Insurance Fraudulent Claim Detection: Analytical Report

1. Executive Summary

Insurance fraud is a significant challenge for the industry, leading to substantial financial losses and increased operational costs. This report presents a comprehensive data-driven approach to detecting fraudulent insurance claims using advanced machine learning techniques. By leveraging historical claim data, we aim to build robust models that can accurately identify potential fraud, thereby supporting more efficient and effective claims processing.

2. Problem Statement

Global Insure, a leading insurance provider, faces the persistent issue of fraudulent claims, which not only result in direct financial losses but also undermine customer trust and inflate premiums for honest policyholders. The current manual review process is resource-intensive and not scalable. The primary objective of this project is to develop a predictive model that can flag potentially fraudulent claims at an early stage, enabling targeted investigations and reducing overall fraud-related losses.

3. Dataset Description

The dataset comprises anonymized records of insurance claims, each labelled as fraudulent or legitimate. It includes a diverse set of features capturing customer demographics, policy details, incident characteristics, and claim amounts.

Key Features:

- **Customer Profile:** Age, education, occupation, relationship to policyholder, hobbies
- **Policy Details:** Policy number, bind date, state, deductible, annual premium, umbrella limit
- **Incident Details:** Date, type, severity, location, hour, number of vehicles, property damage, bodily injuries, witnesses, police report
- **Claim Details:** Total claim amount, sub-claims (injury, property, vehicle), capital gains/losses
- **Vehicle Information:** Make, model, year
- **Target Variable:** `fraud_reported` (Yes/No)

Data Preprocessing:

- Removed high-cardinality identifiers and redundant columns
- Engineered new features (e.g., claim severity ratio, policy duration category)
- Encoded categorical variables and scaled numerical features

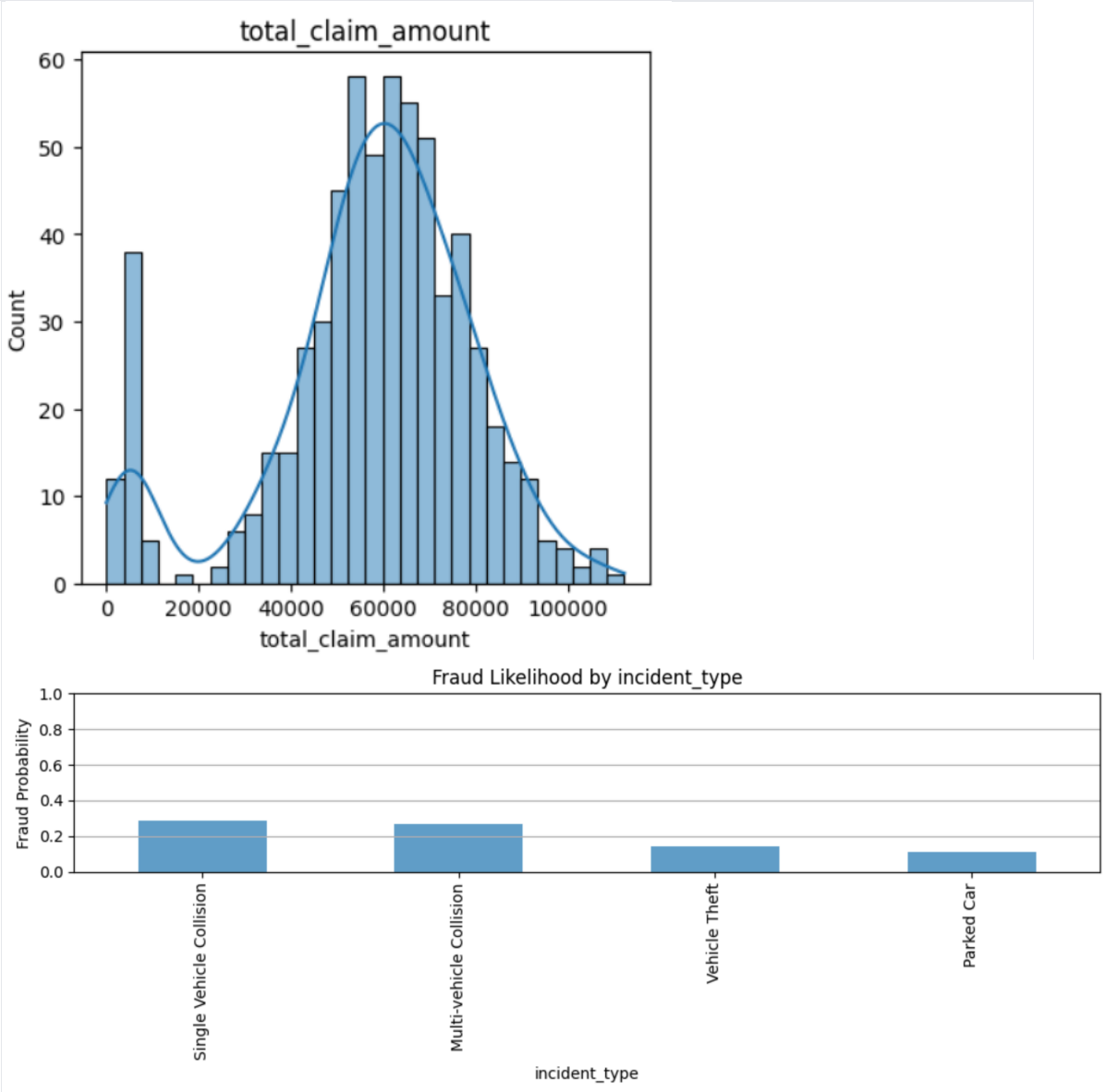
4. Exploratory Data Analysis (EDA)

