Fraudulent Claim Detection Report

Amit Kumar
Amit Kumar Roy

1. Executive Summary

This report presents a comprehensive analysis of fraud detection for Global Insure, a leading insurance company facing significant financial losses due to fraudulent claims. The current manual inspection process is time-consuming and inefficient, often detecting fraud after payments have been made. Using a dataset of 1,000 insurance claims with 40 features, we developed predictive models to classify claims as fraudulent or legitimate at an early stage in the approval process.

Our analysis implemented both Logistic Regression and Random Forest models, with the Random Forest model achieving 79% accuracy on validation data. The most significant predictors of fraud included claim-to-coverage ratios, suspicious timing patterns, and demographic risk factors. Despite reasonable accuracy, both models showed concerning gaps between training and validation performance, indicating opportunities for further refinement.

This report details our methodology, findings, and strategic recommendations for optimizing Global Insure's fraud detection capabilities, with the goal of minimizing financial losses while improving operational efficiency.

2.Problem Statement

Insurance fraud leads to significant financial losses and delayed genuine claim approvals. Manual processes are inefficient in identifying fraud early. This project aims to build a machine learning model that can flag potentially fraudulent claims at the time of submission.

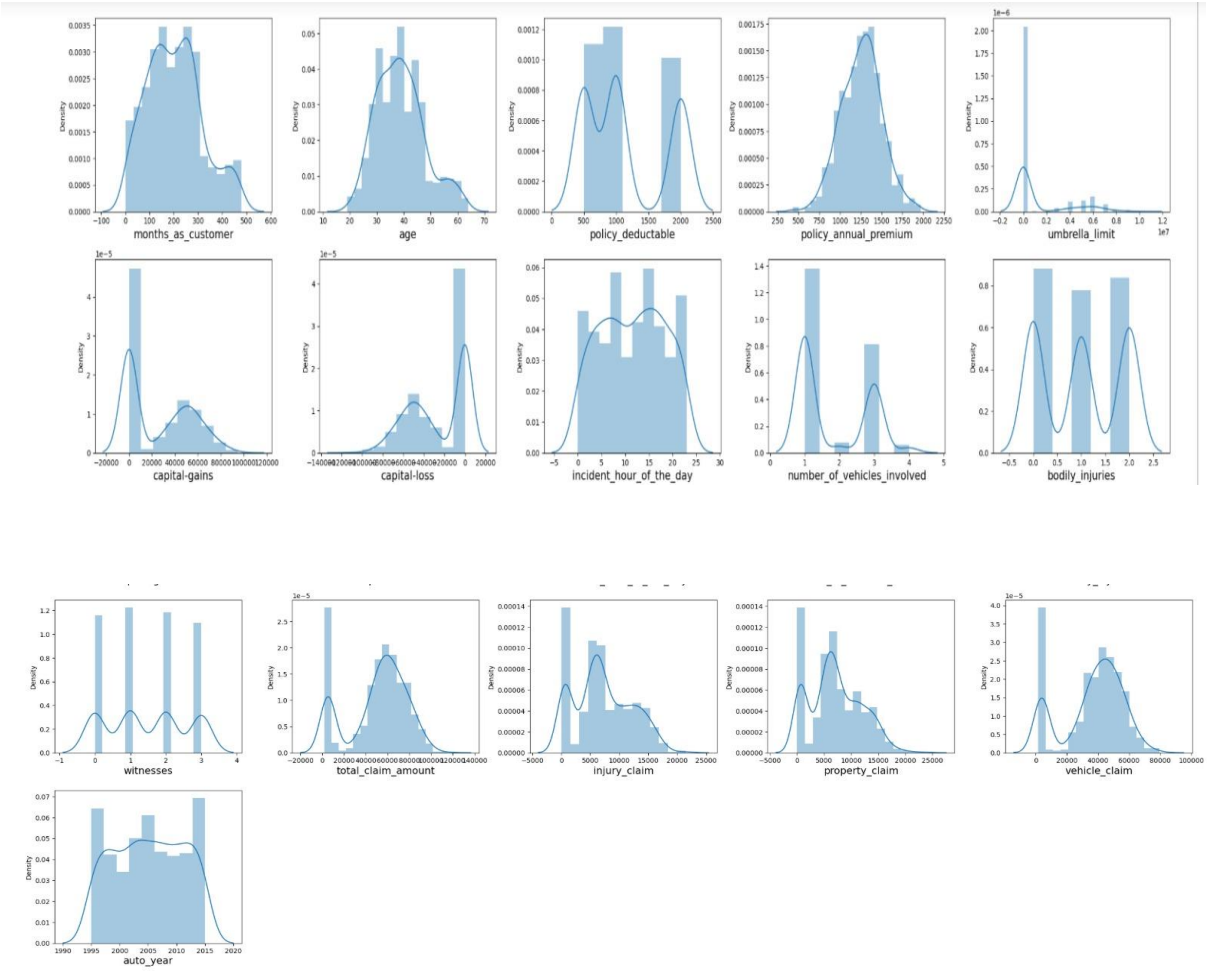
3. Methodology

The process followed to address the problem includes:
- Data loading and preprocessing (handling missing values, encoding categorical variables)
- Exploratory Data Analysis (EDA) to identify key patterns
- Feature selection based on visual and statistical insights
- Model training using Random Forest
- Evaluation of model performance using Accuracy, F1 Score, and ROC AUC

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis:

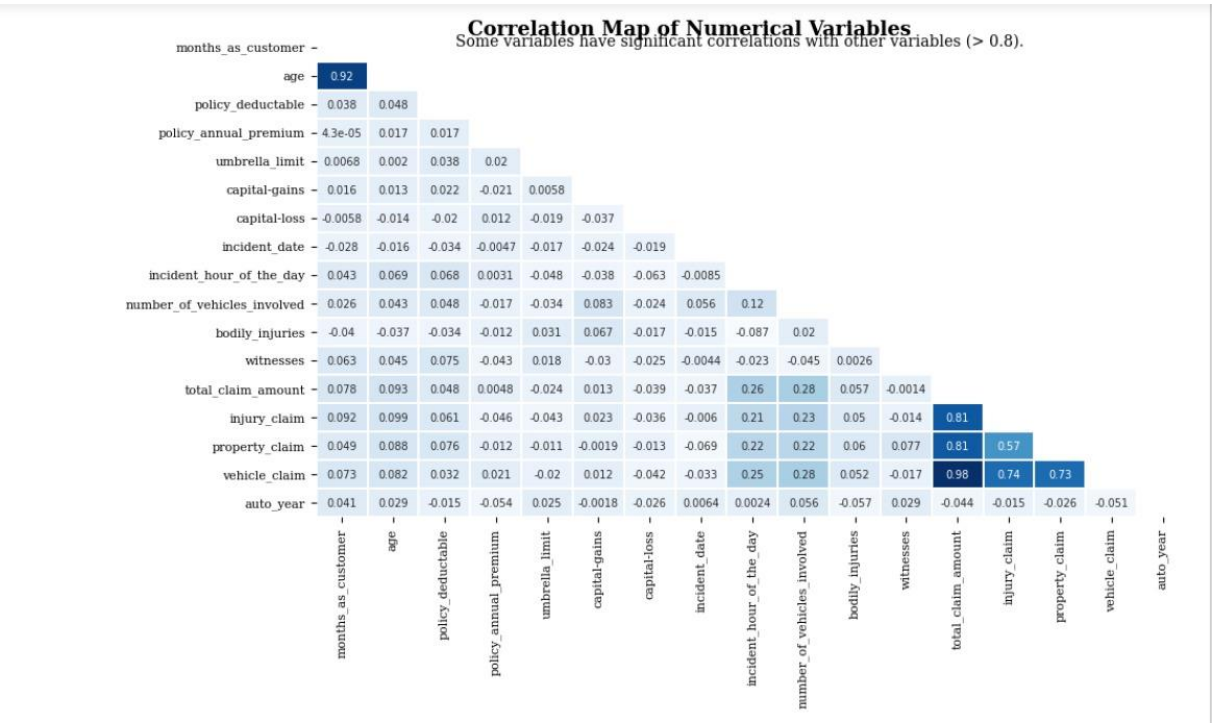
Analysis of numerical features revealed several interesting patterns:



