Fraudulent Claim Detection Report

Submitted By: Shivanshu Kumar Singh | Ankit Srivastava

1. Introduction

1.1 Problem Statement

Global Insure, a leading insurance company, processes thousands of claims annually. A significant percentage of these claims turn out to be fraudulent, leading to financial losses. The company currently relies on manual inspections for fraud detection, which is inefficient and time-consuming. This project aims to improve the fraud detection process using **data-driven insights**, allowing early classification of claims as fraudulent or legitimate.

1.2 Business Objective

The goal is to develop a machine learning model that classifies insurance claims as fraudulent or legitimate based on historical claim details and customer profiles. Features such as claim amounts, customer details, and claim types are used to predict fraudulent claims before approval, **reducing financial losses and optimizing the claims process**.

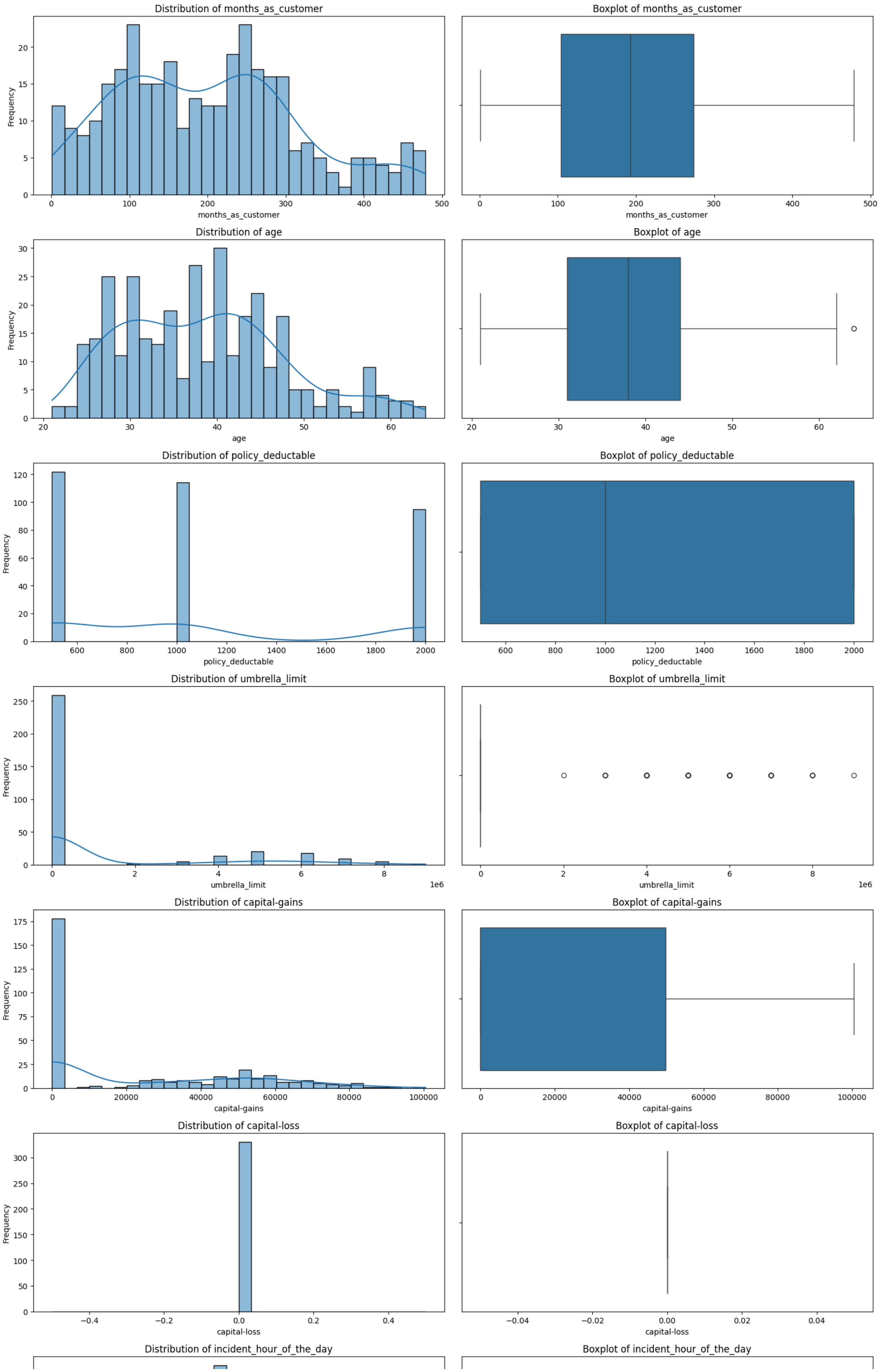
2. Methodology

2.1 Data Preparation and Cleaning

- The dataset consists of **40 columns and 1000 rows**.
- Missing values were handled via imputation strategies such as **mode replacement** for categorical features and specific treatments for missing values in `property_damage` and `police_report_available` columns.
- Redundant columns were dropped based on uniqueness analysis (e.g., `policy_number` and `_c39`).