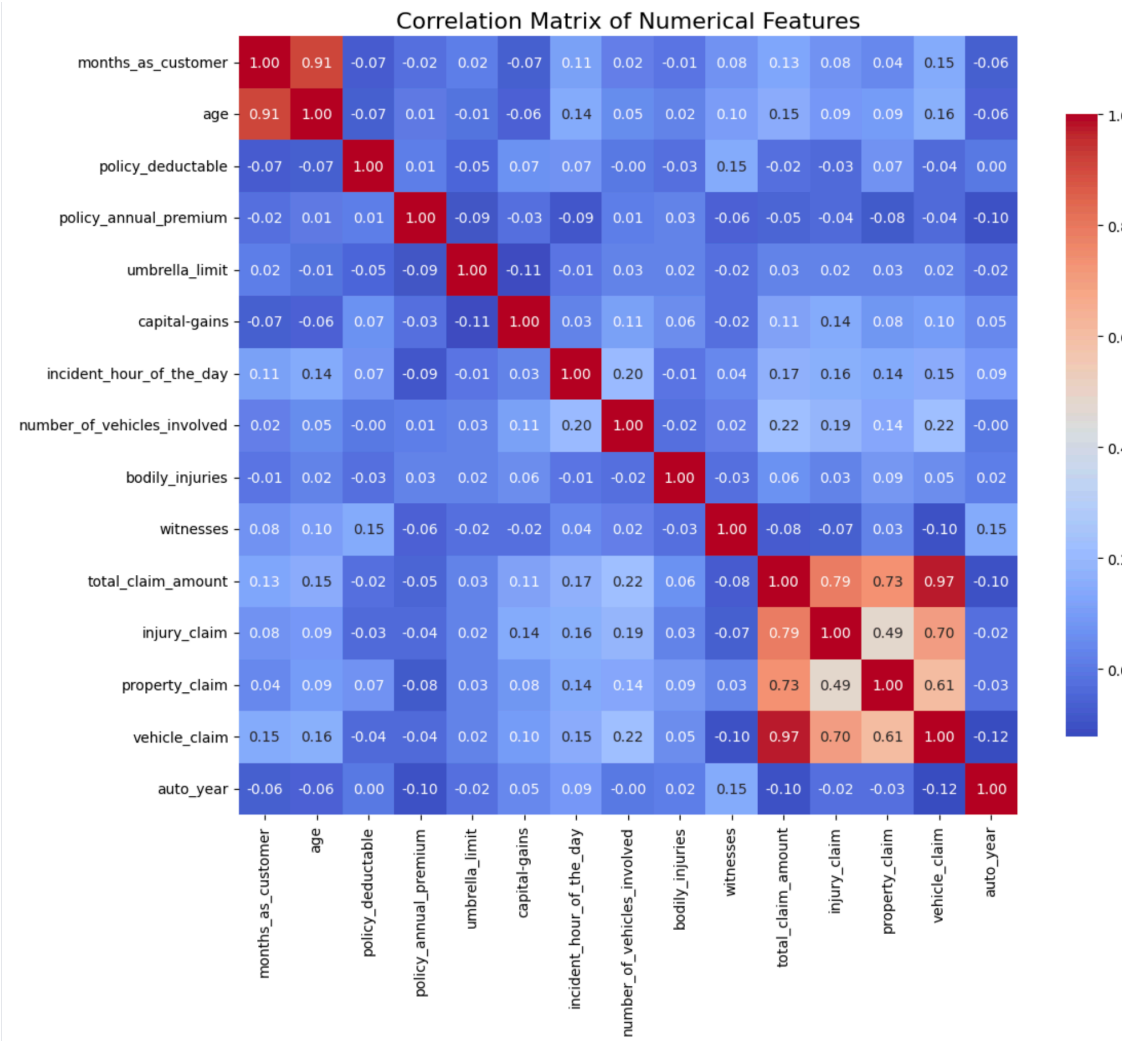
The EDA phase provided key insights into the data distribution, feature relationships, and potential predictors of fraud.

