Fraudulent Claim Detection Case Study

By :

- Akanksha Garg,
- Pooja Singh &
- Tushar Rajput

Business Objective

- How can we analyse historical claim data to detect patterns that indicate fraudulent claims?
- Which features are most predictive of fraudulent behaviour?
- Can we predict the likelihood of fraud for an incoming claim, based on past data?
- What insights can be drawn from the model that can help in improving the fraud detection process?

Business Questions Answered

- How can historical data help detect fraud?
  Pattern recognition using ML models identifies high-risk claims.
- Which features are most predictive of fraud?
  Features like insured hobbies, incident severity, incident type, police report availability, etc.
- Can we predict fraud likelihood for a new claim?
  Yes, with ~82% recall using Logistic Regression.
- What insights can improve fraud detection?
  Focus on high-risk categories, use model explanations, and integrate prediction scores into claim workflows.

Data Overview

01 **Dataset shape: (1000, 40)**

02 **Target variable: fraud_reported (Yes/No)**

03 **Features include demographics, policy details,vehicle info, incident specifics**

Data Cleaning & Preprocessing

- Handled Nulls: Imputed using domain knowledge (e.g., “unknown”, mode)
- Removed Redundancy: Dropped ID-like columns
- Fixed Types: Casted object types appropriately
- Result: Clean, analysis-ready dataset