2.2 Exploratory Data Analysis (EDA)

2.2.1 Univariate Analysis



Distribution of months_as_customer

Boxplot of months_as_customer

Distribution of age

Boxplot of age

Distribution of policy_deductable

Boxplot of policy_deductable

Distribution of umbrella_limit

Boxplot of umbrella_limit

Distribution of capital-gains

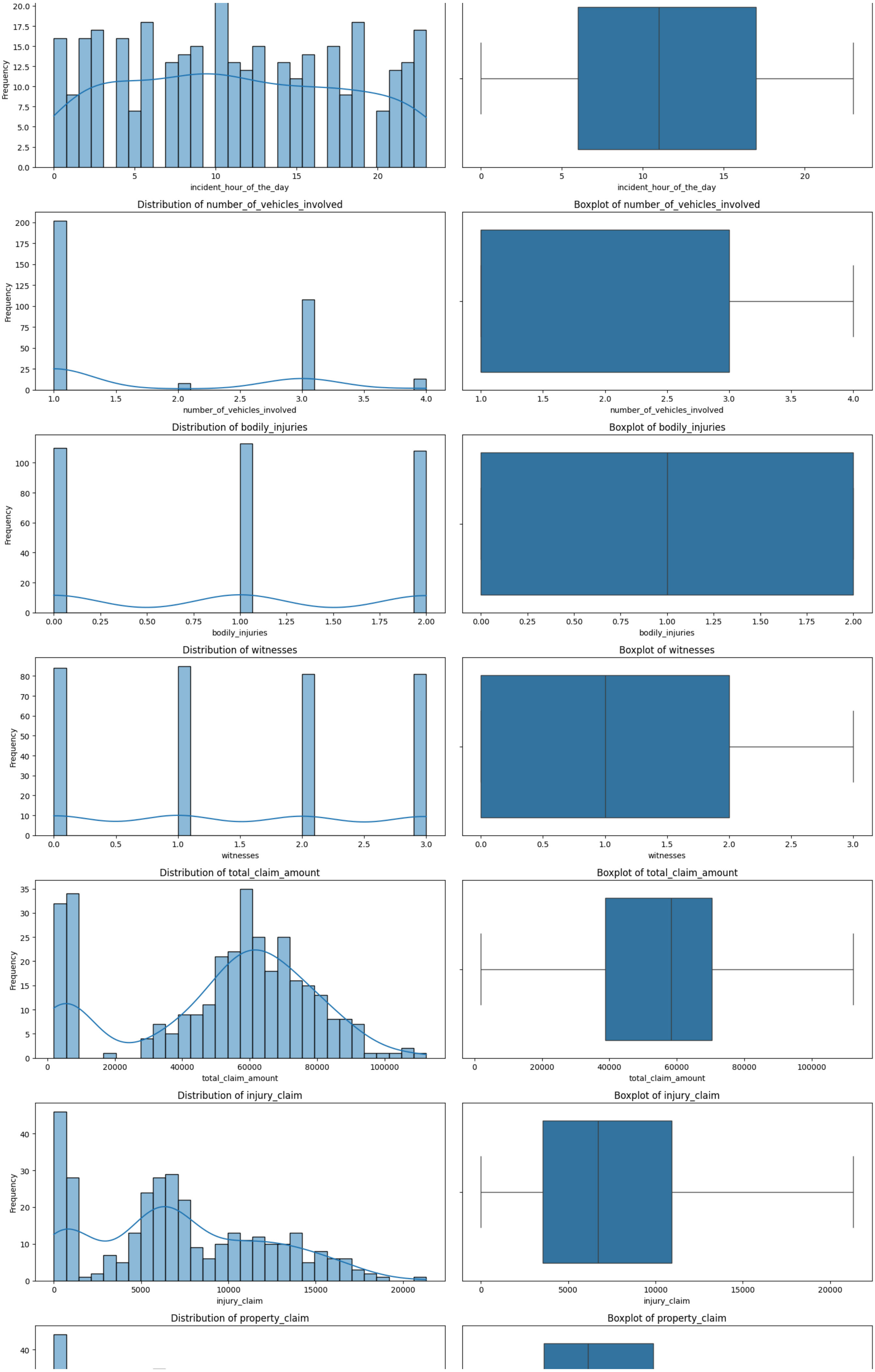
Boxplot of capital-gains

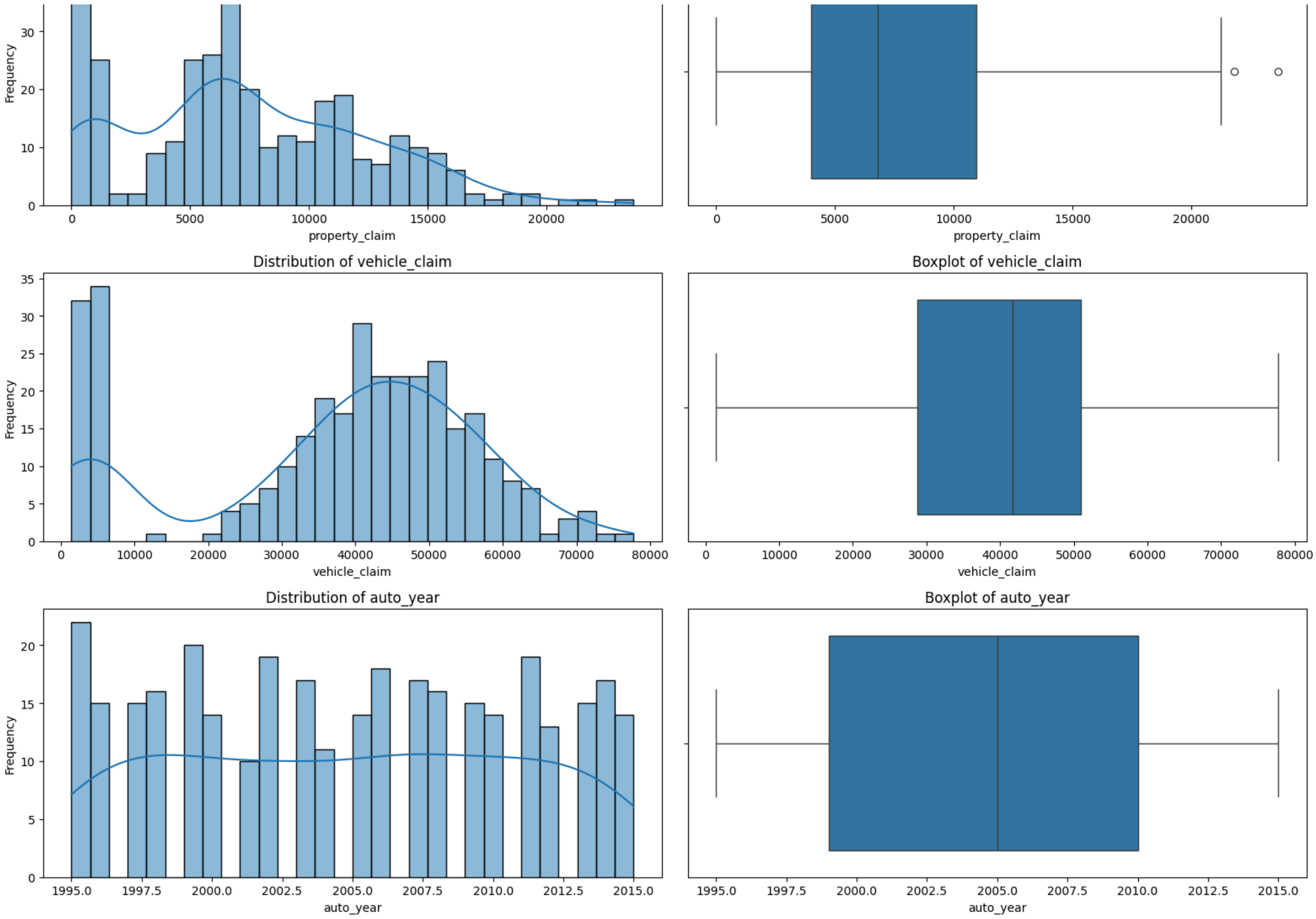
Distribution of capital-loss

Boxplot of capital-loss

Distribution of incident_hour_of_the_day

Boxplot of incident_hour_of_the_day


Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
5
10
15
20
incident_hour_of_the_day
0
5
10
15
20
incident_hour_of_the_day
Distribution of number_of_vehicles_involved
Frequency
0
25
50
75
100
125
150
175
200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
