Fraudulent Claim Detection: A Formal Report

Pradeep, Sai, and Lebo

April 8, 2025

1 Introduction

This report presents a systematic investigation into fraudulent insurance claims using historical claim data. The case study from the Python notebook focuses on extracting valuable insights by leveraging exploratory data analysis (EDA), feature engineering, and predictive modeling techniques. The primary objective is to identify fraud patterns, select highly predictive features, and develop models that enable early detection of fraudulent claims. Our approach encompasses:

- **Data Preparation & Cleaning**: Addressing missing values, invalid entries, and redundant columns.
- **Exploratory Data Analysis (EDA)**: Univariate, bivariate, and correlation analysis.
- **Feature Engineering**: Creating derived features such as `incident_on_weekend` and `injury_claim_ratio`.
- **Model Building**: Training Logistic Regression and Random Forest models, with feature selection using RFECV.

Assumptions made include that all missing values and outliers were handled as described in the notebook and that the domain-specific definitions for the derived features are valid for this dataset.

2 Methodology

2.1 Data Cleaning and Preprocessing

- Removed columns with completely missing values (e.g., `_c39`).
- Dropped rows with missing values in critical columns such as `authorities_contacted`.
- Converted date/time columns (`policy_bind_date` and `incident_date`) to `datetime` objects.
- Applied logical filters to remove rows with negative values in numeric columns.
- Generated new features including:
  - `incident_on_weekend` derived from `incident_date`.
  - `incident_occurred_in_state` by comparing `incident_state` with `policy_state`.
  - Ratios such as `injury_claim_ratio` computed from `injury_claim` and `total_claim_amount`.

2.2 Exploratory Data Analysis (EDA)

Visualizations were generated to analyze both numerical and categorical features. Each visualization is accompanied by key insights that highlight specific variables.

3 Visualizations and Key Insights

3.1 Numerical Variable Distributions

Visualization: Histograms were created for numerical columns including:

- `months_as_customer`, `policy_deductable`, `policy_annual_premium`, and `total_claim_amount`.

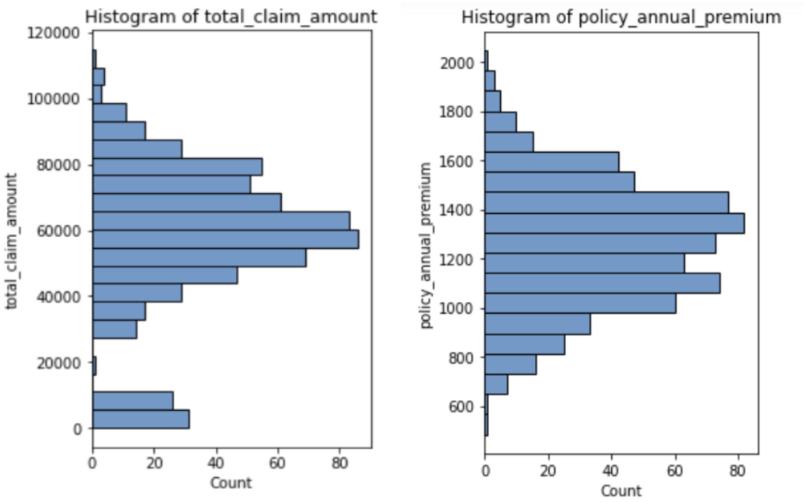


Figure 1: Histogram of Numerical Features

Key Insights:

1. The distribution of `total_claim_amount` reveals that most claims lie below a specific threshold, suggesting potential cutoffs for manual review.
2. `policy_annual_premium` exhibits a consistent spread, indicating uniform pricing strategies.

3.2 Correlation Analysis of Numerical Features

Visualization: A correlation heatmap was generated for numerical variables such as `policy_deductable`, `policy_annual_premium`, and `total_claim_amount`.