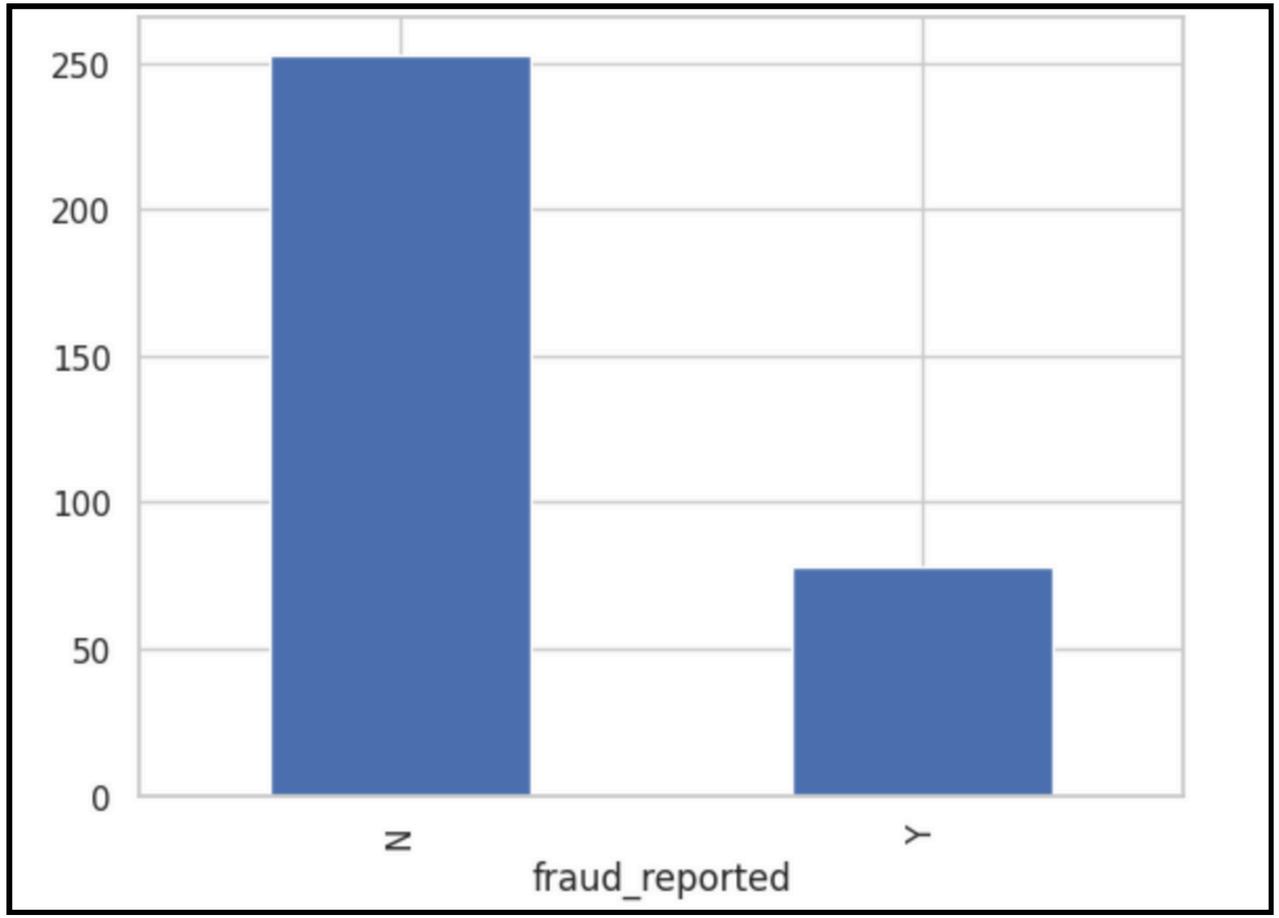
Train-Validation Split

- Performed stratified 70/30 split
- Maintains fraud proportion in both sets
- Class imbalance addressed during modeling