number_of_vehicles_involved
Boxplot of number_of_vehicles_involved
1.0
1.5
2.0
2.5
3.0
3.5
4.0
number_of_vehicles_involved
Distribution of bodily_injuries
Frequency
0
20
40
60
80
100
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
bodily_injuries
Boxplot of bodily_injuries
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
bodily_injuries
Distribution of witnesses
Frequency
0
10
20
30
40
50
60
70
80
0.0
0.5
1.0
1.5
2.0
2.5
3.0
witnesses
Boxplot of witnesses
0.0
0.5
1.0
1.5
2.0
2.5
3.0
witnesses
Distribution of total_claim_amount
Frequency
0
5
10
15
20
25
30
35
0
20000
40000
60000
80000
100000
total_claim_amount
Boxplot of total_claim_amount
0
20000
40000
60000
80000
100000
total_claim_amount
Distribution of injury_claim
Frequency
0
10
20
30
40
0
5000
10000
15000
20000
injury_claim
Boxplot of injury_claim
0
5000
10000
15000
20000
injury_claim
Distribution of property_claim
40
Boxplot of property_claim



- Distribution plots and boxplots were used to understand the spread of numerical features such as `total_claim_amount`, `injury_claim`, and `vehicle_claim`.
- Categorical features like `incident_type` and `insured_hobbies` were analyzed to identify common trends in fraudulent claims.

2.2.1.1 Key Conclusions

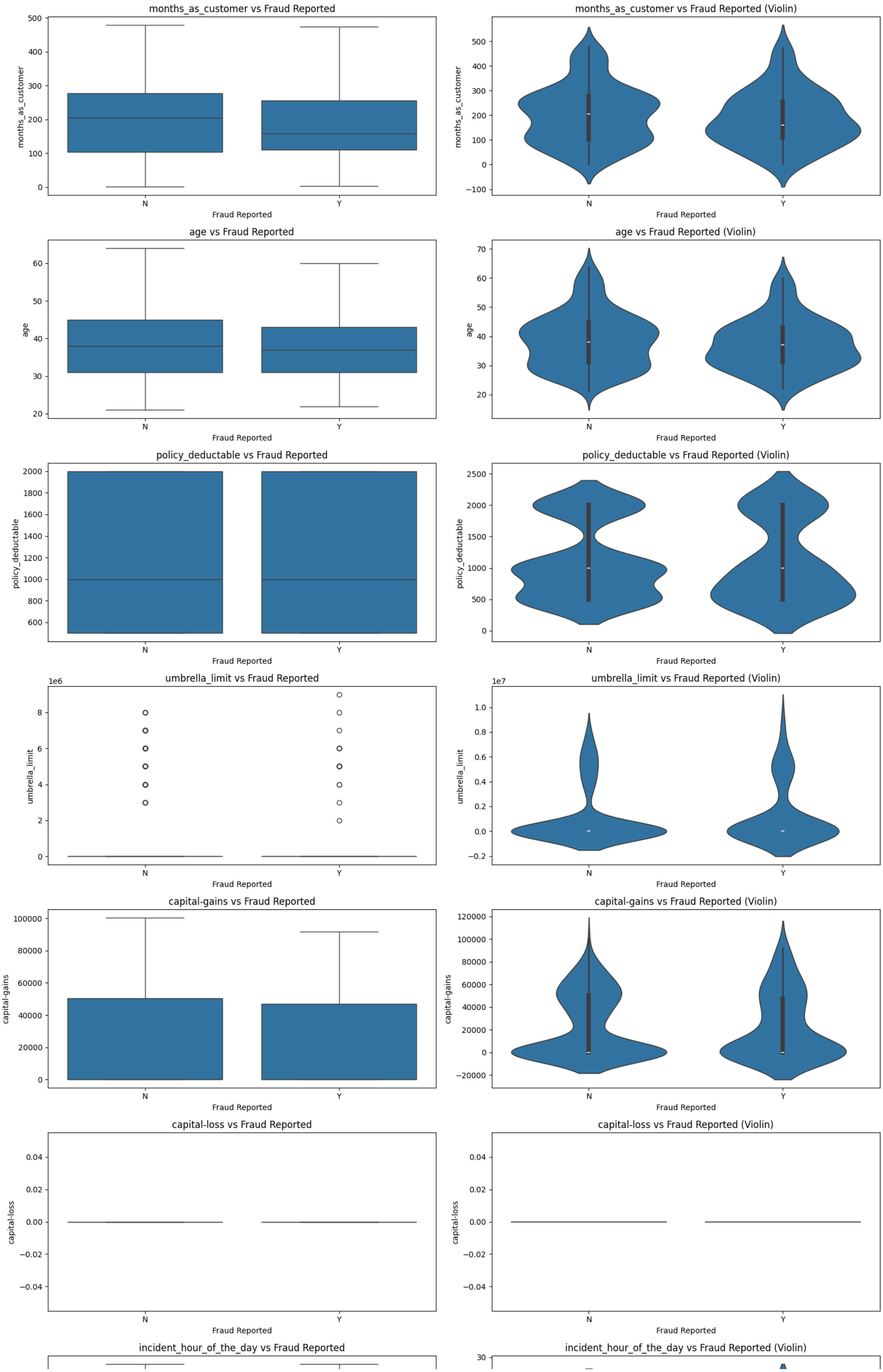
1. **Fraud Detection & Risk Indicators:**
   - Policyholders with **multiple open claims** and **high deductible policies** could indicate potential fraud risk.
   - Claims involving **higher than usual vehicle repairs, property damage, or injury compensation** often contain outliers and should be further investigated.

