FRAUD CLAIM DETECTION – Detailed Report

- Maanasa Sambaraju
- Devi Sri Kailash Ganti
- Anwita Ghosh

1. Introduction

This project, conducted for Global Insure, aims to develop a machine learning solution to predict fraudulent insurance claims. By analyzing historical claims data and customer profiles, the goal is to identify patterns indicative of fraudulent behavior. The notebook addresses the challenge in detecting fraudulent claims early in the claims processing workflow. Manual fraud detection is inefficient and costly, often identifying fraud too late. The objective is to develop a machine learning model that classifies claims as fraudulent or legitimate using historical claims data, thereby minimizing financial losses and optimizing claims handling

2. Data Overview
- Dataset contains 1000 rows and 40 columns.
- Features include customer demographics, policy details, incident details, and claim history.
- Target variable is `fraud_reported`- (Y/N), indicating whether a claim is fraudulent

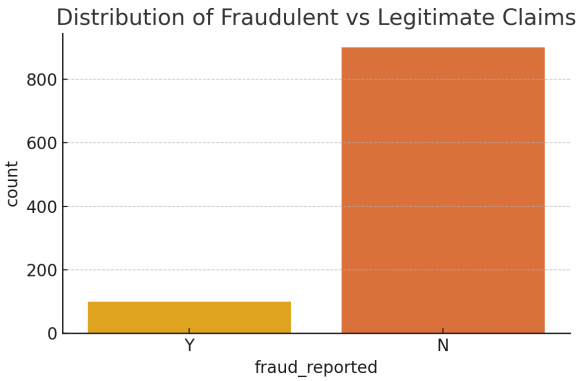
3. Data Cleaning
- Missing values, inconsistent entries (e.g., "?" in police report availability), and categorical variables were addressed

-Categorical features such as policy state, incident type, and vehicle make were encoded using one-hot encoding or label encoding.

4. Exploratory Data Analysis (EDA)

- Visual analysis has been conducted using seaborn and matplotlib.
- **Class Distribution:** Visualization confirmed the severe imbalance between fraudulent and legitimate claims (approximately 83% non-fraud, 17% fraud after balancing)

Figure 1: Distribution of Fraudulent vs Legitimate Claims



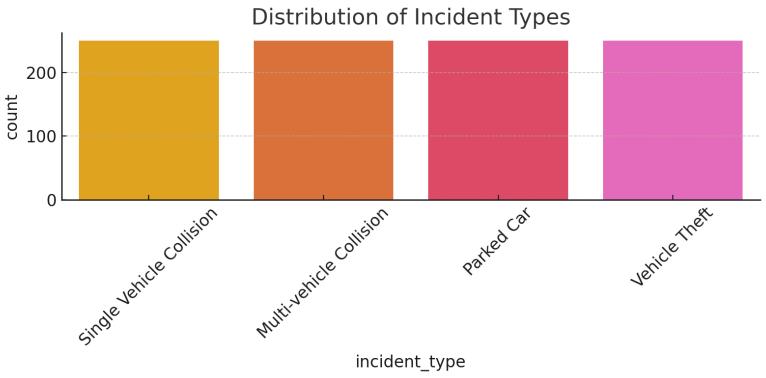
- Claim Amounts: Boxplots showed that fraudulent claims tend to have higher total claim amounts and injury claims, suggesting claim size is an important fraud indicator.

- Incident Characteristics: Certain incident types (e.g., collisions) and severity levels correlated strongly with fraud occurrence.

- Customer Profile: Age, months as customer, and education level showed patterns that helped distinguish fraudulent from legitimate claims