EDA —
Target Distribution & Class Imbalance

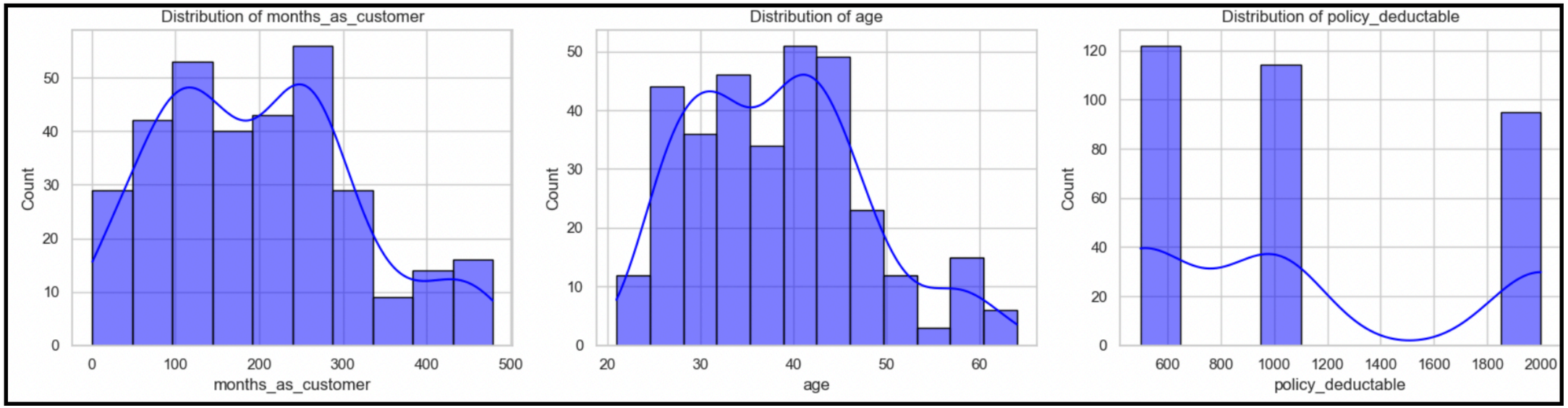
Fraud class is only ~23.56%

Strong imbalance → needs resampling

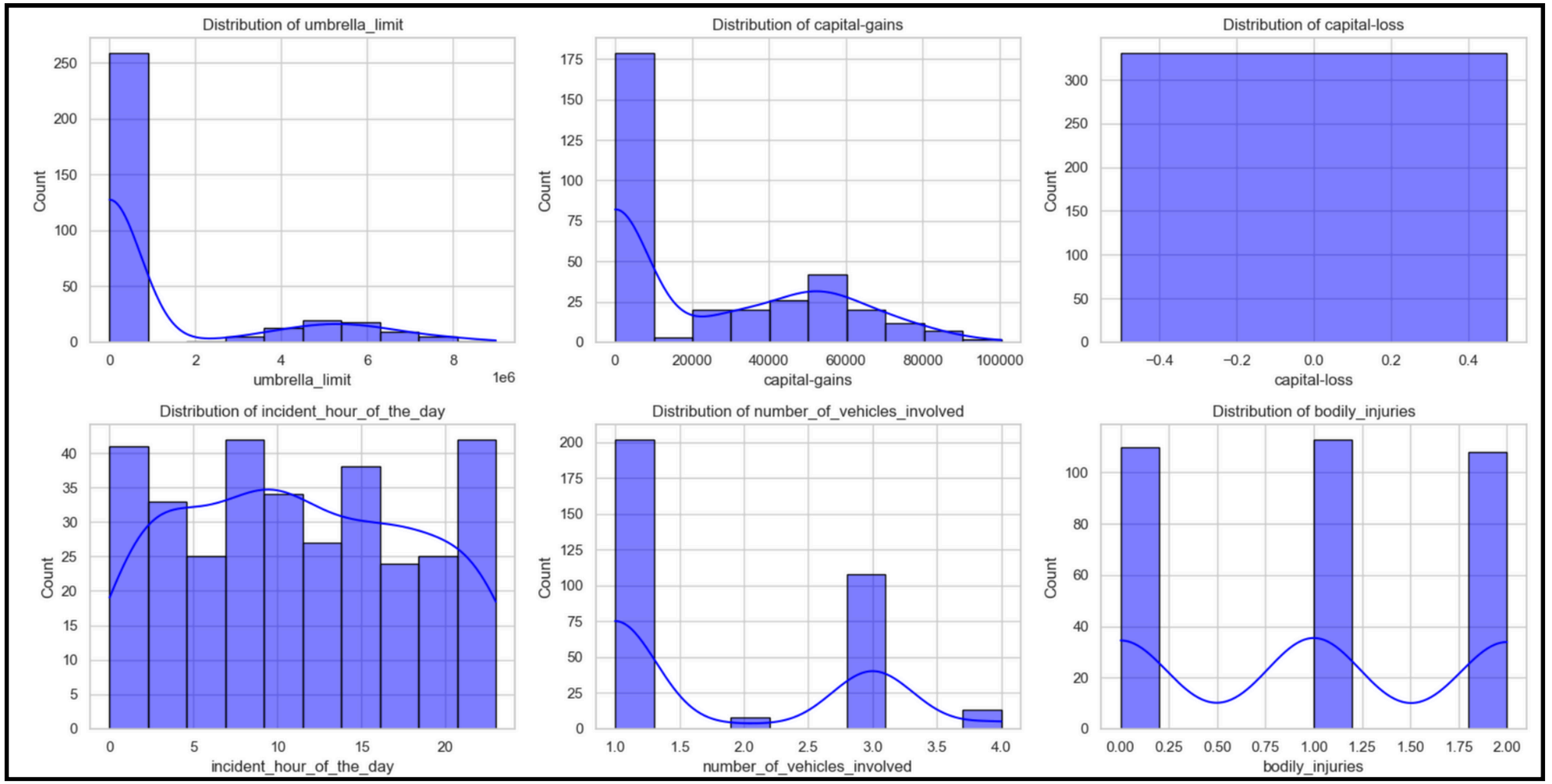


EDA
Univariate & Correlation Analysis

