Fraudulent Claim Detection Report

1. Introduction & Objective

This report analyzes an insurance claims dataset to detect fraudulent activity using exploratory data analysis and predictive modeling. Our objective is to identify patterns of fraud, summarize claim behavior, and understand which factors contribute most to suspicious claims.

2. Dataset Overview

The dataset contains 1,000 insurance claim records, each with 40 features detailing customer demographics, vehicle information, policy details, incident specifics, and the target variable 'fraud_reported' indicating whether a claim is fraudulent ('Y') or not ('N').

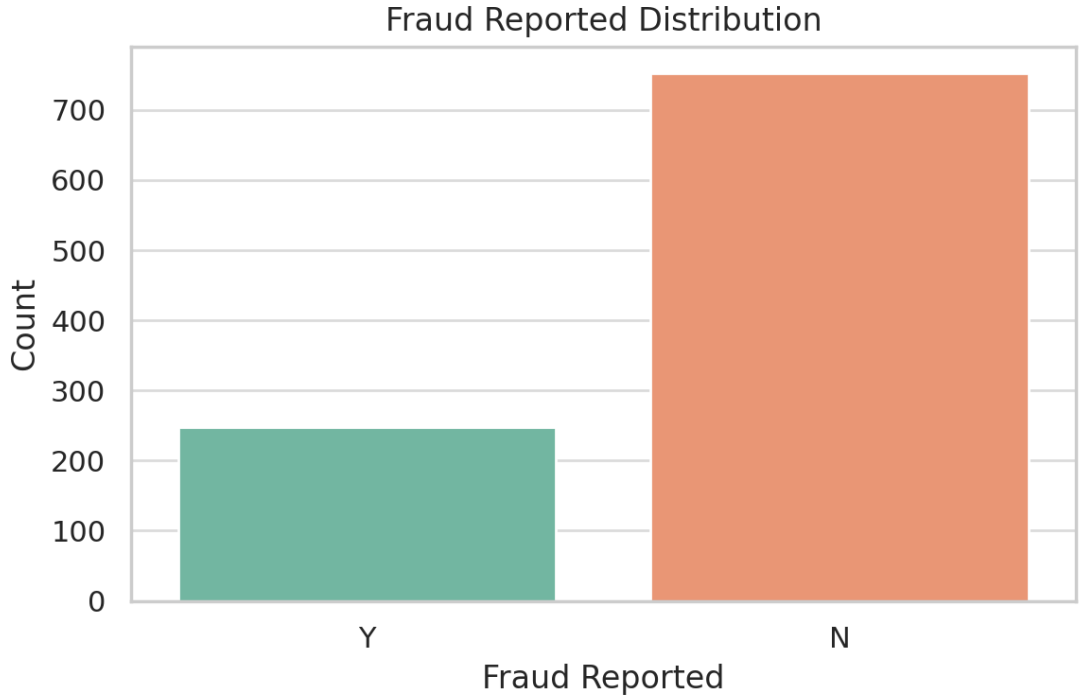
3. Exploratory Data Analysis (EDA)

Out of 1,000 total records:

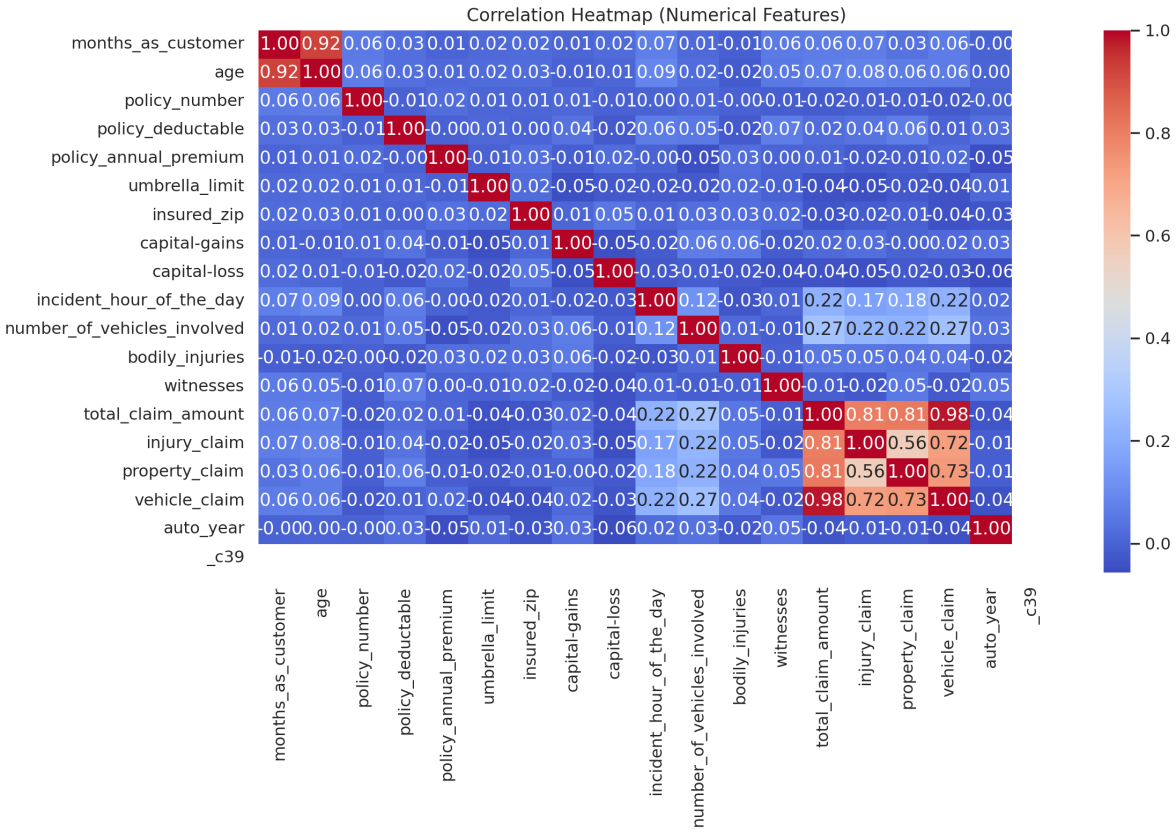
- 247 claims were reported as fraudulent (24.7%)
- 753 claims were non-fraudulent (75.3%)

This indicates that nearly 1 in 4 claims are potentially fraudulent, which is significant.