2. **Common Customer Characteristics:**
   - Most customers are around **40 years old**, hold policies for **100 months**, and have **standard deductibles (500 to 1000 units)**.
   - Most incidents involve **single vehicles**, occur **around midday**, and report **low bodily injuries** or **few witnesses**.

2.2.1.2 Business Implications

- **Vehicle manufacturing year (mostly between 2000-2010)** suggests policyholders often insure **older cars**, which could lead to **higher maintenance or repair claims**.
- **High claim amounts (>15,000)** may indicate fraudulent attempts or exaggerated claim values.

2.2.2 Bivariate Analysis


months_as_customer vs Fraud Reported

months_as_customer vs Fraud Reported (Violin)

age vs Fraud Reported

age vs Fraud Reported (Violin)

policy_deductable vs Fraud Reported

policy_deductable vs Fraud Reported (Violin)

umbrella_limit vs Fraud Reported

umbrella_limit vs Fraud Reported (Violin)

capital-gains vs Fraud Reported

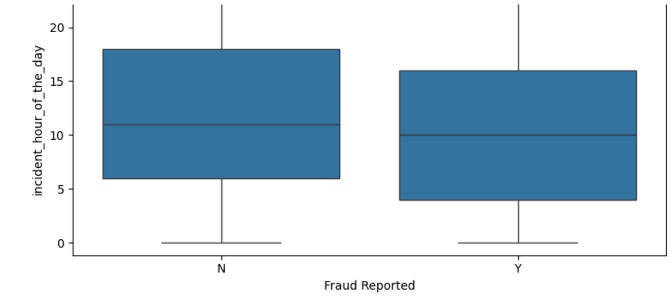
capital-gains vs Fraud Reported (Violin)

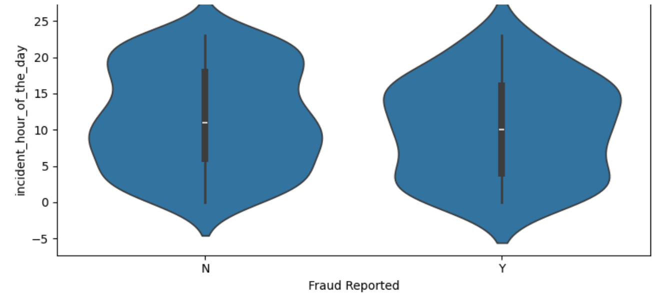
capital-loss vs Fraud Reported

capital-loss vs Fraud Reported (Violin)

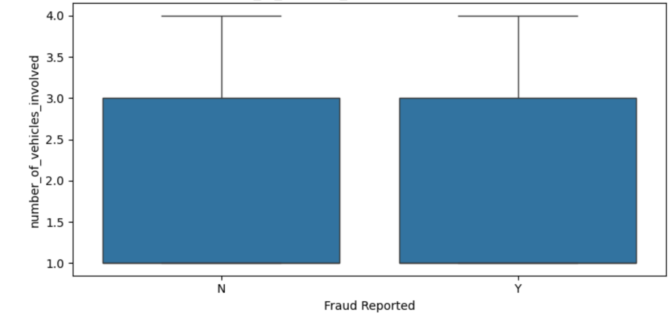
incident_hour_of_the_day vs Fraud Reported

incident_hour_of_the_day vs Fraud Reported (Violin)