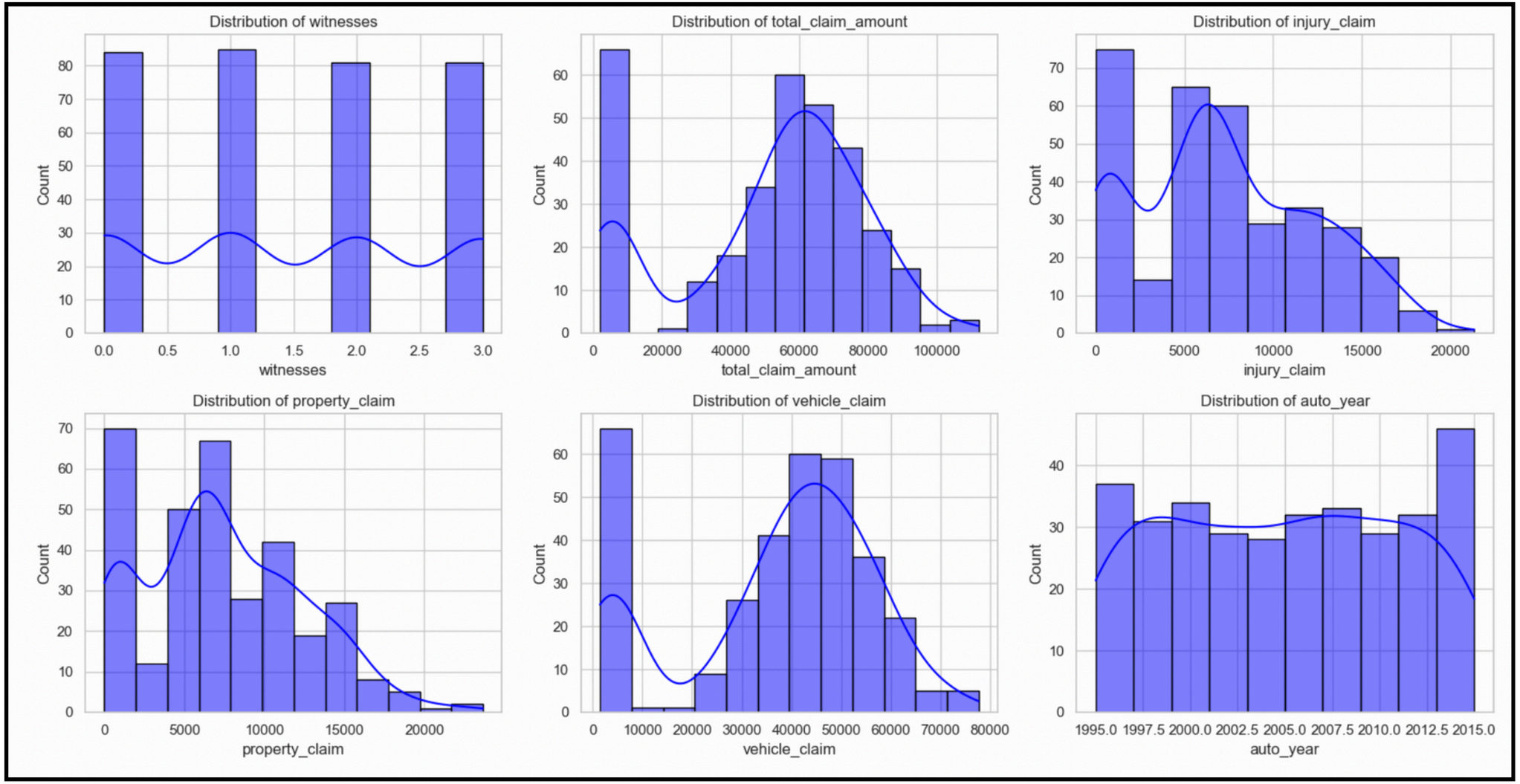
- Visualized categorical and numerical features
- Detected key patterns in:
  insured_hobbies (e.g., Chess, Cross-fit → >75% fraud rate), incident_severity, incident_type



