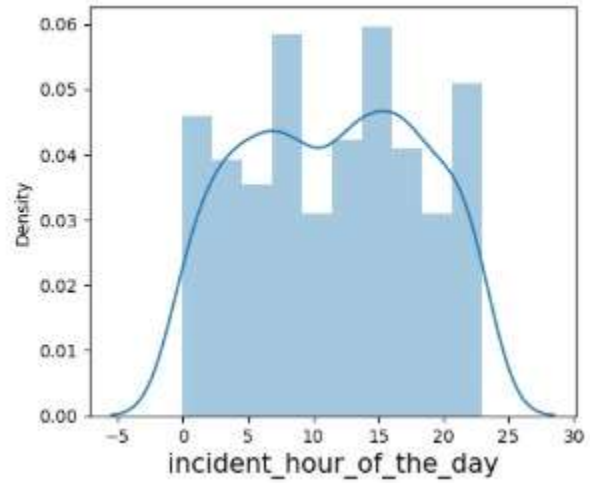
- Age followed a normal distribution centered around 35-40 years
- Customer tenure (months_as_customer) showed a right-skewed multi-modal distribution Capital gains/losses exhibited significant zero-inflation patterns

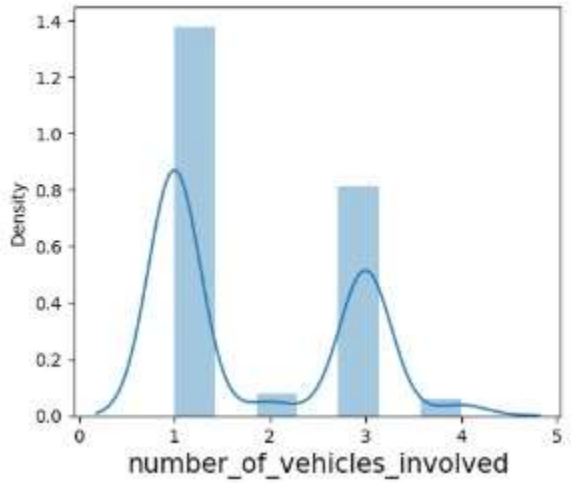


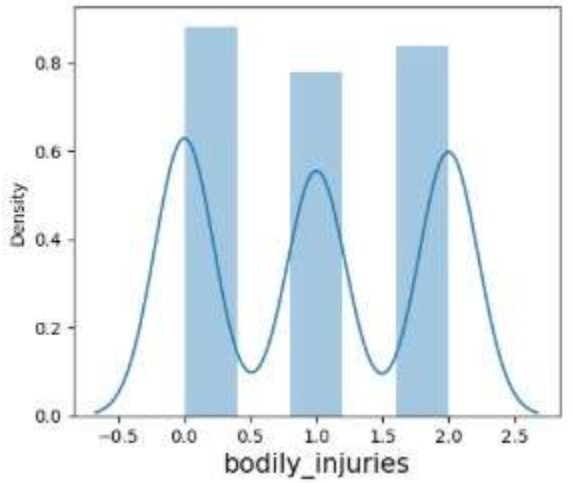


Incident Characteristics

- 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

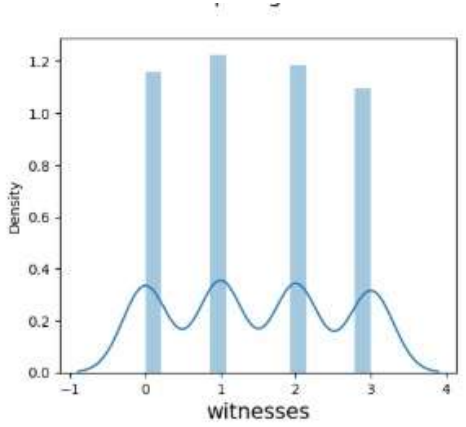


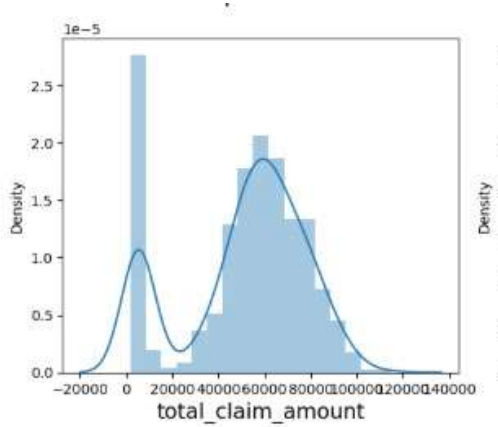


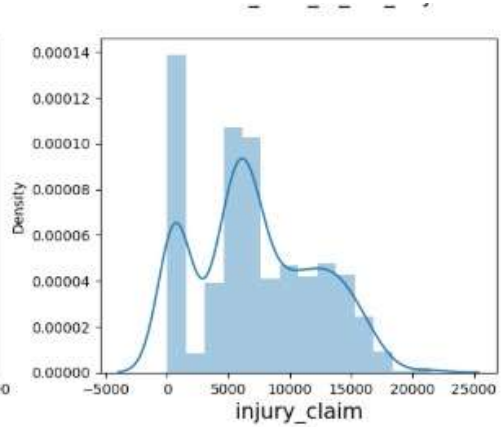


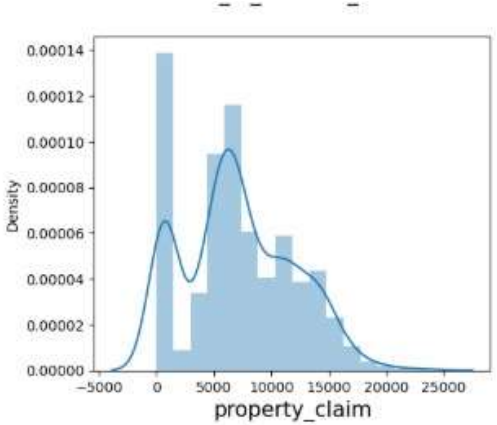
Claims Information

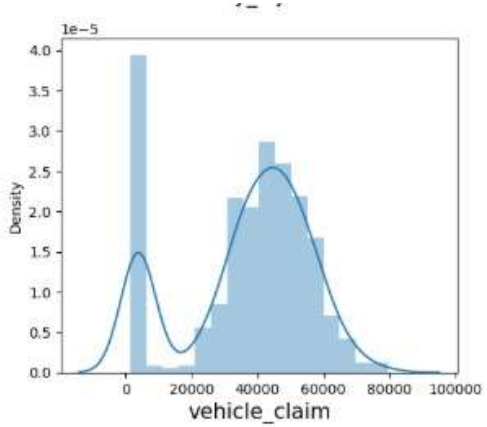
- Total claim amounts showed bimodal distribution with small claims around $0 and larger claims at $60,000-80,000
- The claim components (injury, property, vehicle) all showed distinctive bimodal patterns Suspicious clustering at specific claim amounts suggested potential fraud patterns









