AUTO INSURANCE FRAUD DETECTION

OUR TEAM

1. ASHUTOSH DAS - 22CSEAIML071
2. MEETALI SINHA - 22CSEAIML034
3. SMARANEEKA - 22CSEAIML058
4. BHABANI SHANKAR PRADHAN - 22ECE101

INTRODUCTION AND PROBLEM STATEMENT

Problem: Auto insurance fraud costs billions, increasing premiums for honest policyholders.

Impact: Financial losses for insurers, inefficiency, eroded trust.

Our Goal: Develop a machine learning model to accurately identify fraudulent claims, improving efficiency and reducing losses.

OUR APPROACH (METHODOLOGY)

- **Phased Approach**: Data Understanding -> Preprocessing -> EDA -> Feature Engineering -> Model Building -> Evaluation -> Prediction.
- **Tools:** Python (Pandas, Scikit-learn, XGBoost, Matplotlib, Seaborn, Imblearn).

DATA SOURCES & INITIAL UNDERSTANDING

File01.csv: Training Data (with Fraud_Ind).

File02.csv: Testing Data (with Fraud_Ind for evaluation).

File03.csv: Prediction Data (unlabeled, model will predict Fraud_Ind here).

Initial State: Raw, mixed data types, some missing values, class imbalance.

DATA PREPROCESSING - PHASE 1: CLEANING & STRUCTURING

- Objective: To make data clean, consistent, and ready for analysis.

Actions:

- Loaded all three files to train_df, test_df, predict_df.
- Standardized column names (e.g., Bind_Date1 to Bind_Date).
- Converted Fraud_Ind (Y/N) to numerical (1/0).
- Filled specific missing values (authorities_contacted, Police_Report) with 'Not Contacted'/'No'.
- Removed any duplicate rows.
- Outcome: Clean, structured dataframes.

EXPLORATORY DATA ANALYSIS (EDA)

Accident Characteristics: 'Parked Car' accidents and 'Minor' severities show higher fraud rates. (Show Accident Type/Severity Plots)

Claim Amount Patterns: Fraudulent claims might have distinct claim value distributions. (Show Total Claim Plot)

Feature Relationships: Identified numerical features correlated with fraud. (Show Correlation Heatmap)

MODEL BUILDING & TRAINING

Objective: Train a high-performing fraud detection model.

Addressing Imbalance (Critical Step)

Technique Used: SMOTE (Synthetic Minority Over-sampling Technique)

Applied On: Training data only

Purpose: Balances the rare fraud cases by generating synthetic samples

Why Important: Prevents model bias toward the majority (non-fraud) class

Model Choice

Selected Model: XGBoost Classifier

Reason: Handles class imbalance well, high performance with structured/tabular data

Advantages: Robust to overfitting, supports early stopping, feature importance available

Solutions

CONCLUSION

This project successfully demonstrates the use of AI and predictive modeling to detect fraudulent auto insurance claims. By performing thorough data cleaning, handling missing values, and addressing class imbalance using SMOTE, we ensured a robust foundation for model training. The XGBoost Classifier was chosen for its effectiveness in handling tabular data and imbalanced classes, achieving promising results in identifying rare fraud cases.

Such a model can greatly support insurance companies in automating claims assessment, reducing financial losses, and improving decision-making, making claims management smarter and more efficient.

THANK YOU