EDA -
Univariate & Correlation Analysis

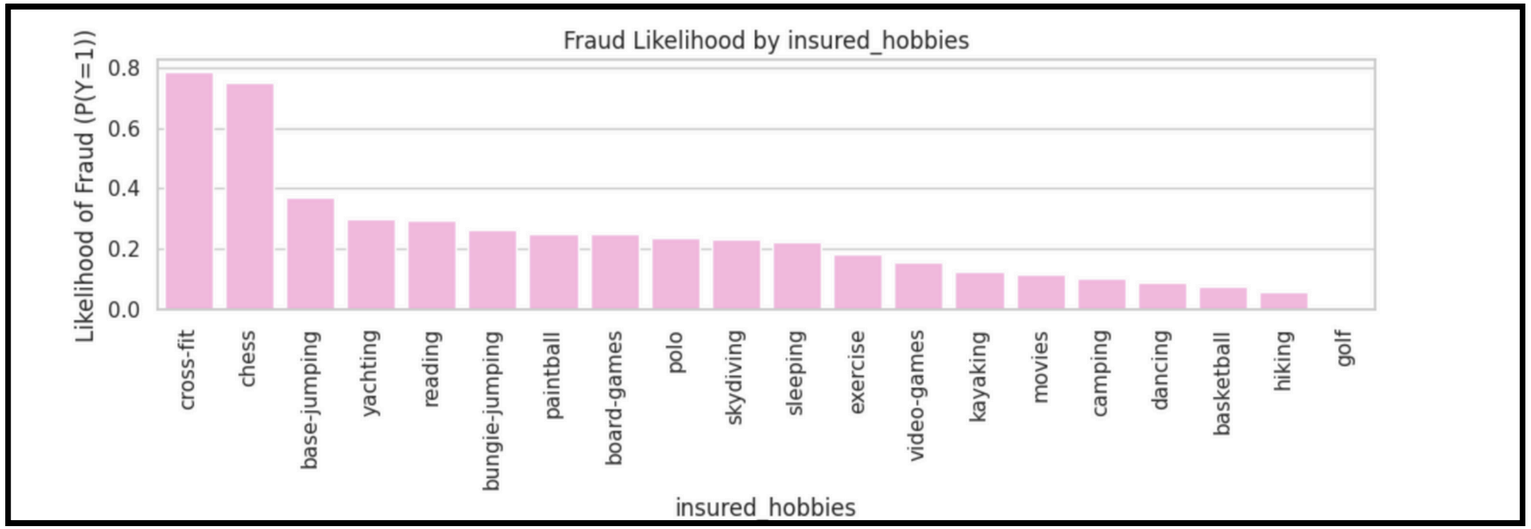


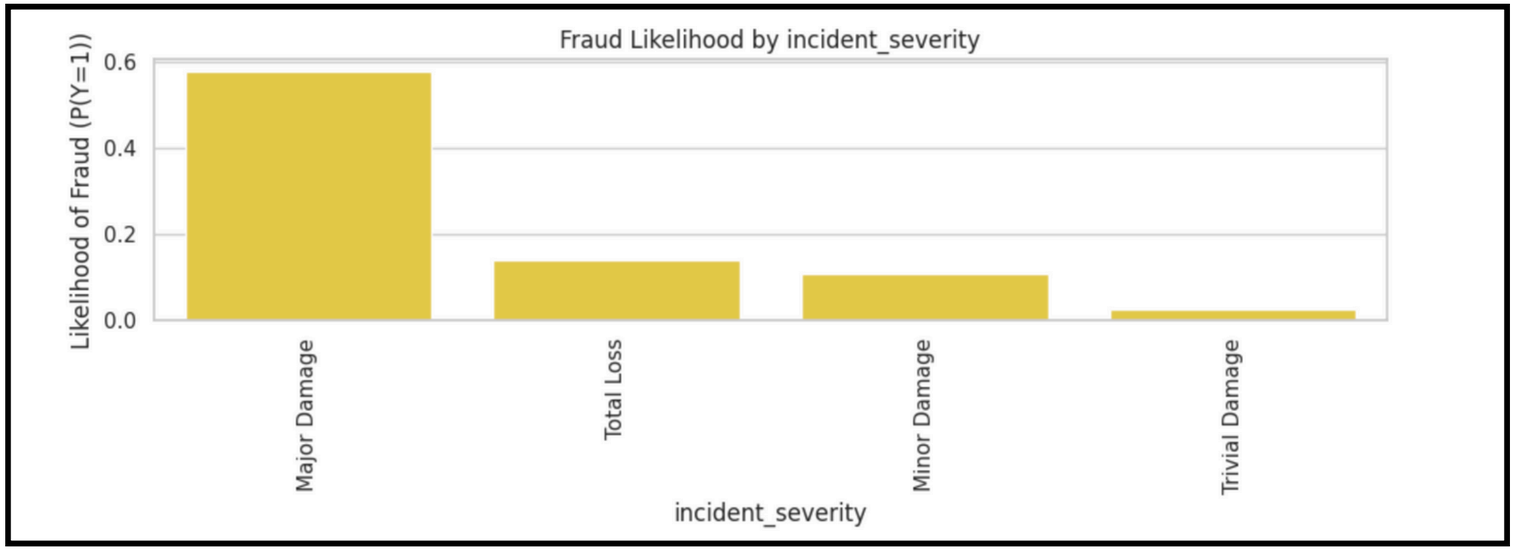
EDA -
Univariate & Correlation Analysis



Target Likelihood Analysis

- Performed likelihood mapping for high-cardinality categorical features
- Found strong indicators of fraud in: Incident_severity, auto_make, insured_hobbies, etc