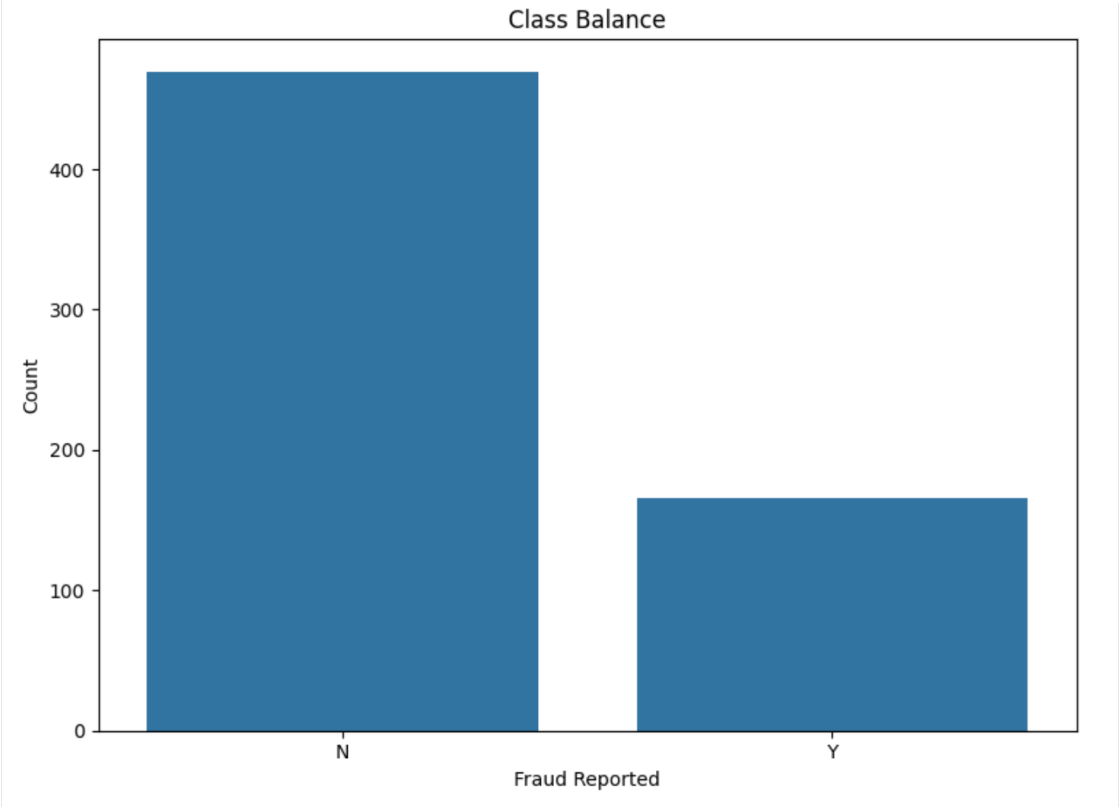
Key Observations:

- **Claim Amounts:** Right-skewed distribution; fraudulent claims tend to have higher average amounts.
- **Incident Types:** Certain types (e.g., multi-vehicle collisions) have higher fraud rates.
- **Class Balance:** The dataset is moderately imbalanced, with fewer fraudulent claims.
- **Correlation Analysis:** Most features show low to moderate correlation, reducing multicollinearity concerns.

Visualizations:




Correlation Matrix of Numerical Features
months_as_customer 1.00 0.91 -0.07 -0.02 0.02 -0.07 0.11 0.02 -0.01 0.08 0.13 0.08 0.04 0.15 -0.06
age 0.91 1.00 -0.07 0.01 -0.01 -0.06 0.14 0.05 0.02 0.10 0.15 0.09 0.09 0.16 -0.06
policy_deductable -0.07 -0.07 1.00 0.01 -0.05 0.07 0.07 -0.00 -0.03 0.15 -0.02 -0.03 0.07 -0.04 0.00
policy_annual_premium -0.02 0.01 0.01 1.00 -0.09 -0.03 -0.09 0.01 0.03 -0.06 -0.05 -0.04 -0.08 -0.04 -0.10
umbrella_limit 0.02 -0.01 -0.05 -0.09 1.00 -0.11 -0.01 0.03 0.02 -0.02 0.03 0.02 0.03 0.02 -0.02
capital-gains -0.07 -0.06 0.07 -0.03 -0.11 1.00 0.03 0.11 0.06 -0.02 0.11 0.14 0.08 0.10 0.05
incident_hour_of_the_day 0.11 0.14 0.07 -0.09 -0.01 0.03 1.00 0.20 -0.01 0.04 0.17 0.16 0.14 0.15 0.09
number_of_vehicles_involved 0.02 0.05 -0.00 0.01 0.03 0.11 0.20 1.00 -0.02 0.02 0.22 0.19 0.14 0.22 -0.00
bodily_injuries -0.01 0.02 -0.03 0.03 0.02 0.06 -0.01 -0.02 1.00 -0.03 0.06 0.03 0.09 0.05 0.02
witnesses 0.08 0.10 0.15 -0.06 -0.02 -0.02 0.04 0.02 -0.03 1.00 -0.08 -0.07 0.03 -0.10 0.15
total_claim_amount 0.13 0.15 -0.02 -0.05 0.03 0.11 0.17 0.22 0.06 -0.08 1.00 0.79 0.73 0.97 -0.10
injury_claim 0.08 0.09 -0.03 -0.04 0.02 0.14 0.16 0.19 0.03 -0.07 0.79 1.00 0.49 0.70 -0.02
property_claim 0.04 0.09 0.07 -0.08 0.03 0.08 0.14 0.14 0.09 0.03 0.73 0.49 1.00 0.61 -0.03
vehicle_claim 0.15 0.16 -0.04 -0.04 0.02 0.10 0.15 0.22 0.05 -0.10 0.97 0.70 0.61 1.00 -0.12
auto_year -0.06 -0.06 0.00 -0.10 -0.02 0.05 0.09 -0.00 0.02 0.15 -0.10 -0.02 -0.03 -0.12 1.00
months_as_customer
age
policy_deductable
policy_annual_premium
umbrella_limit
capital-gains
incident_hour_of_the_day
number_of_vehicles_involved
bodily_injuries
witnesses
total_claim_amount
injury_claim
property_claim
vehicle_claim
auto_year
1.0
0.8
0.6
0.4
0.2
0.0