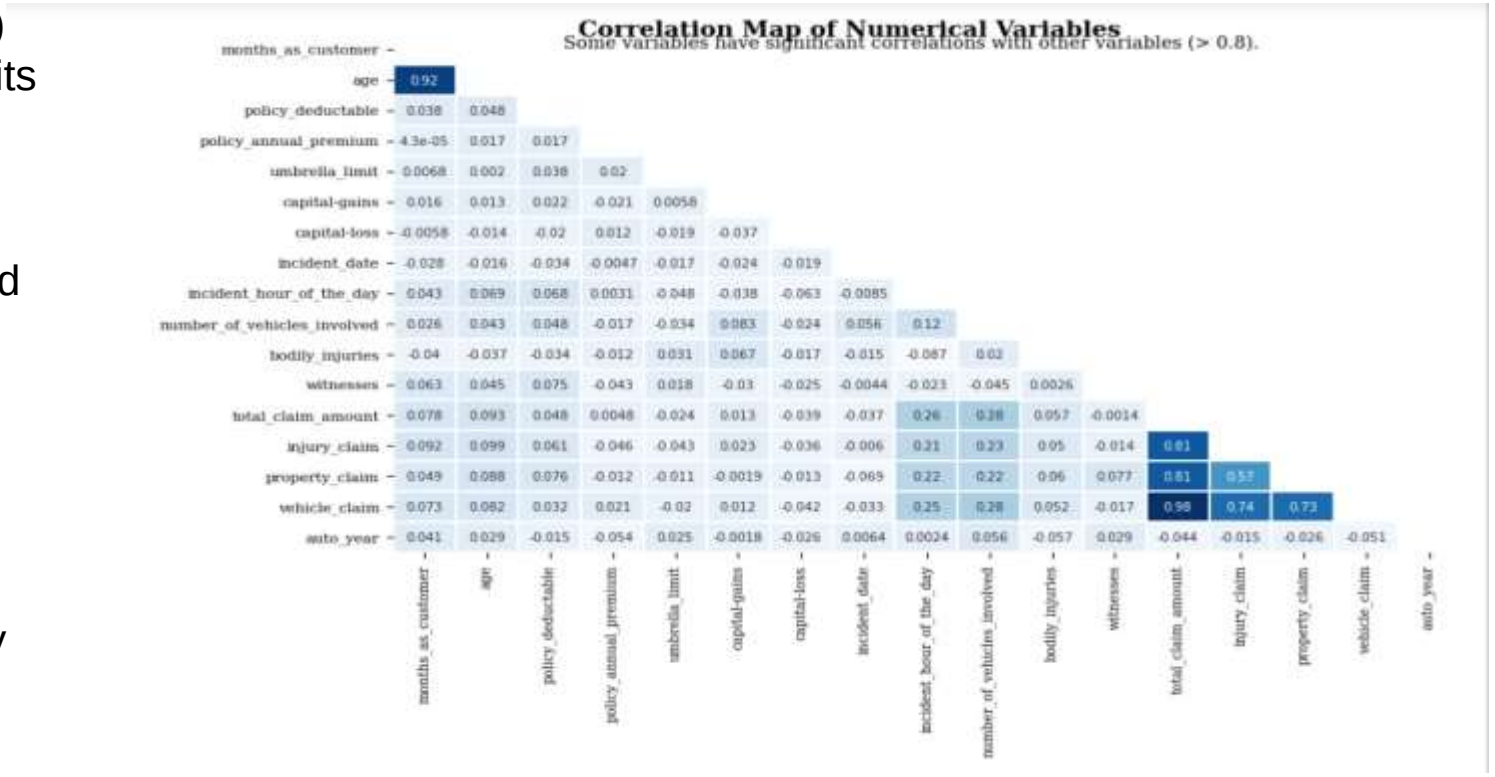


Fraud Patterns

- Umbrella limits were highly right-skewed with most policies having minimal coverage.
- 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
- Vehicle involvement had strong bimodal distribution with peaks at 1 and 3 vehicle
- Total claim amounts showed bimodal distribution with small claims around $0 and larger claims at $60,000-80,000

Correlation Analysis

- Strong multicollinearity (0.74-0.98) between total_claim_amount and its components (injury_claim, property_claim, vehicle_claim)
- Very high correlation (0.92) between months_as_customer and age
- Number of vehicles involved correlated moderately (0.23-0.28)with claim amounts
- Incident hour showed relationshipswithclaim amounts (0.14-0.18), suggesting time-ofday patterns