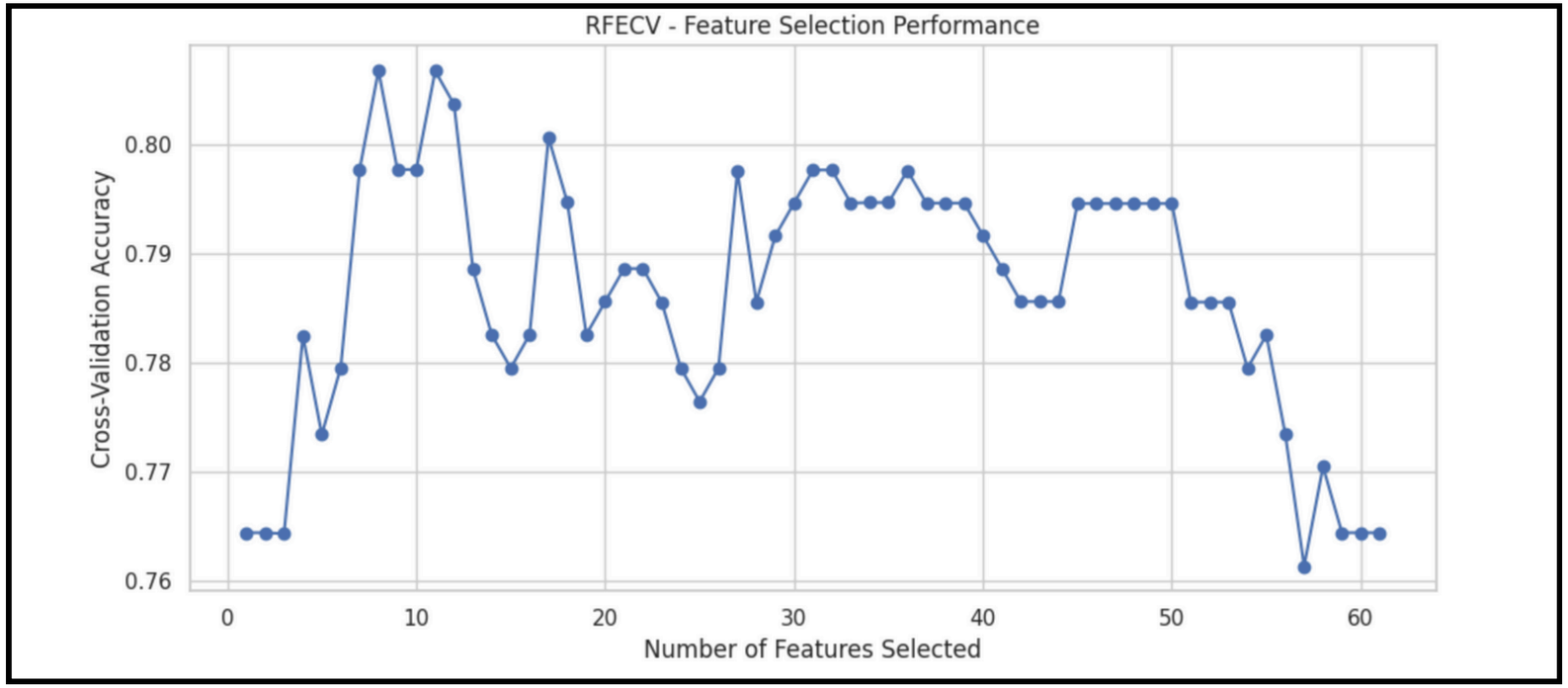


Feature Engineering

- Created: new features like claim_to_policy_ratio
- Grouped rare categories: e.g., consolidated hobbies/auto models
- Dummies created for categorical variables
- Standardized numerical features