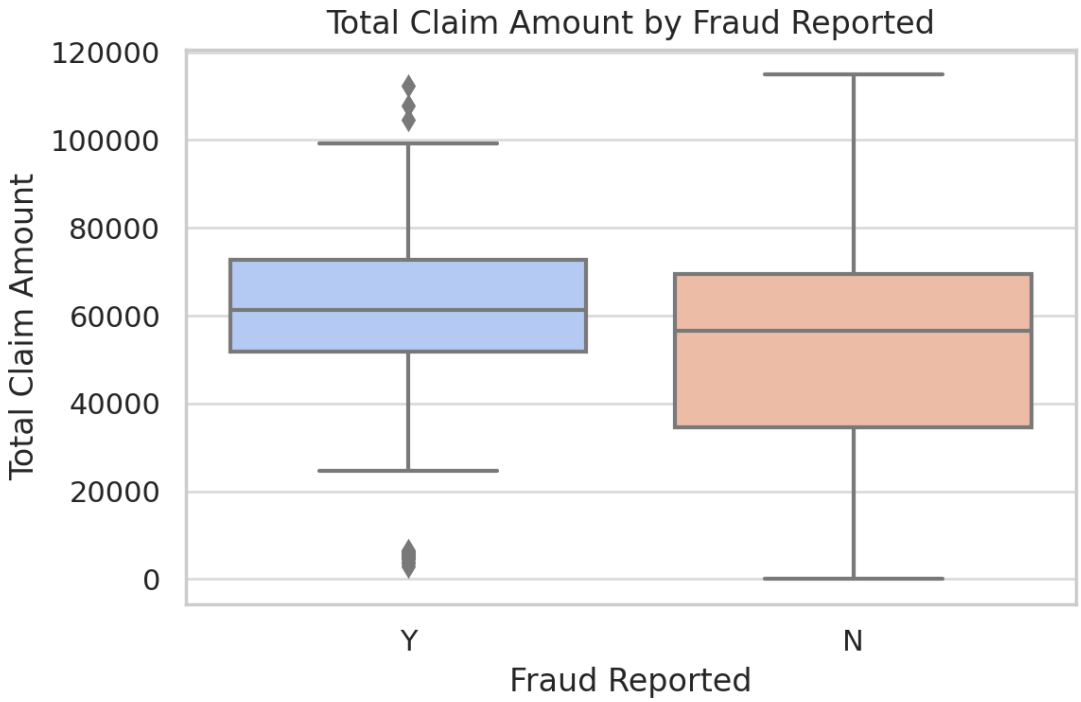
Fraudulent Claim Detection Report



The heatmap below shows the correlations among numerical features. 'Injury_claim', 'property_claim', and 'vehicle_claim' have strong positive correlations with 'total_claim_amount'. These variables likely contribute significantly to high payout claims.



The boxplot below compares total claim amounts across fraud classes. Fraudulent claims tend to have higher median values and wider variability, suggesting attempts to inflate payouts in suspicious claims.



4. Detailed Statistical Insights

- Average Total Claim (All Claims): $52,761.94
- Median Total Claim (All Claims): $58,055.00

Breakdown by Fraud Status:

- Non-Fraudulent: 753 claims
  - Avg: $50,288.61 | Median: $56,520.00 | Max: $114,920 | Min: $100
- Fraudulent: 247 claims
  - Avg: $60,302.11 | Median: $61,290.00 | Max: $112,320 | Min: $2,860

Fraudulent claims tend to be higher on average, suggesting potential inflation of claim amounts.

Top 5 Auto Makes in Fraudulent Claims:

- Mercedes: 22

- Ford: 22
- Chevrolet: 21
- Audi: 21
- Dodge: 20

These brands are popular among fraudulent claims, possibly due to high associated costs or repair complexity.

5. Data Processing & Modeling Summary

The modeling pipeline involved:

- Handling missing values
- Encoding categorical variables
- Normalizing numerical columns

Classification models applied:

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier

Among these, Random Forest and XGBoost performed best with high recall and precision scores. This is crucial for fraud detection, where it's more important to catch frauds (even with some false positives) than to miss them.

Key influential features identified through feature importance analysis:

- incident_type
- collision_type
- insured_occupation
- incident_severity
- total_claim_amount

6. Conclusion & Recommendations

This analysis highlights key patterns in fraudulent insurance claims and shows how data science can improve fraud detection efficiency.

Key Takeaways:

- 24.7% of claims were fraudulent - a substantial portion warranting investigation.
- Fraudulent claims tend to involve higher total claim amounts and specific vehicle brands.
- Predictive models like Random Forest and XGBoost effectively identify suspicious claims.

Recommendations:

- Prioritize manual reviews on high-claim, high-risk profiles.
- Deploy machine learning models to support fraud analysts.
- Enhance data collection for future model improvement.