Fraudulent Claim Detection

Authors:
Astha Bansal
Ashwini Agalave

SAY NO TO FRAUD

Machine Learning Strategies for Detecting Insurance Claim Fraud

Overview

- **Problem Statement:**

  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. This would minimize financial losses and optimize the overall claims handling process.

- **Business Objective:**

- Global Insure aims to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles.
- The goal is to predict which claims are likely to be fraudulent before they are approved, minimizing financial losses and optimizing the claims handling process.

- **Why to solve this problem:**

  Identify the circumstances which lead to fraudulent claims.

  Reduce reliance on human investigators to determine.

  Correctly identify legal claims to ensure payout to rightful policyholders.

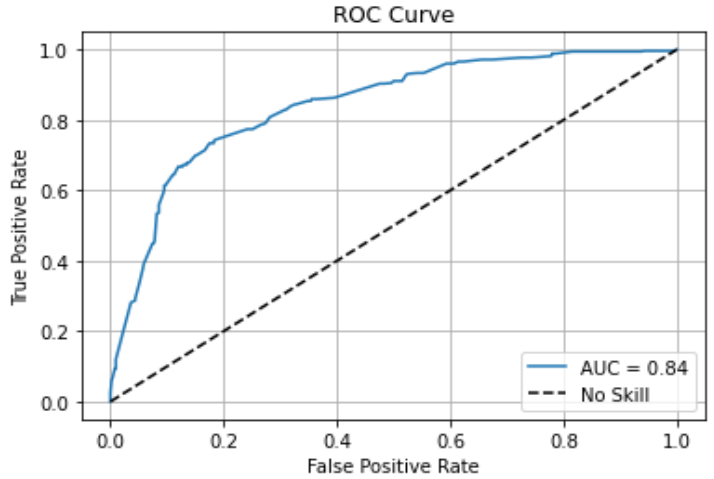
  May also help to determine what type of policyholders the company should avoid.

Methodology

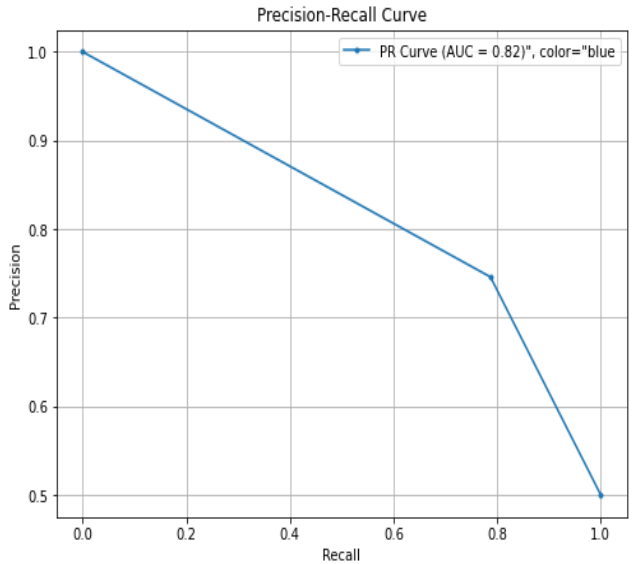
1. **Data Preparation**:
    1. The dataset consists of 40 columns and 1000 rows, with attributes such as claim amounts, customer profiles, and claim types.
    2. Data preprocessing involved cleaning, transforming, and engineering features from raw claim data to ensure effective model learning.
2. **Exploratory Data Analysis (EDA)**:
    1. Univariate and bivariate analyses were conducted to understand the distribution and relationships between features.
    2. Correlation analysis was used to identify multicollinearity among features.
3. **Feature Engineering**:
    1. Resampling techniques like RandomOverSampler were used to address class imbalance.
    2. New features were created from existing ones to enhance the model's ability to capture patterns in the data.

Models and Approaches

1. **Model Building**:
    1. Two models were built: Logistic Regression and Random Forest.
    2. Logistic Regression used RFECV for feature selection, while Random Forest involved hyperparameter tuning using grid search.
2. **Model Evaluation**:
    1. Models were evaluated using metrics such as accuracy, sensitivity, specificity, precision, recall, and F1-score.
    2. Predictions were made on both training and validation data to assess model performance.