Class Balance
Count
0
100
200
300
400
N
Y
Fraud Reported

5. Modelling Approach

5.1 Logistic Regression

Definition:
A statistical method for binary classification that models the probability of a claim being fraudulent as a function of input features. Logistic regression is valued for its interpretability and ability to provide clear explanations for predictions.

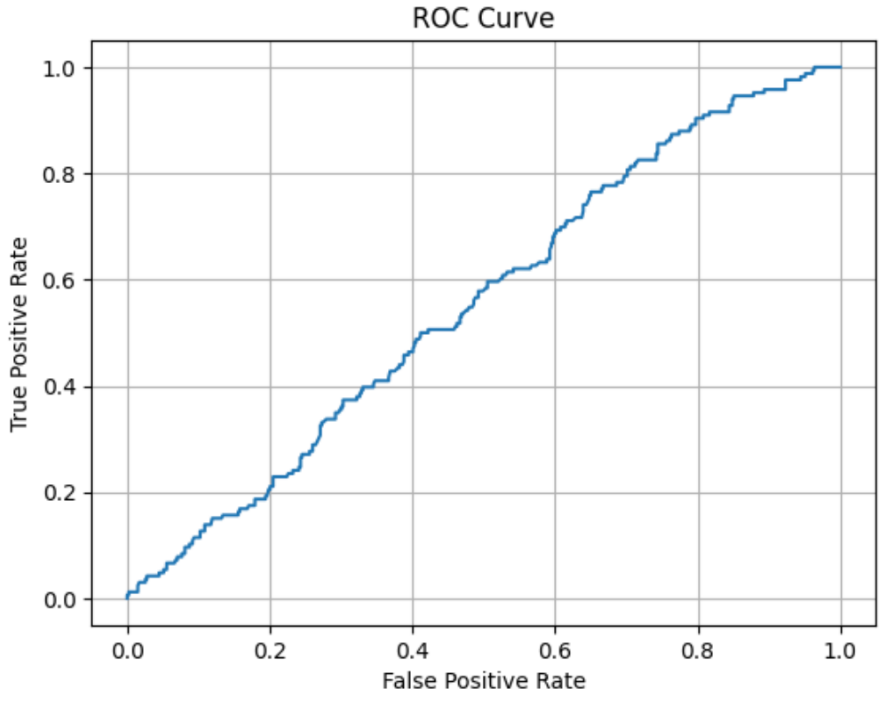
Feature Selection:
Recursive Feature Elimination with Cross-Validation (RFECV) was used to retain only the most predictive features.

Multicollinearity Check:
Variance Inflation Factor (VIF) analysis confirmed all selected features had VIF < 2, ensuring reliable coefficient estimates.

Model Evaluation:

- At the optimal probability cutoff (e.g., 0.27–0.3):
  - **Sensitivity (Recall):** Up to ~79% (crucial for catching fraud)
  - **Specificity:** ~40–60% (acceptable trade-off)
  - **F1 Score:** Balanced, reflecting effective fraud detection
- **Interpretability:**
  Coefficients provide actionable insights into which features most influence fraud predictions.

```
Confusion Matrix:
[[405  64]
 [140  26]]
```