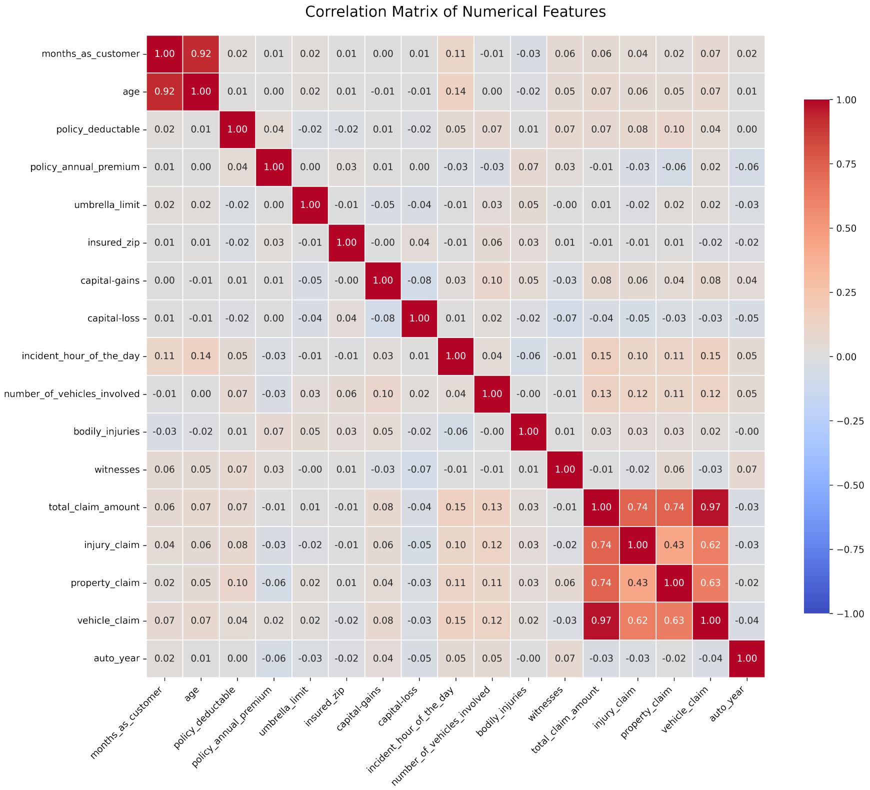


Figure 2: Correlation Matrix of Numerical Features

Key Insights:

1. A strong positive correlation between `policy_annual_premium` and `total_claim_amount` suggests that higher premium policies may be linked to larger claim amounts.
2. The low correlation between `policy_deductable` and `total_claim_amount` indicates that deductibles have limited influence on claim size.

3.3 Class Imbalance

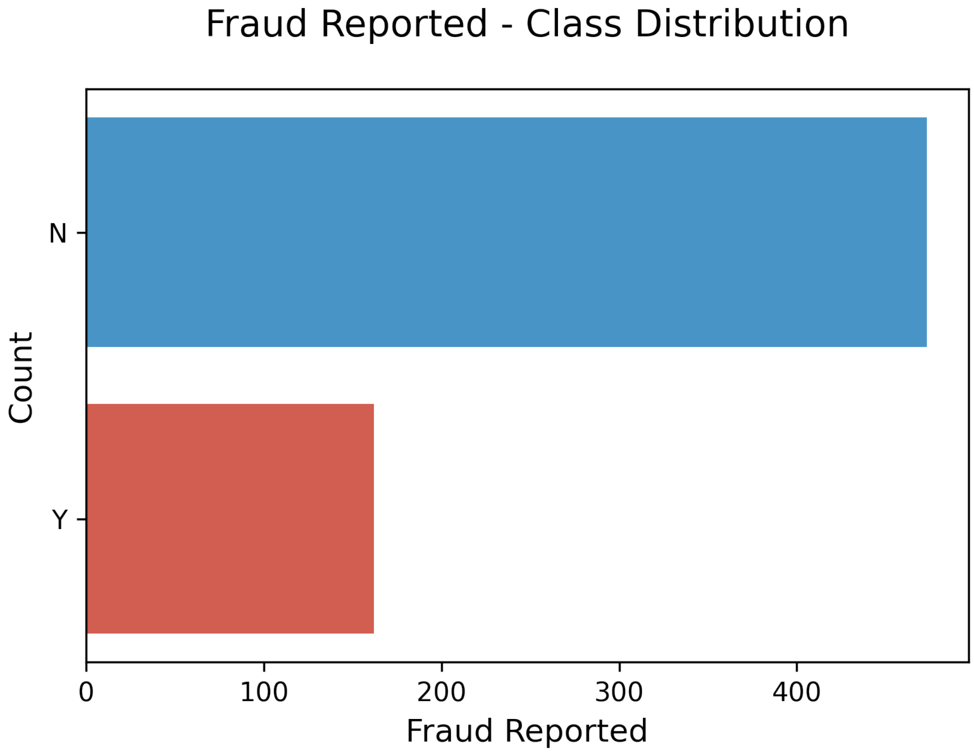


Figure 3: Class Imbalance: The bar graph shows that the number of 'N' entries far exceeds 'Y', underscoring the need for resampling techniques (e.g., Random Over Sampler).

3.4 Target Likelihood Analysis for Categorical Variables (Part 1)

Visualization: Bar plots were generated for categorical features such as `incident_type` and `collision_type`. Each plot presents the fraud likelihood (i.e., probability of fraud as computed from `fraud_reported`) for the top categories.