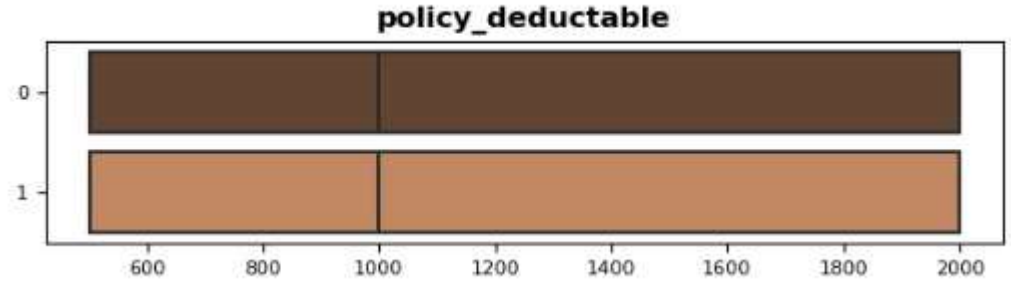
Numerical feature Analysis

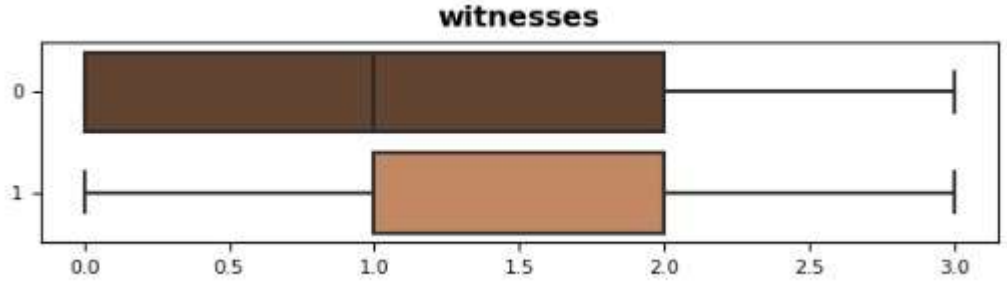
Policy Deductible

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

Witness Count

- Fewer witnesses present in fraudulent claim scenarios.