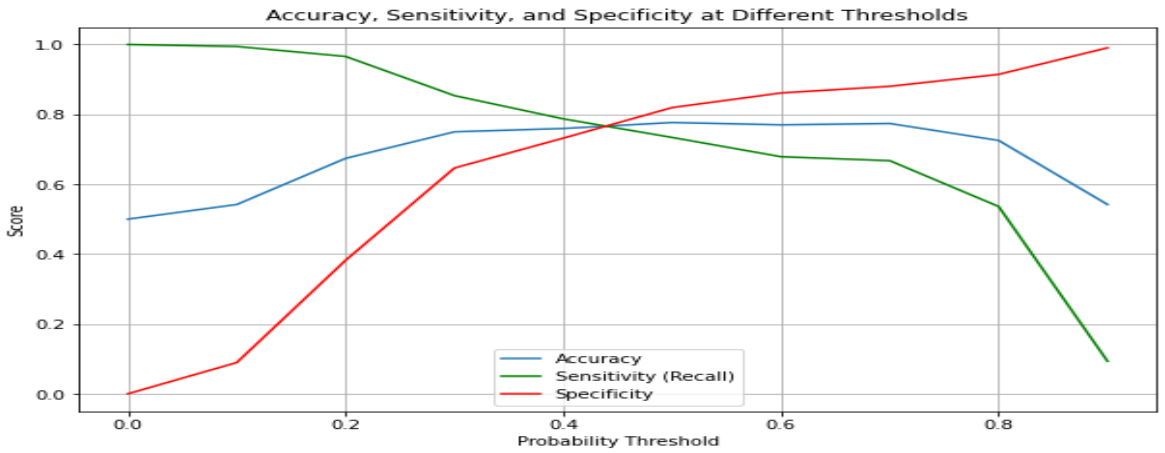


Sensitivity and Specificity tradeoff



📊 Final Model Evaluation Metrics

Metric	Value
Sensitivity (Recall)	0.92
Specificity	0.96
Precision	0.96
Recall	0.92
F1-Score	0.94



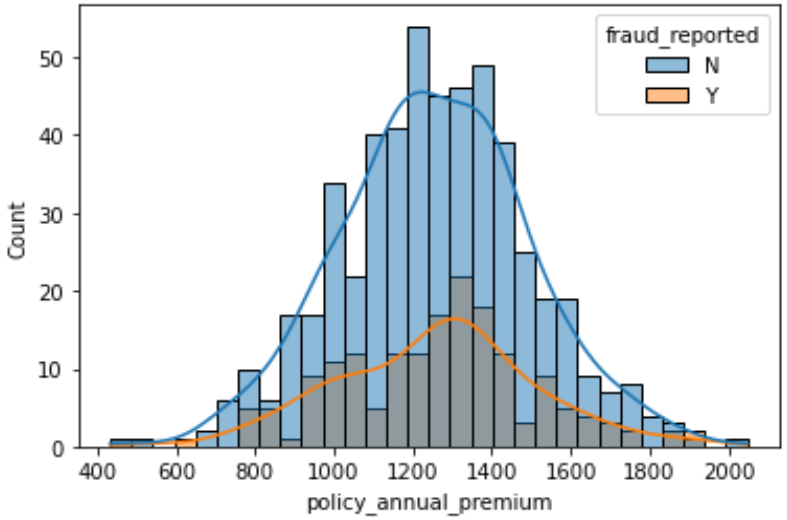
Choosing a optimal cutoff- -the cutoff which maximizes accuracy, sensitivity and specificity.

Model Comparison on Validation Set

Metric	Logistic Regression	Random Forest (Tuned)
Accuracy	0.71	0.78
Recall (Sensitivity)	0.70	0.56
Precision	0.45	0.64
Specificity	0.72	0.88

| F1 Score | 0.55 | 0.59

1. Random Forest (tuned) performs better than logistic regression across in accuracy and F1 score.
2. F1 Score: Balances the trade-off between catching fraud and minimizing false positives.



Summary

A. By following below steps, insurers can leverage historical claim data to build robust models that detect patterns indicative of fraudulent claims. This approach not only helps in minimizing financial losses but also supports efficient and data-driven claim triaging. The report emphasizes the importance of using advanced analytics and machine learning techniques to enhance the fraud detection process:

1. Data Preparation: Data Cleaning, Data Transformation
2. Exploratory Data Analysis (EDA): Univariate Analysis, Bivariate Analysis, Correlation Analysis
3. Feature Engineering: Feature Creation, Resampling
4. Model Building: Logistic Regression and Random Forest
5. Model Evaluation: Performance Metrics, Cross-Validation
6. Insights and Recommendations: Predictive Features, Business Strategies

B. Top predictive features suggested to achieve this objective by random forest:

C. Yes, based on past data, it is possible to predict the likelihood of fraud for an incoming claim using the methodologies outlined.