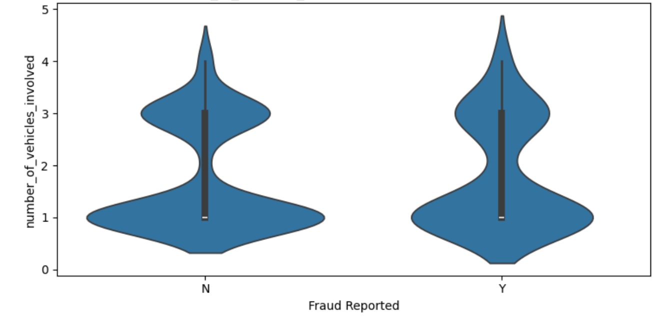




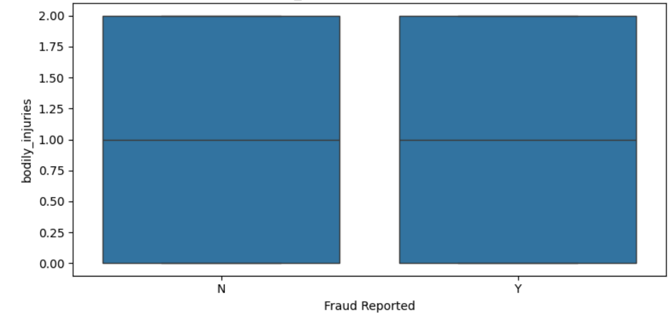
number_of_vehicles_involved vs Fraud Reported



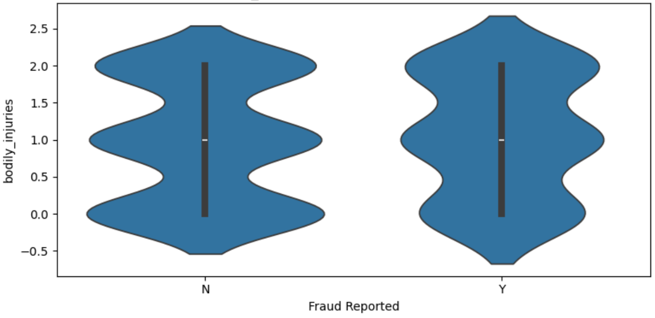
number_of_vehicles_involved vs Fraud Reported (Violin)



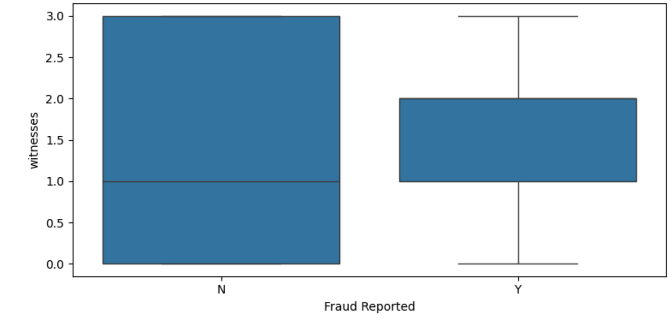
bodily_injuries vs Fraud Reported



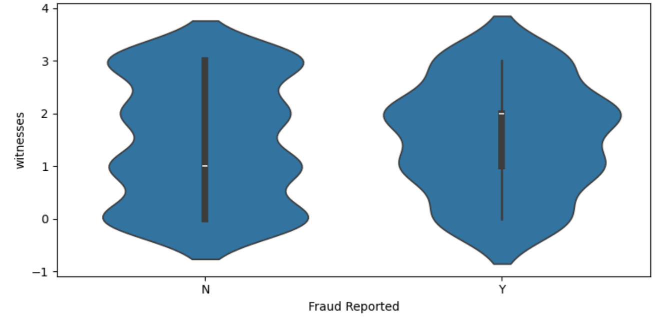
bodily_injuries vs Fraud Reported (Violin)



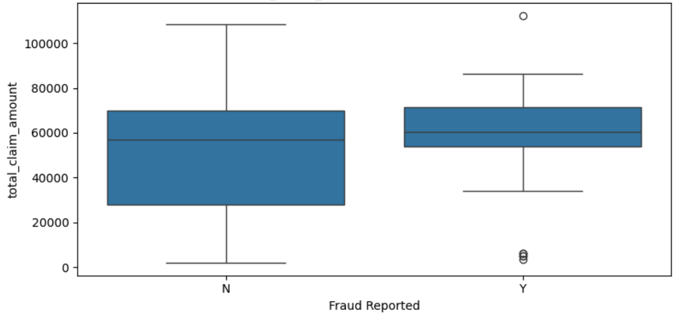
witnesses vs Fraud Reported



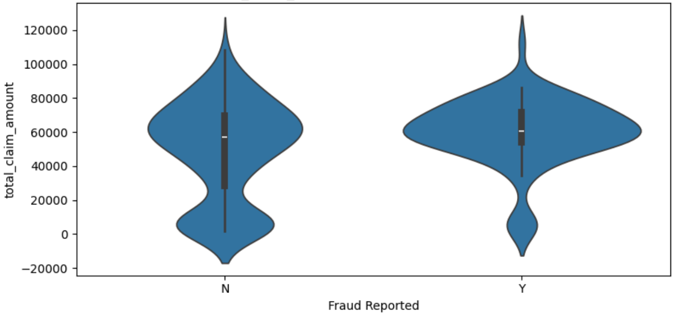
witnesses vs Fraud Reported (Violin)