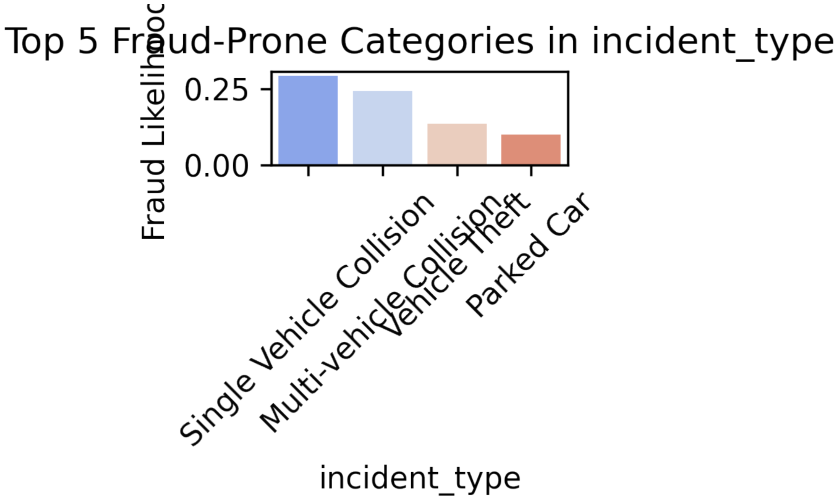


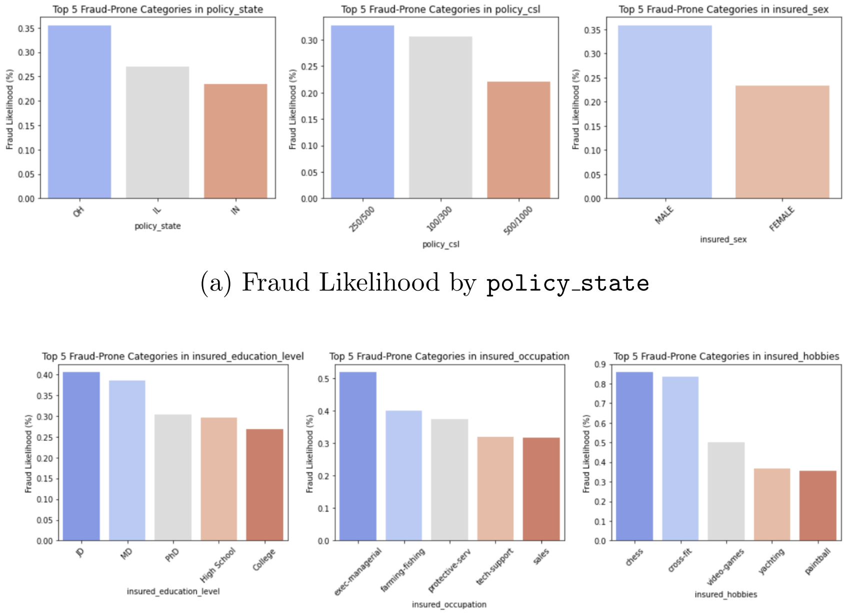
Figure 4: Fraud Likelihood by `incident_type`

Key Insights:

1. The category `Multi-vehicle Collision` in `incident_type` shows a higher fraud likelihood compared to `Single Vehicle Collision`.
2. Within `collision_type`, `Rear Collision` is notably associated with fraudulent claims.

3.5 Target Likelihood Analysis for Categorical Variables (Part 2)

Visualization: Using subfigures to display two related plots for categorical features `policy_state` and `insured_hobbies`.



(a) Fraud Likelihood by `policy_state`

(b) Fraud Likelihood by `insured_hobbies`

Figure 5: Fraud Likelihood Analysis for Selected Categorical Variables

Key Insights:

1. Analysis of `policy_state` suggests that claims in OH exhibit a higher fraud likelihood compared to other states.

2. Among `insured_hobbies`, the hobby `chess` is associated with a higher likelihood of fraud.

3.6 Target Likelihood Analysis for Numerical Variables

Visualization: A boxplot was generated for the numerical variable `capital_loss`.