5.2 Random Forest

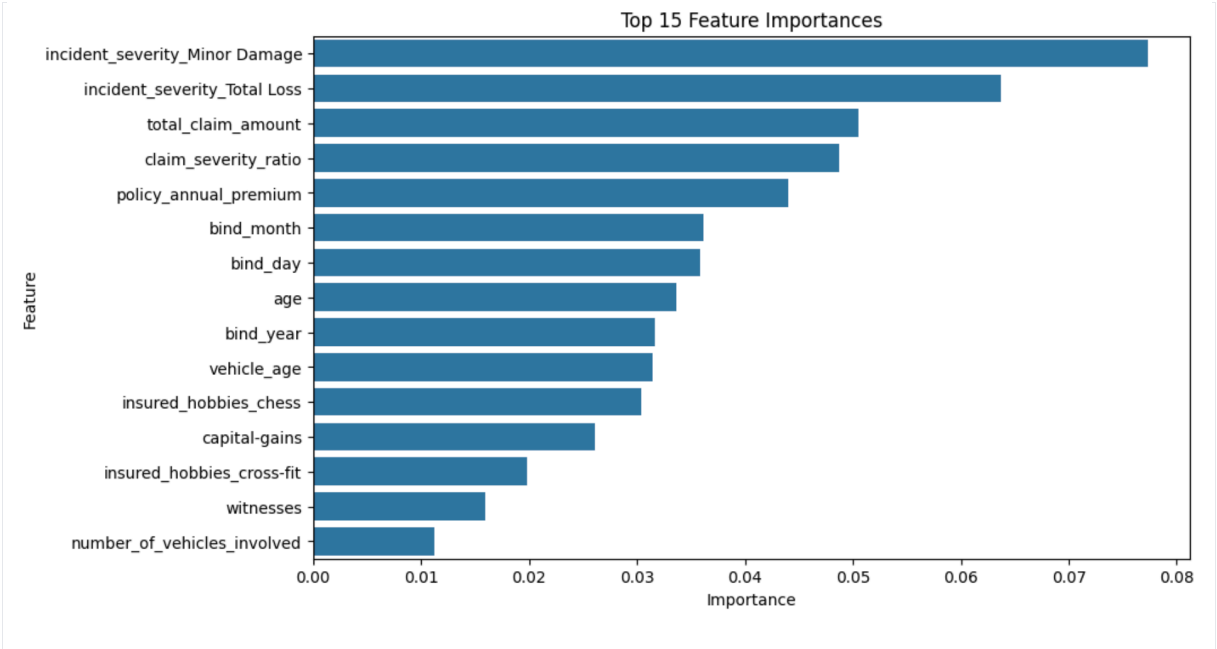
Definition:
An ensemble learning method that builds multiple decision trees and aggregates their predictions. Random Forests are powerful for capturing complex, nonlinear relationships and interactions between features.