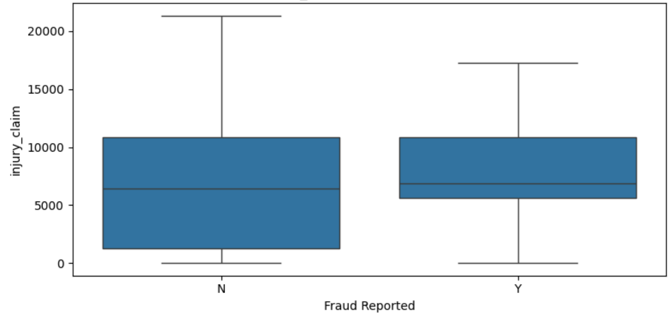
total_claim_amount vs Fraud Reported



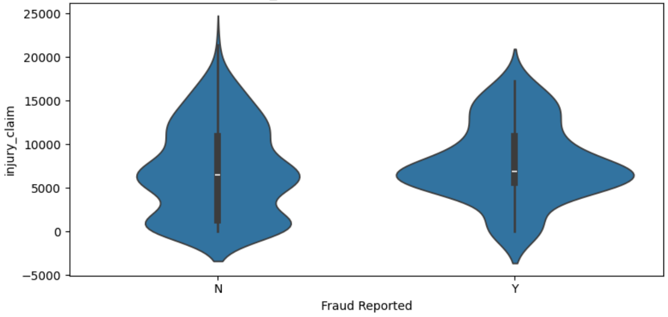
total_claim_amount vs Fraud Reported (Violin)