	Varname	Imp
2	total_claim_amount	0.205215
3	incident_days_since_policy	0.169116
0	policy_annual_premium	0.146011
5	incident_severity_Minor Damage	0.143435
6	incident_severity_Total Loss	0.138651
1	capital-gains	0.088709
7	incident_severity_Trivial Damage	0.053081
9	witnesses_2	0.016434
8	incident_state_WV	0.015532
10	has_Umbrella_Umbrella	0.012660
4	policy_deductable_1000	0.011156

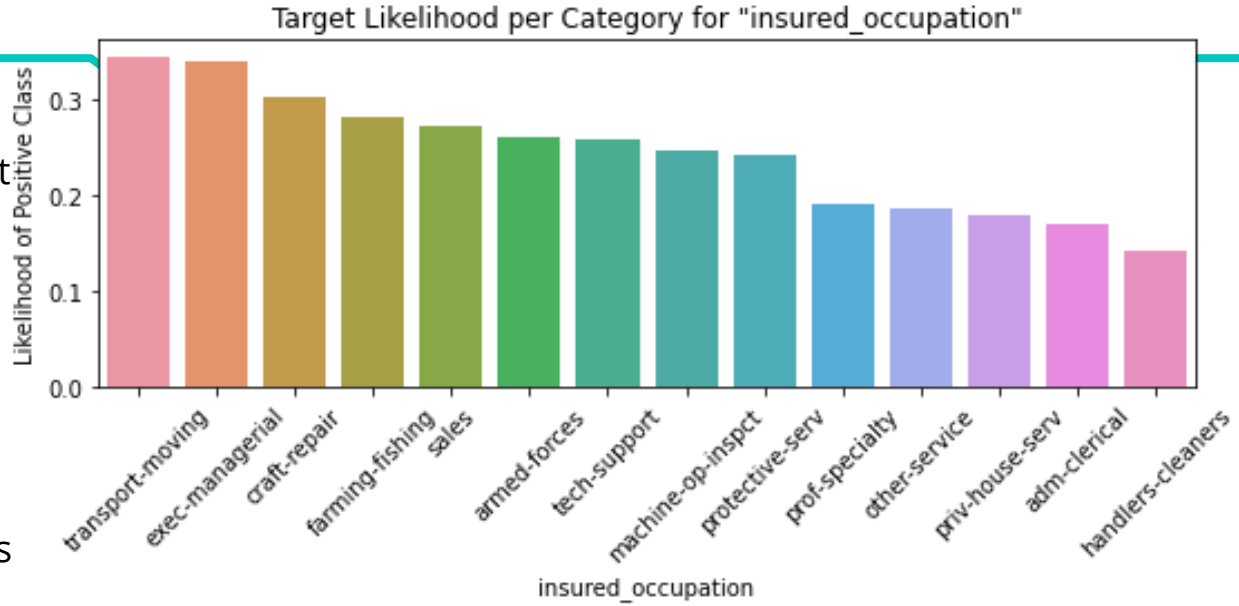
Insights for Business:

- **Claim-related variables** like total_claim_amount are highly influential in detecting fraud.
- **Customer behavior and policy details**(e.g.,*incident_days_since_policy*, *policy premium*, incident_severity) also play a strong role.
- Targeted fraud detection strategies can be developed using these high-importance variables for early and effective intervention.
- **Recommended Model**: Random Forest with Hyperparameter Tuning

1. High generalization to unseen claims
2. Strong fraud detection capability (recall = 0.56)
3. Good balance of precision and recall (F1 = 0.59)

- **Business Impact:**

1. Increases early detection of fraudulent claims
2. Reduces financial losses
3. Supports efficient and data-driven claim triaging
4. Focus on the main features suggested to reach the objective.



Recommendations for Improvement:

1. Continuous Model Refinement
2. Integration with Business Processes
3. Focus on High-Importance Features
4. Leverage Advanced Analytics

- The Random Forest (tuned) model outperforms logistic regression across all metrics, especially in recall and F1 score, which are critical for fraud detection.
- The tuned Random Forest model generalizes well to unseen data, making it the most reliable model for deployment.