- 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

4.2 Correlation Analysis:

The correlation analysis revealed several important relationships:



Highly correlated features:

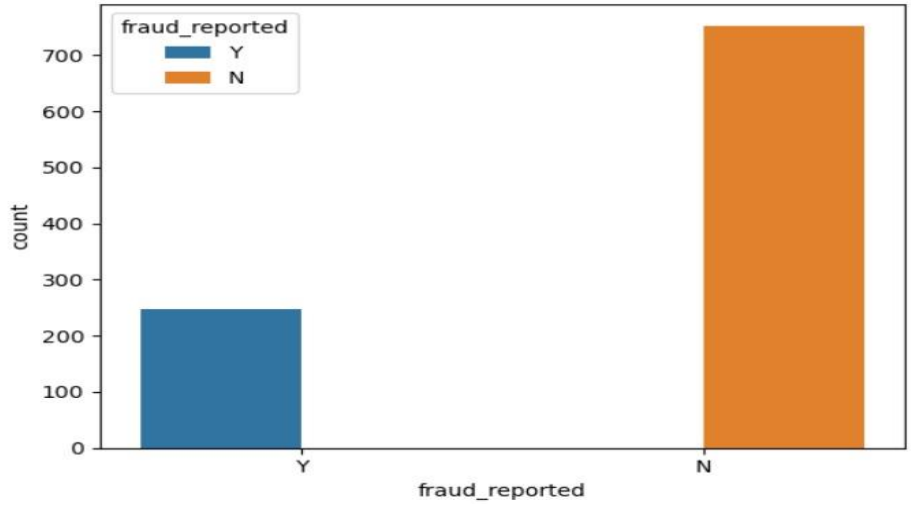
- Strong multicollinearity (0.74-0.98) between total_claim_amount and its components (injury_claim, property_claim, vehicle_claim)
- Very high correlation (0.92) between months_as_customer and age

Moderate correlations:

- 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

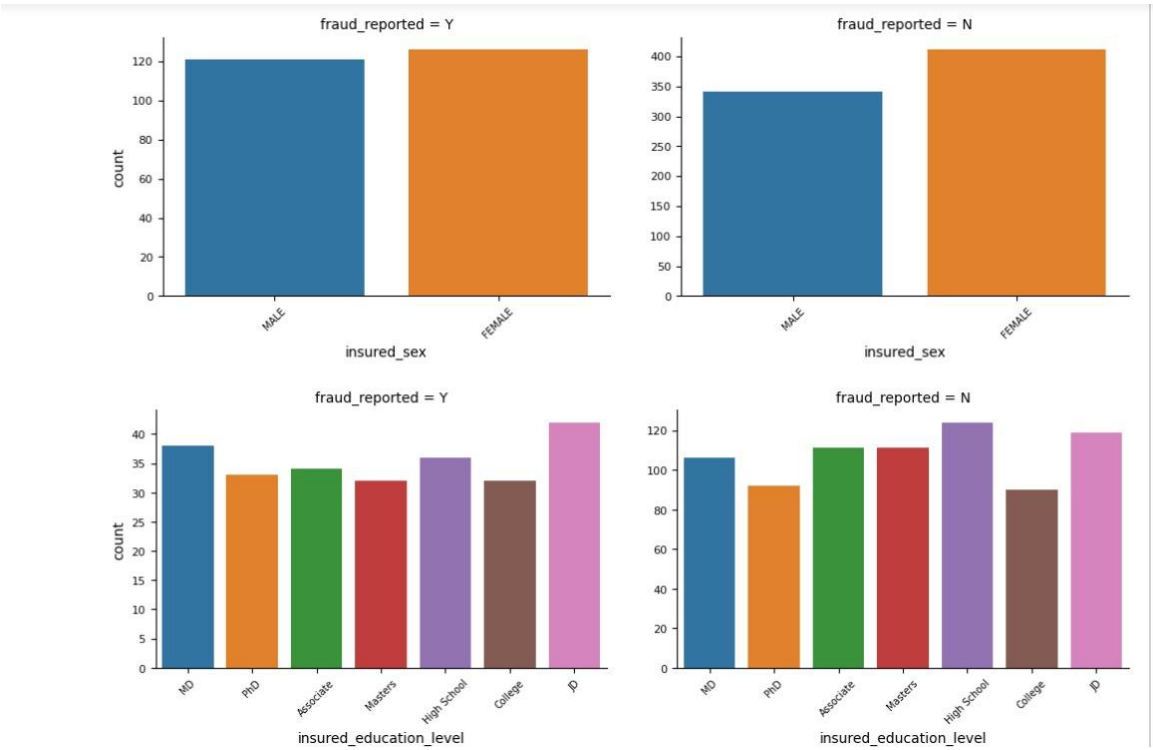
4.3 Class Balance Analysis

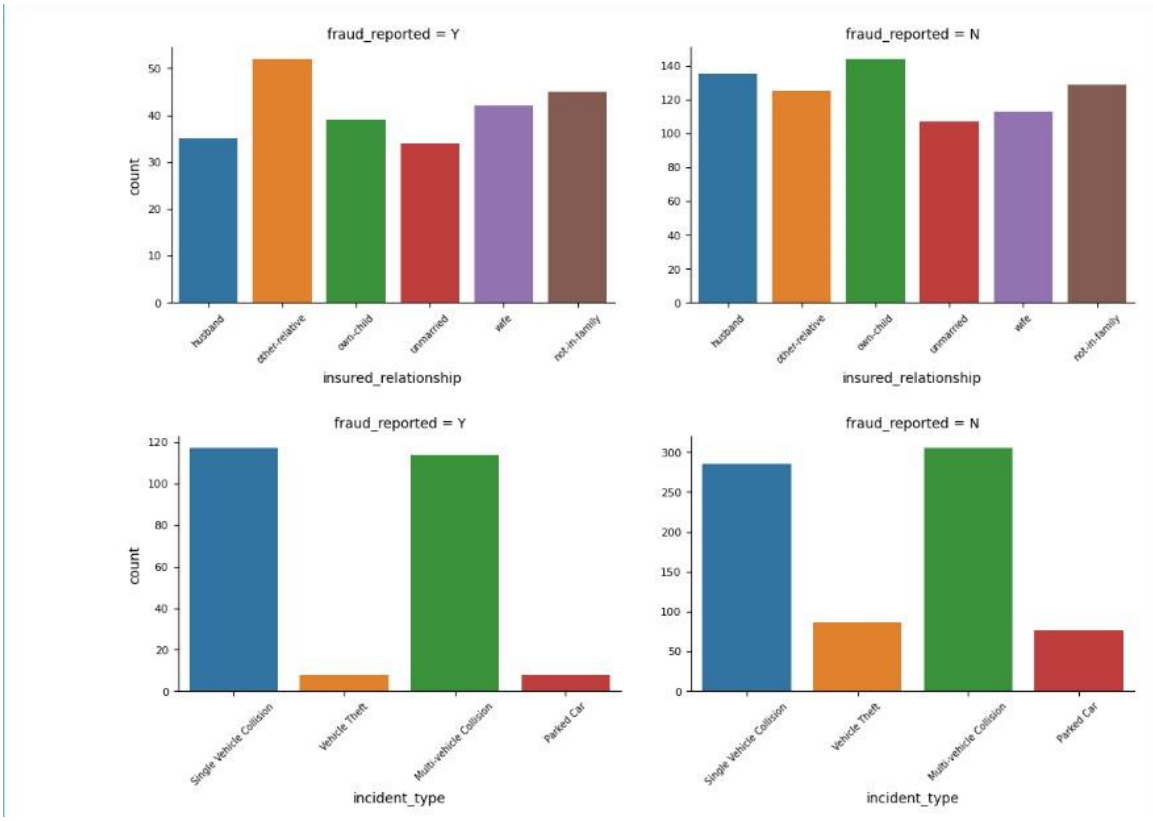
The dataset exhibits class imbalance with a higher proportion of legitimate claims compared to fraudulent ones.