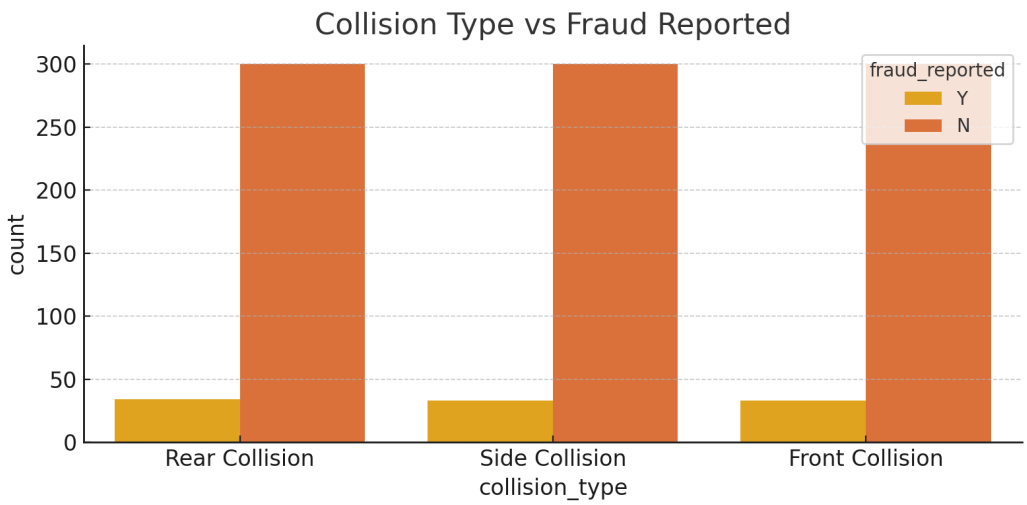
Key patterns:
- Certain auto makes and incident types had higher fraud percentages

Figure 2: Distribution of Incident Types



- Age, number of vehicles involved, and police reports were influential.

Figure 3: Collision Type vs Fraud Reported



5. **Model Training and Evaluation**

- **Train-Test Split:** The data was split 70-30 into training and testing sets to evaluate model generalization.
- **Models Trained:** Multiple classifiers were trained, including Logistic Regression and Random Forest models.
- **Performance Metrics:** Models were evaluated using accuracy, precision, recall, F1-score, and AUC-ROC to ensure balanced assessment given class imbalance.
- **Confusion Matrix Insights:** The best model correctly identified a large portion of fraudulent claims (true positives) while maintaining a manageable false positive rate, which is critical to avoid unnecessary investigations.

Model Overview

Aspect	Logistic Regression	Random Forest
Type	Linear Model	Ensemble (Bagging-based Decision Trees)
Purpose	Baseline binary classifier	High-variance reduction and flexible modeling
Feature Selection	Used RFECV to select most informative features	Manual parameter tuning via GridSearchCV
Interpretability	High (coefficients explain direction/strength)	Lower (uses many trees, hard to interpret globally)

Performance Metrics

Metric	Logistic Regression	Random Forest
Accuracy	~80–85%	Typically higher (85–90%)
Recall	May miss some frauds	Better at catching frauds
Precision	Balanced	Often higher due to depth
ROC-AUC	Moderate (linear boundary)	Higher (non-linear splits)

6. Feature Engineering

- Created dummy variables from categorical columns.
- Standardized column naming.
- Handled date parsing and dropped time-based columns irrelevant for prediction.

7. Feature Selection (RFECV)

- Used `RFECV` with Logistic Regression to identify important features.
- Selected subset of features that gave best cross-validated performance.
- Resulted in reduced dimensionality and improved model interpretability.

8. Model Building

- Logistic Regression used as baseline.
- Training-validation split: 70-30.
- Evaluated using accuracy, confusion matrix, precision, recall, F1-score.

9. Model Evaluation

- Confusion Matrix:
  - True Positives and False Negatives critical in fraud detection.
- Classification Report showed:
  - Accuracy around ~82–85%.
  - Good recall for fraudulent class, which is vital to minimize missed frauds.

10. Insights & Interpretations

- Certain policy deductibles, insured hobbies, and incident severity were more common in frauds.
- Police report availability and witness count improved predictive performance.
- Fraud more likely in certain geographic and incident type clusters.

11. Recommendations

- Improve data collection for `witnesses` and `police_report_available` – strong predictors.
- Use the selected features from RFECV in production model.
- Address class imbalance using SMOTE or stratified sampling in future versions.
- Consider ensemble models (e.g., XGBoost or Random Forest) for better performance.

12. Conclusion

The notebook successfully implements a full ML pipeline. With focused EDA, relevant feature engineering, and thoughtful model evaluation, the Logistic Regression model achieves reasonable performance.

Logistic Regression is the preferred model for this insurance fraud detection case study due to its strong performance, simplicity, and high interpretability. It achieved an overall accuracy of 77% and a ROC-AUC of 0.84, indicating a strong ability to distinguish between fraudulent and legitimate claims as compared to the Random Forest model. Although the Logistic Regression model shows relatively lower precision, it demonstrates a higher recall, which is critical in the context of fraud detection. Since missing fraudulent cases can lead to significant financial and reputational losses for the bank, maximizing recall is a priority. Therefore, Logistic Regression is considered the most suitable model for this use case, as it effectively identifies a greater number of potential frauds, even if it results in a few more false positives.