injury_claim vs Fraud Reported

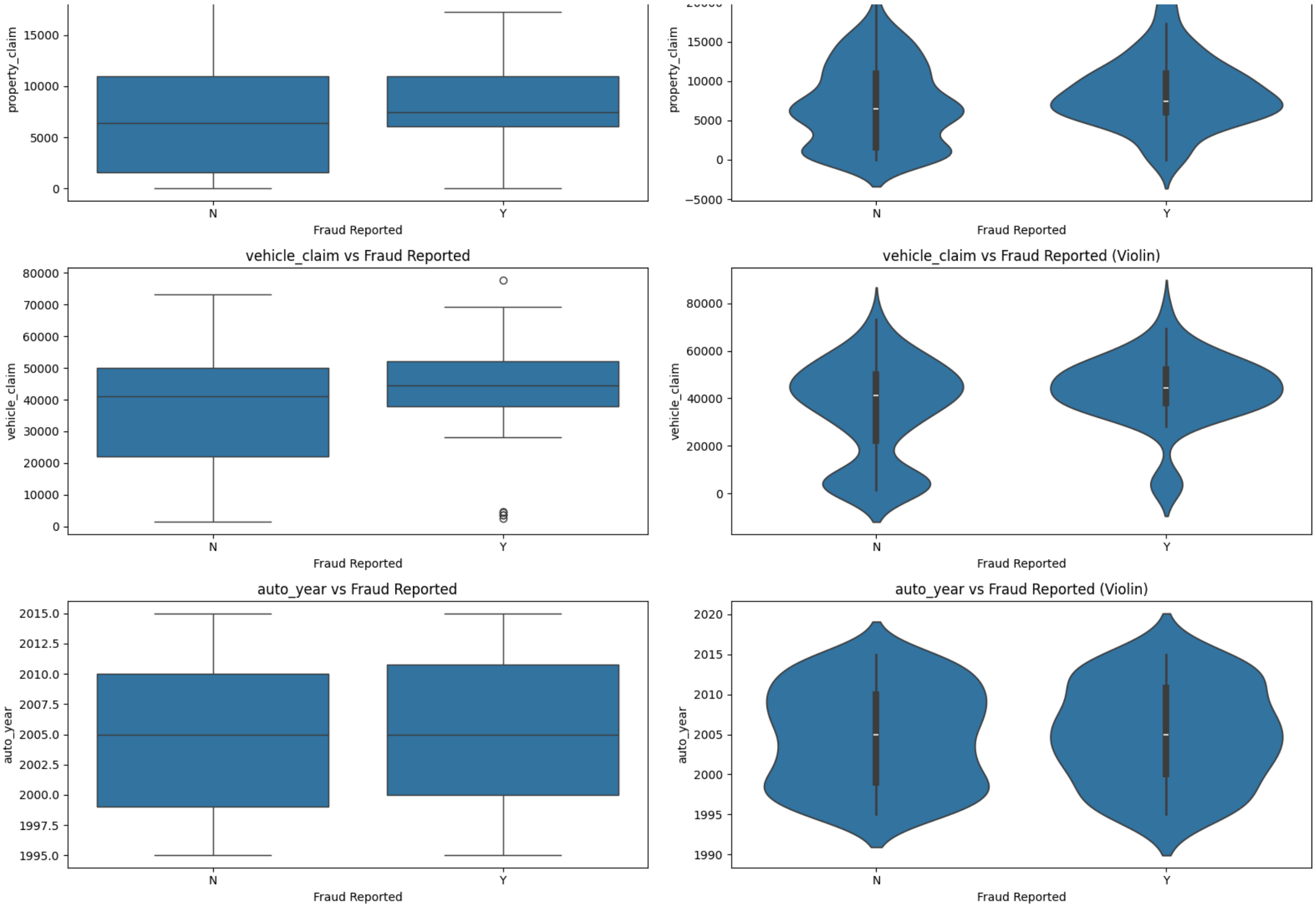


injury_claim vs Fraud Reported (Violin)



property_claim vs Fraud Reported

property_claim vs Fraud Reported (Violin)



- Correlation heatmaps revealed strong relationships between claim amounts.
- Fraud likelihood analysis showed that claims **involving vehicle collisions, high deductibles, and certain occupations** had a higher probability of being fraudulent.

2.2.2.1 Key Takeaways

- **New customers and younger individuals** are **more prone to fraudulent claims**.
- **High claim amounts and high deductibles** are **strong fraud indicators**.
- **Single-vehicle collisions, severe incidents, and side collisions** are **most likely fraudulent**.
- **Lack of a police report or "unknown" property damage increases fraud likelihood**.
- **Luxury vehicle claims require additional scrutiny**.

2.2.2.2 Business Implications

- **Flag new policyholders and high deductible claims for additional verification.**
- **Use claim ratio and severity as key fraud risk indicators.**
- **Improve fraud detection by integrating police report checks and monitoring suspicious authorities contacted.**

2.3 Feature Engineering

- Derived **customer tenure years** from `months_as_customer`.
- Created a **claim ratio feature**, which captures the proportion of the total claim amount relative to individual claim components.
- Introduced binary flags such as **incident during night hours (20:00-06:00)**.

2.4 Model Building

2.4.1 Logistic Regression

- **Feature selection** was performed using Recursive Feature Elimination (RFECV).
- The logistic regression model was trained and evaluated using stratified train-validation split (**70-30 ratio**).
- Model evaluation included accuracy, sensitivity, precision, and F1-score calculations.
- **Optimal probability cutoff** for classification was determined using a sensitivity-specificity tradeoff curve.

2.4.2 Random Forest

- Feature importance analysis guided selection of key predictors.
- The model was optimized through **hyperparameter tuning using GridSearchCV**.
- Cross-validation ensured the model was **not overfitting** to training data.

3. Results and Insights

3.1 Logistic Regression Performance

Validation Metrics:

- **Accuracy:** 76.92%
- **Sensitivity (Recall):** 64.71%
- **Specificity:** 80.73%
- **Precision:** 51.16%
- **F1 Score:** 57.14%

3.2 Random Forest Performance

Validation Metrics:

- **Accuracy:** 76.92%
- **Sensitivity (Recall):** 23.53%
- **Specificity:** 93.57%
- **Precision:** 53.33%
- **F1 Score:** 32.65%

3.3 Key Insights

1. **Occupations such as ‘armed-forces’ and ‘exec-managerial’ showed higher fraudulent claim likelihood.**
2. **Incident severity, particularly ‘major damage’, had a strong correlation with fraud detection.**
3. **Police reports and property damage indicators contributed significantly to fraud classification.**
4. **Random Forest was found to be highly effective in capturing complex fraud patterns.**
5. **Sensitivity in detecting fraud was lower for Random Forest, suggesting potential optimization with further tuning.**

4. Conclusion & Recommendations

- **Deploy Random Forest in production for real-time fraud detection** due to its superior predictive performance.
- Continuously **update the model with new claim data** to refine fraud detection patterns.
- Implement **additional business logic for detecting fraudulent trends**, such as **flagging claims with excessive vehicle damage requests**.
- Further **fine-tune model hyperparameters** to **improve sensitivity** without compromising specificity.