FRAUDULENT INSURANCE CLAIM DETECTION PROJECT REPORT

EXECUTIVE SUMMARY

Global Insure, a leading insurance provider, faces significant financial losses due to fraudulent claims, currently identified through inefficient manual inspections. This project leverages historical claim data and advanced machine learning techniques to develop a predictive model for early fraud detection. Using a dataset of 1,000 claims with 40 features, we built and evaluated Logistic Regression and Random Forest models. The Random Forest model outperformed with an accuracy of 80.33% on validation data, identifying key predictive features such as claim amounts, incident severity, and customer tenure. This report outlines our methodology, findings, and recommendations to enhance Global Insure's fraud detection process, potentially reducing financial losses and improving operational efficiency.

INTRODUCTION

Global Insure processes thousands of claims annually, with a notable portion being fraudulent, leading to substantial financial losses. The current manual fraud detection process is slow and often identifies fraud post-payout. This project aims to address the following objectives:

1. Analyze historical claim data to detect patterns indicative of fraud.

2. Identify the most predictive features of fraudulent behavior.

3. Predict the likelihood of fraud for incoming claims using past data.

4. Provide insights to improve the fraud detection process.

We utilized a dataset containing 1,000 rows and 40 columns, including customer profiles, policy details, and incident specifics, to build data-driven solutions.

METHODOLOGY

1. DATA PREPARATION AND CLEANING

1. ***Dataset Overview:*** Loaded a CSV file with 1,000 claims and 40 features (e.g., months_as_customer, total_claim_amount, fraud_reported).
2. ***Cleaning:*** Handled missing values (e.g., replacing "?" with NaN), dropped irrelevant columns (e.g., _c39), and converted data types (e.g., dates to datetime).
3. ***Outcome:*** A clean dataset ready for analysis.

2. EXPLORATORY DATA ANALYSIS (EDA)

Training test split: Split the dataset into 70% training (700 rows) and 30% validation (300 rows).

Key Observations:

- Fraudulent claims (~25%) showed higher claim amounts (vehicle_claim, property_claim).
- Severe incidents (incident_severity_Total Loss) were more associated with fraud.
- Shorter customer tenure (months_as_customer) correlated with fraudulent claims.

3. FEATURE ENGINEERING

1. ***New Features:*** Created vehicle_age (from auto_year), incident_period_of_day (from incident_hour_of_the_day), and buckets for hobbies and auto makes.

2. ***Encoding:*** Applied one-hot encoding to categorical variables (e.g., incident_severity, collision_type).
3. ***Outcome:*** Enhanced dataset with 20+ predictive features.

4. Model Building

Logistic Regression:

1. Iteratively refined using statistical significance (p-values < 0.05).
2. Final model used 27 features, achieving 95.63% accuracy on training data.

Random Forest:

1. Tuned via GridSearchCV with parameters (e.g., max_depth, n_estimators).
2. Selected top 20 features (e.g., vehicle_claim, incident_severity_Total Loss), achieving 95.63% accuracy on training data with balanced resampling.

5. Prediction and Evaluation

Validation Performance:

1. Logistic Regression: 72.67% accuracy, sensitivity 72.97%, specificity 72.57%, F1-score 0.5684.
2. Random Forest: 80.33% accuracy, sensitivity 64.86%, specificity 85.40%, F1-score 0.6194.
3. Key Metrics: Random Forest excelled in accuracy and specificity, making it more reliable for minimizing false positives (legitimate claims flagged as fraud)

FINDINGS

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

1. High claim amounts (vehicle_claim, property_claim, injury_claim) often linked to fraud.
2. Lack of witnesses (witnesses = 0) in severe incidents (incident_severity_Total Loss) is suspicious.
3. Short customer tenure (months_as_customer < 100) frequently associated with fraudulent claims.

Which features are most predictive of fraudulent behavior?

Top Features (Random Forest):

1. Financial: vehicle_claim, property_claim, injury_claim, capital-gains, capital-loss.
2. Incident: incident_severity_Total Loss, incident_severity_Minor Damage, collision_type_Rear Collision.
3. Customer: months_as_customer, hobby_bucket_Mental/Indoor, hobby_bucket_Fitness/Active.
4. Policy: policy_deductable, policy_csl_250/500.

Can we predict the likelihood of fraud for an incoming claim based on past data?

Yes. The Random Forest model assigns a fraud probability score (e.g., a claim with high injury_claim, minor damage, and no witnesses scores high).Also it has validation accuracy , around 80.33%, demonstrating reliable predictive capability.

What insights can be drawn to improve the fraud detection process?

Key Insights:

1. Claims with high financial stakes and minor/total loss severity warrant closer scrutiny.
2. Policies with high deductibles or specific coverage (policy_csl_250/500) are more fraud-prone.
3. Customer behaviors (e.g., hobbies, short tenure) and mainstream vehicle makes correlate with fraud risk.

RECOMMENDATIONS

1. Implement Predictive Model:

1. Deploy the Random Forest model to score incoming claims in real-time, flagging high-probability fraud cases for review.
2. Example: A claim with vehicle_claim > $50,000, incident_severity_Minor Damage, and witnesses = 0 should trigger an alert.

2. Enhance Process Efficiency:

1. Prioritize manual reviews based on model scores, reducing workload and focusing on high-risk claims.
2. Automate low-risk claim approvals to streamline operations.

3. Targeted Investigations:

1. High claim amounts disproportionate to incident severity.
2. Short customer tenure (<100 months).
3. Specific policy traits (e.g., high deductibles, 250/500 coverage).

4. Continuous Improvement:

1. Regularly update the model with new claim data to maintain accuracy.
2. Monitor emerging fraud patterns (e.g., new hobbies or vehicle makes) for feature inclusion.

CONCLUSION

This project successfully developed a data-driven solution to detect fraudulent claims at Global Insure. The Random Forest model, with 80.33% accuracy by focusing on key features like claim amounts, incident severity, and customer profiles, Global Insure can transition from reactive manual inspections to proactive fraud prevention. Implementing these recommendations will optimize the claims handling process, reduce payouts on fraudulent claims, and enhance overall operational efficiency.

TEAM MEMBERS

1. ***Krishnakumar V***
2. ***Kante Shiva Chandra***