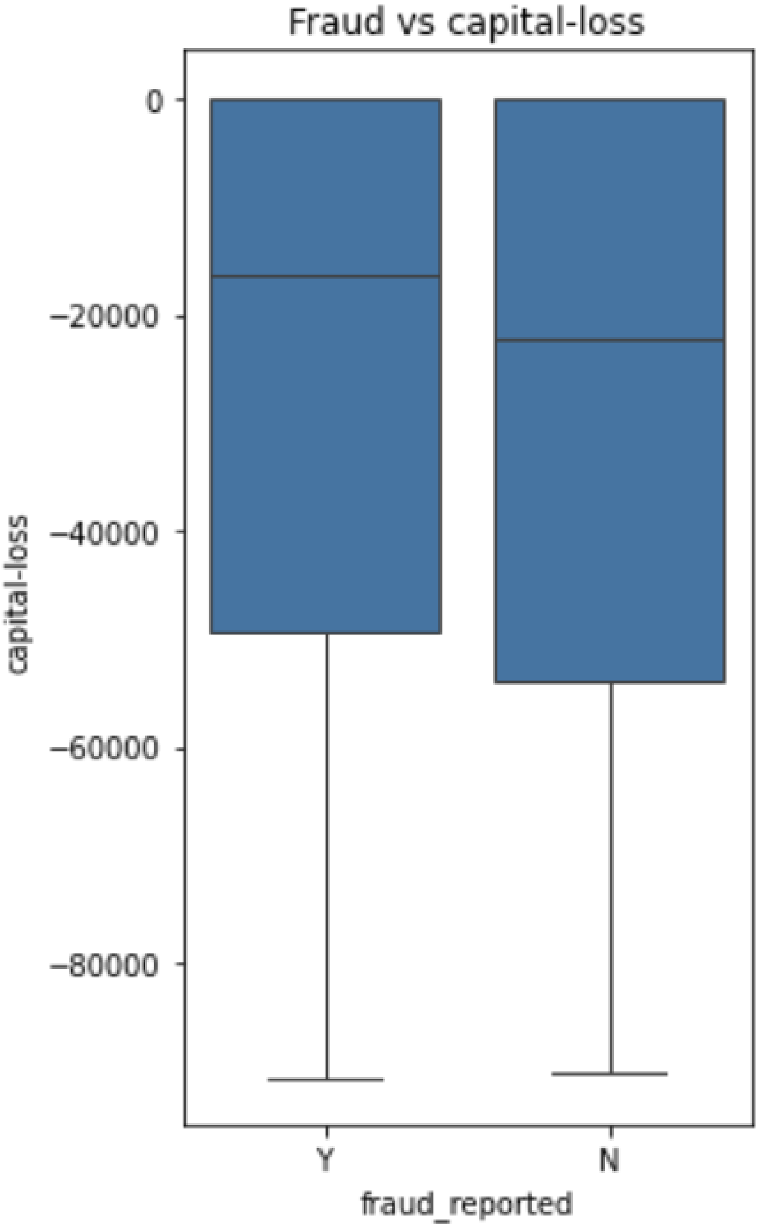


Figure 6: Fraud Likelihood by `capital_loss`

Key Insights:

1. The boxplot indicates that major capital losses are frequently associated with fraudulent claims.

4 Modeling and Feature Selection

4.1 Feature Selection

Using Recursive Feature Elimination (RFE), fifteen highly predictive variables were retained. Notable features include:

- `incident_occurred_in_state`
- `insured_occupation_handlers-cleaners`
- `insured_hobbies_board-games`
- `incident_severity_Minor Damage`
- `age_category_young`

4.2 Model Building

Two models were developed and evaluated:

1. **Logistic Regression:** This model provided probability scores for classifying claims, with cutoff optimization performed to maximize performance.

2. **Random Forest:** This ensemble method not only classified claims but also provided feature importance metrics, reinforcing the predictive power of the selected features.

5 Recommendations

Based on the analysis and model insights, the following actionable recommendations are made:

- **Early Fraud Flagging:** Incorporate the logistic regression model into the claims processing pipeline to flag high-risk claims (e.g., those with elevated fraud likelihood as indicated by `incident_occurred_in_state` and `insured_hobbies_board-games`) for manual review.
- **Enhanced Data Collection:** Improve data capture for critical variables (e.g., ensure complete records for `collision_type` and `incident_severity`) to boost model performance.
- **Periodic Model Updates:** Continuously update the models with new data and re-calibrate probability thresholds to reflect emerging fraud patterns.

6 Conclusion

This report outlines a comprehensive approach to fraudulent claim detection through robust data cleaning, targeted feature engineering, and rigorous model development. Visualizations revealed actionable insights—such as the strong link between `policy_annual_premium` and `total_claim_amount`, and the elevated fraud likelihood in OH and among insured individuals with the hobby `chess`. The recommendations proposed are realistic and actionable, aiming to integrate advanced analytics into the fraud detection process for improved efficiency and accuracy.