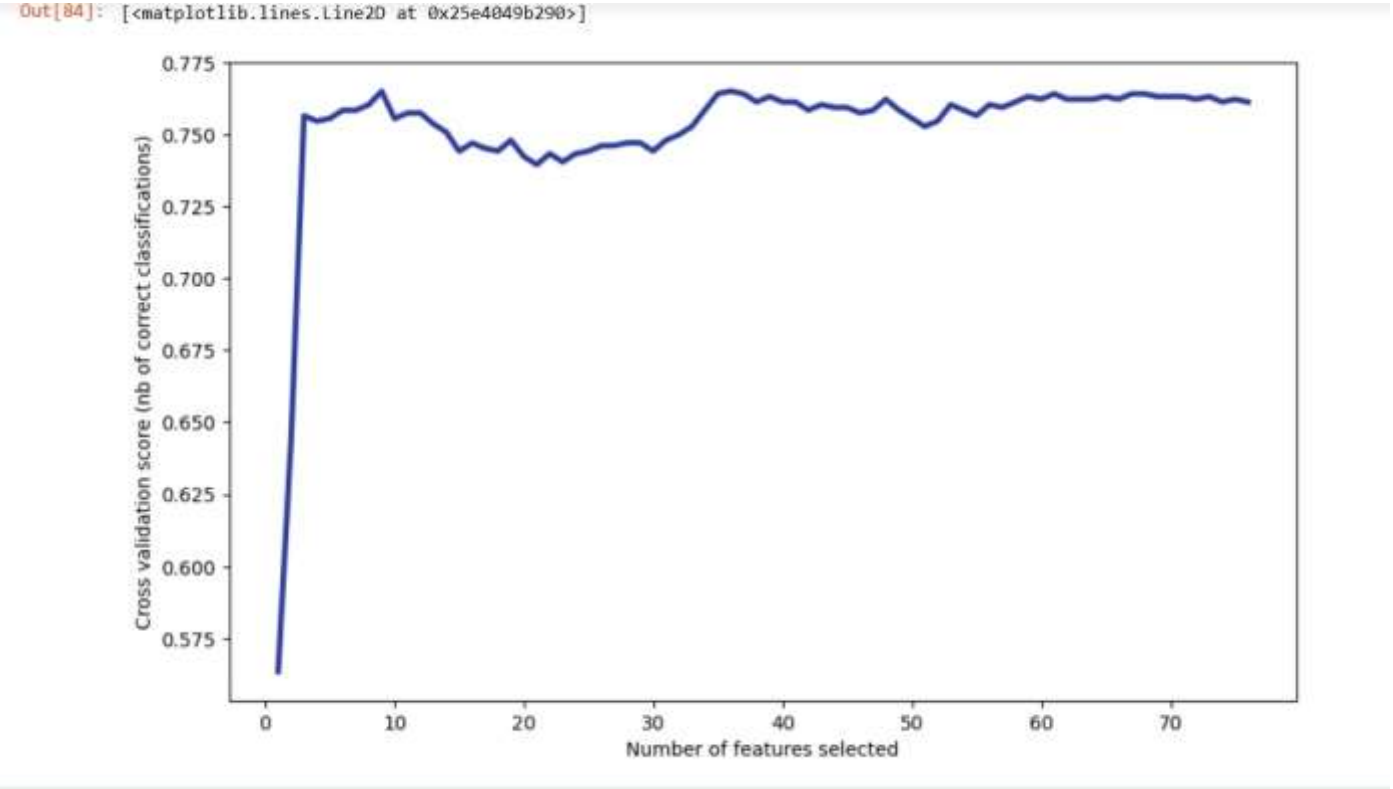




Feature Engineering

Feature Selection:

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



Logistics Regression

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

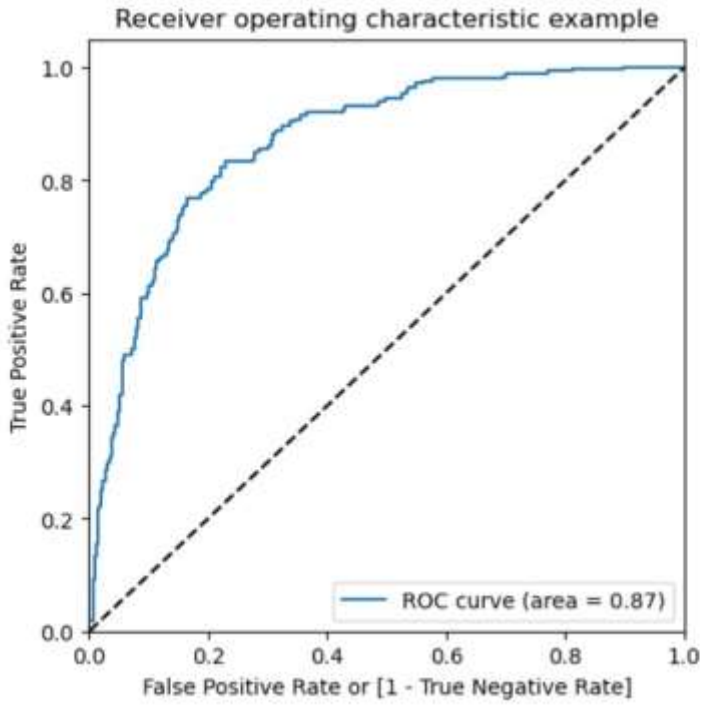
- Evaluated p-values to assess feature significance
- Calculated Variance Inflation Factors (VIFs) to detect multicollinearity