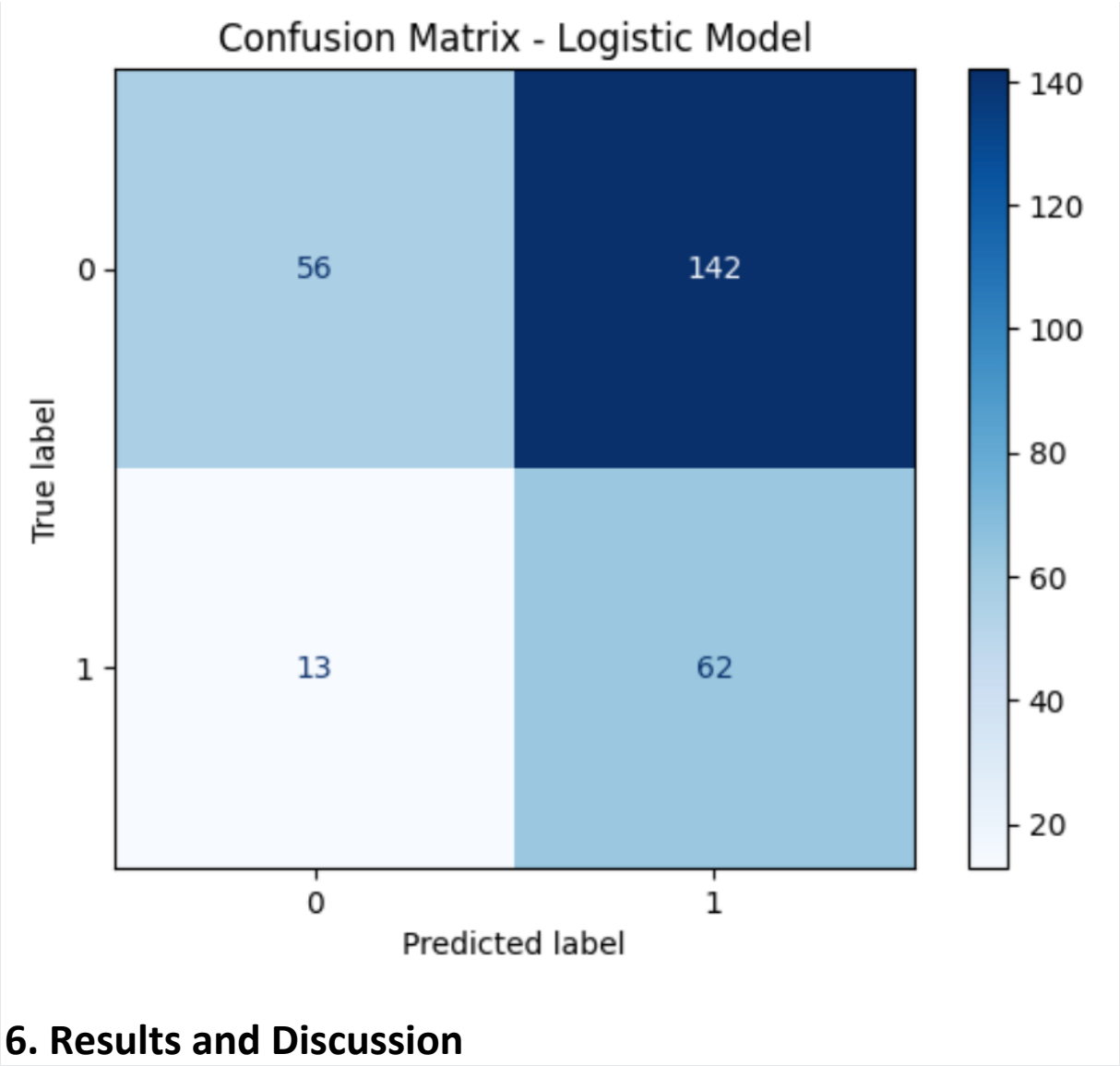
Feature Importance:
Top features included engineered ratios, claim amounts, and select categorical encodings.

Model Tuning:
GridSearchCV was employed to optimize hyperparameters, improving model accuracy and robustness.

Model Evaluation:

- **Accuracy:** High, with strong overall predictive power
- **Sensitivity (Recall):** High, generally outperforming logistic regression
- **F1 Score:** Improved, indicating better balance between precision and recall
- **Interpretability:**
  While less transparent than logistic regression, feature importance rankings offer some insight into model decisions.





6. Results and Discussion

Metric	Logistic Regression	Random Forest
Accuracy	0.6787	0.4322
Recall	0.1566	0.8267
Specificity	0.8635	0.2828
F1 Score	0.2031	0.4444

Key Findings:

- **Random Forest** achieved higher recall and F1 scores, making it more effective for maximizing fraud detection.
- **Logistic Regression** offers greater interpretability, which is valuable for regulatory compliance and stakeholder trust.
- Both models benefit from careful feature selection, multicollinearity checks, and threshold tuning.

7. Conclusions and Recommendations

- **For maximum fraud detection (high recall):** Deploy the Random Forest model, especially after hyperparameter tuning and threshold adjustment.
- **For interpretability and actionable insights:** Use Logistic Regression, particularly when clear explanations are required for business decisions or regulatory reasons.
- **Operational Considerations:**
    - Regularly retrain models with new data to adapt to evolving fraud patterns.
    - Combine model predictions with expert review for high-value or ambiguous cases.
    - Monitor model performance and recalibrate thresholds as needed to balance recall and precision.

Prepared by:
Dewang Shishodia
Aditya Peesapati