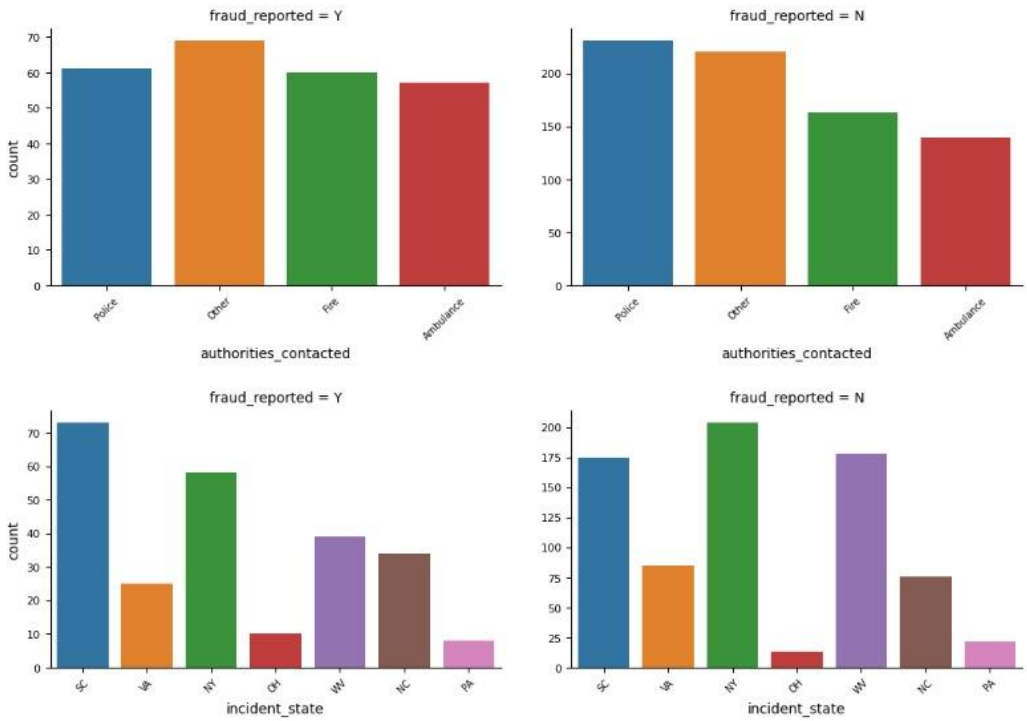


4.4 Bivariate Analysis:

The relationship between categorical features and fraud revealed several insights:

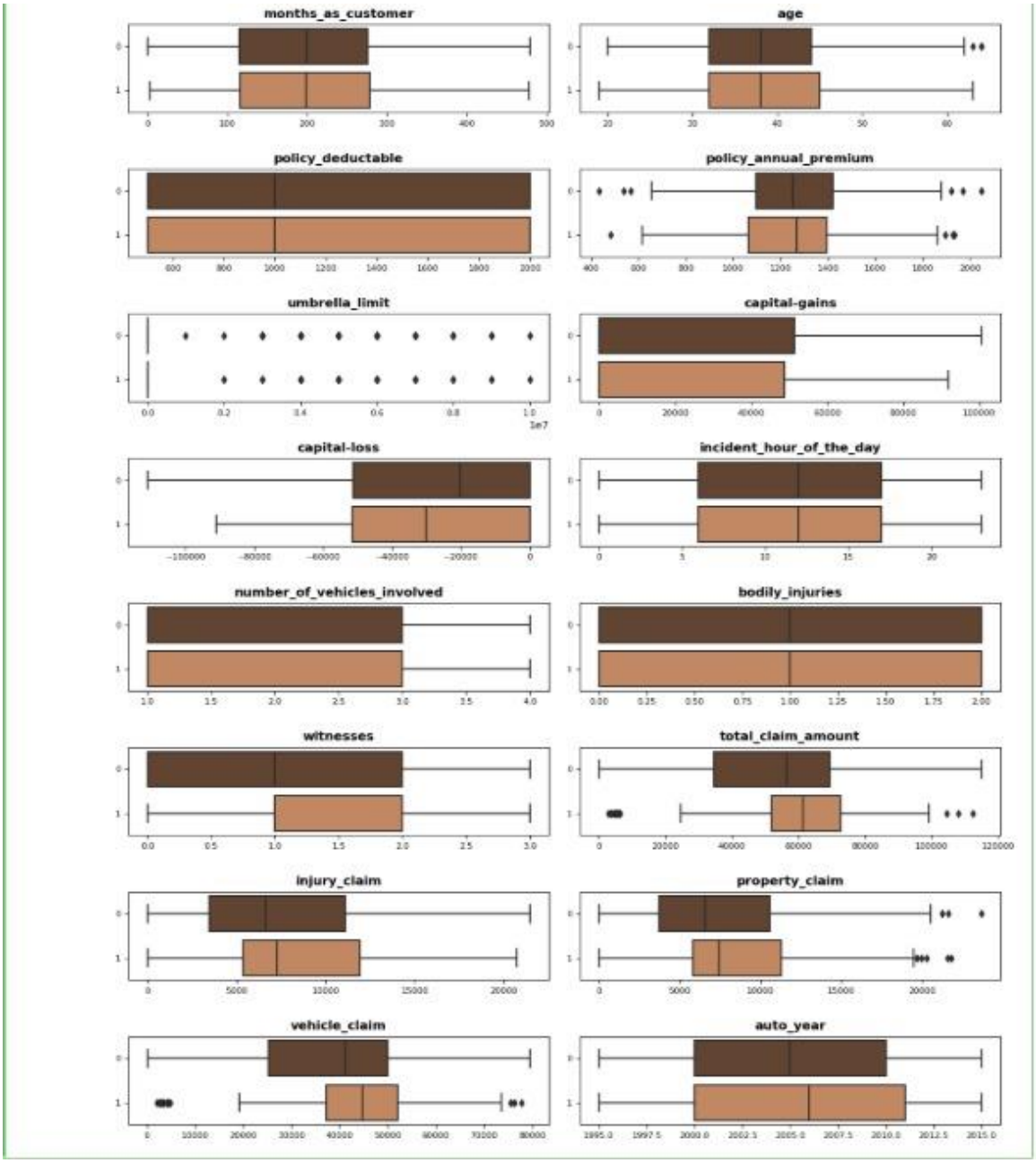






- Gender differences: Male insured individuals showed higher fraud rates than females
- Education impact: High variation in fraud likelihood across education levels, JD showing the highest fraud rate (~45%)
- Incident types: "Vehicle Theft" & Parked Car showed very low fraud rates compared to other incident types
- Authority contacts: Interactions with certain authorities (ambulance, fire) and providing police reports correlated strongly with fraud likelihood

For numerical features, boxplot analysis revealed:



- Higher policy deductibles for fraudulent claims
- Geographic patterns with higher zip codes for fraudulent claims
- Later incident hours for fraudulent claims
- Fewer witnesses for fraudulent claims
- Higher severity of bodily injuries for fraudulent claims

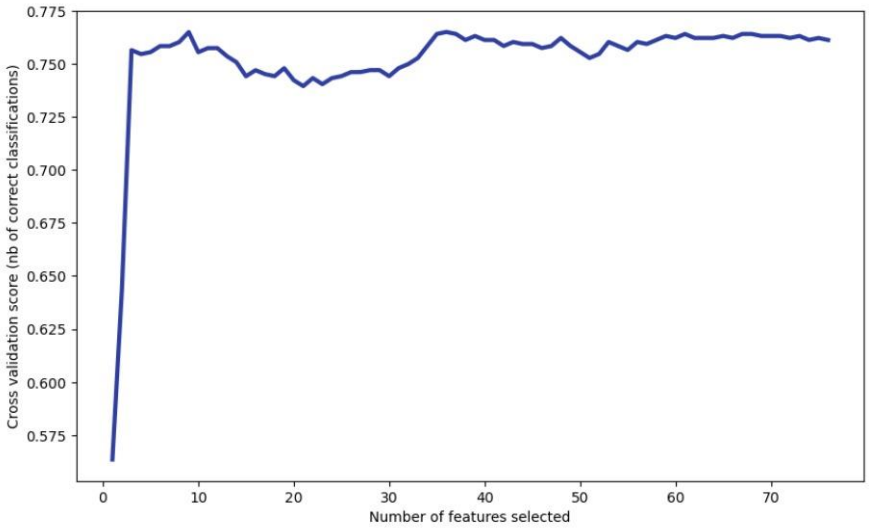
5. Model Building and Evaluation

5.1 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 n-fold cross-validation