Iteratively removed variables with high p-values (>0.05) and high VIFs (>10)
- Achieved a final model with all variables significant ($p<0.05$) and VIFs <5

Logistics Regression Model

The initial logistic regression model achieved:

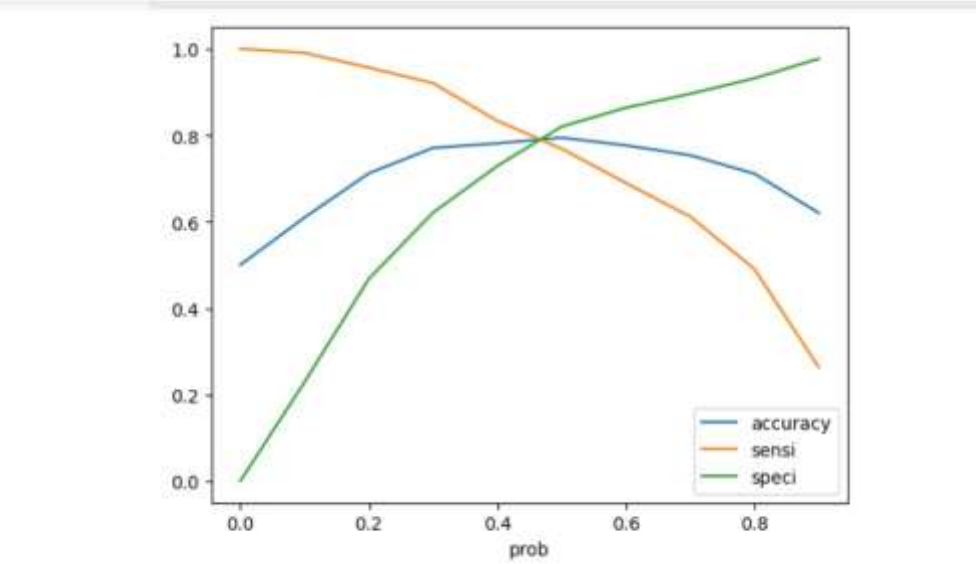
- 79% accuracy on the training set
- 76.89% sensitivity
- 82% specificity
- 81% precision
- 78.91% F1 score