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



5.2 Logistic Regression Model

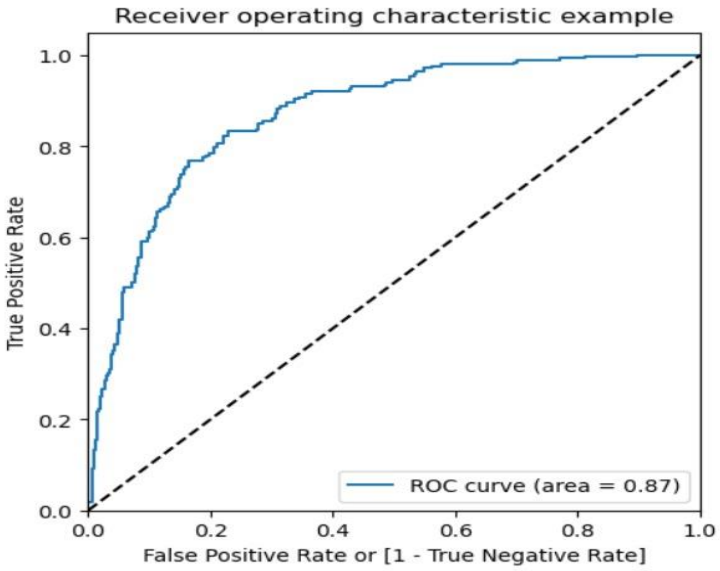
We built a logistic regression model using Statsmodels to enable detailed statistical analysis:

- 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

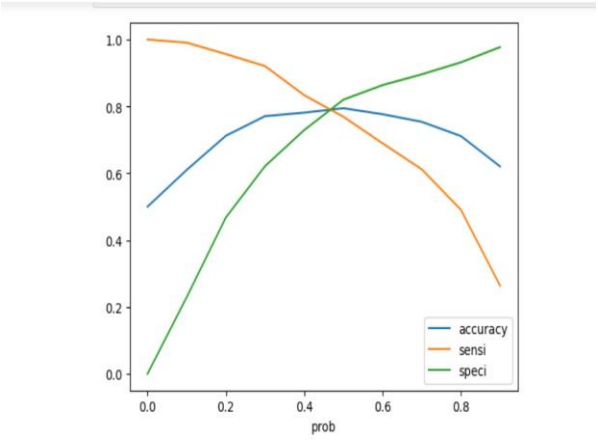
The initial logistic regression model achieved:

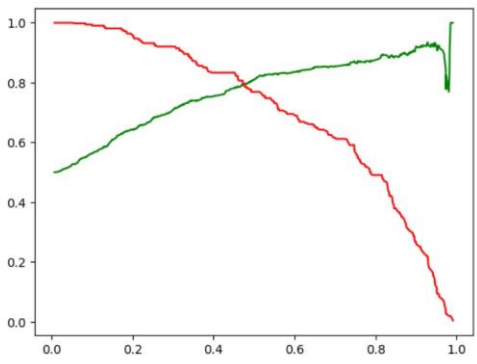
- 79% accuracy on the training set
- 76.89% sensitivity
- 82% specificity
- 81% precision
- 78.91% F1 score

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.



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.





5.3 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

Used grid search for hyperparameter 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

5.4 Validation Performance

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.

6. Model Performance Analysis

Our fraud detection models 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 non-linear relationships, despite more severe overfitting.

7.Methodological Insights

Several aspects of our methodology warrant discussion:

1. **Resampling effects:** The use of RandomOverSampler may have introduced artificial patterns that don't generalize well to new data, potentially contributing to overfitting.
2. **Feature engineering impact:** The creation of domain-specific features (like claim-to-coverage ratio and suspicious timing flags) proved valuable, outweighing many raw variables in predictive power.
3. **Categorical grouping effectiveness:** Grouping categorical variables by risk level successfully reduced dimensionality while maintaining or improving predictive power.
4. **Feature selection timing:** Performing feature selection before hyperparameter tuning may have led to suboptimal feature subsets, as feature importance can change with different model configurations

8. Conclusion

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.

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

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.

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.