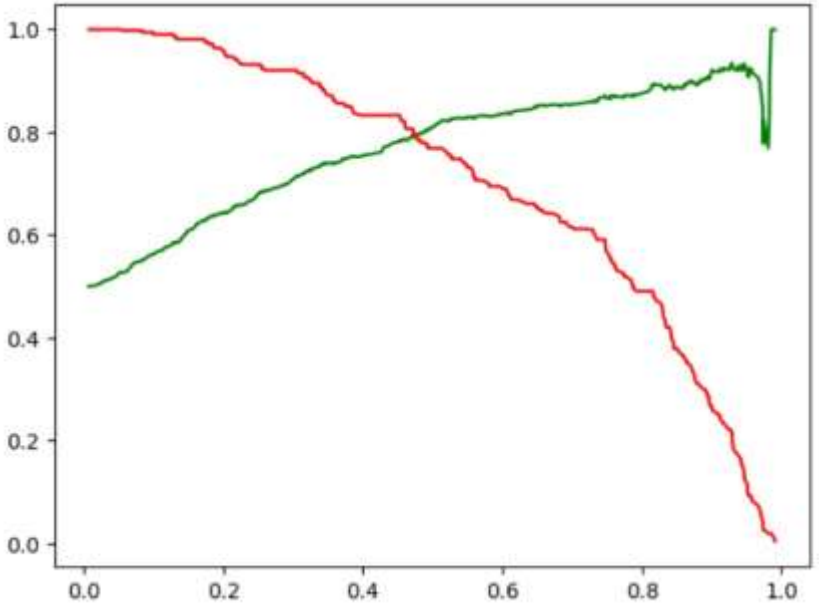
We also plotted ROC curves to find the optimal probability cutoff, with the area under the ROC curve reaching 0.87, indicating strong discriminatory power.



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

Usedgridsearchforhyperparameter tuning:
rf_best=grid_search.best_estimator_
The tuned Random Forest model achieved exceptional training performance:
● 79.5% accuracy ● 97.1% sensitivity ● 89.96% specificity ● 90.63% precision ● 93.78% F1 score

Evaluation On Validation Data

When evaluating both models on the validation set, we observed a significant performance drop:
Logistic Regression on validation data: ● 72% accuracy ● 66% sensitivity ● 73% specificity ● 45% precision ● 54% F1 score

Random Forest on validation data: ● 77% accuracy ● 62% sensitivity ● 82% specificity ● 54% precision ● 58% F1 score

The tuned Random Forest model shows strong training performance but a significant drop on validation data, especially in sensitivity, precision, and F1 score. This indicates overfitting, where the model learns the training data too well but fails to generalize. Regularization and model simplification are needed to improve validation performance.

Validation Performance

Our fraud detectionmodels achieved reasonable accuracy (77-79%) on validation data, which represents a significant improvement over random classification. However, several performance concerns emerged:

1. Overfitting: Both models showed substantial performance drops between training and validation data, with the Random Forest model exhibiting more extreme overfitting (92% training accuracy vs. 79% validation accuracy).
2. Precision challenges: The relatively low precision(53-55%) on validation data indicates that approximately half of the claims flagged as fraudulent were actually legitimate, which could lead to customer dissatisfaction if implemented without careful review.
3. Recall-precision tradeoff: While achieving reasonable sensitivity (69-72%), this came at the cost of precision, highlighting the inherent challenge in fraud detection—balancing false positives and false negatives.
4. Model comparison: The Random Forest model slightly outperformed Logistic Regression on validation data in most metrics, suggesting some benefit from capturing nonlinear relationships, despite more severe 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
. 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.