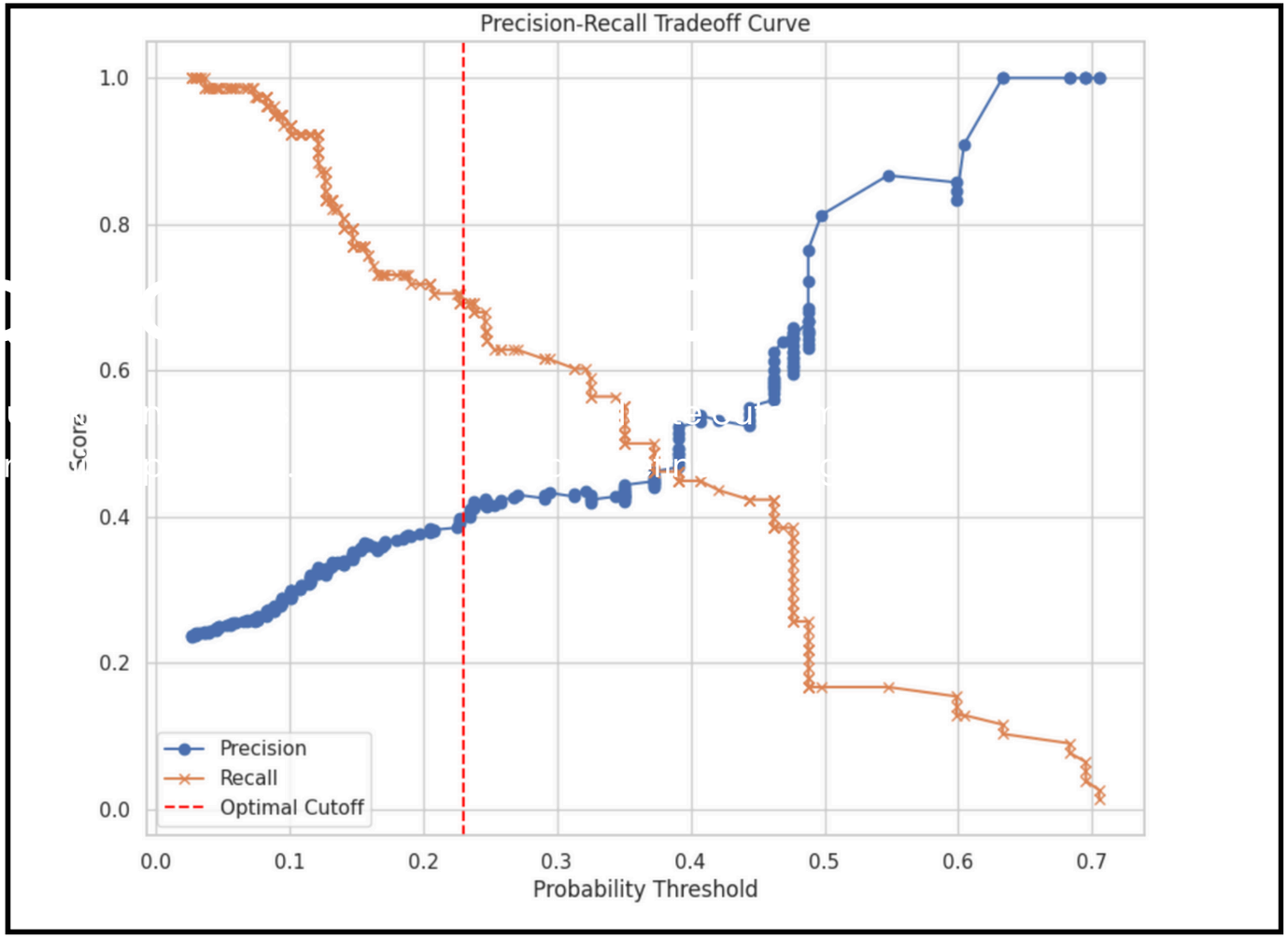
Feature Selection

- Used RFECV with Logistic Regression
- Selected top ~20 features
- Dropped low-importance ones to reduce noise



Model 1 – Logistic Regression

- Optimized using Grid Search for hyperparameters
- Best classification threshold based on F1 score from Precision-Recall Trade-off: 0.23



Logistic Regression:
Model Evaluation (Validation Data)

- Optimal Cutoff Used: 0.23
- Accuracy: 73.43%
- Precision: 46.67%
- Recall (Sensitivity): 82.35%
- Specificity: 70.64%
- F1 Score: 0.5957
- 📊 Confusion Matrix:

  TN: 77 FP: 32 FN: 6 TP: 28

Random Forest:
Model Evaluation (Validation Data)

- Accuracy: 76.92%
- Precision: 57.14%
- Recall (Sensitivity): 11.76%
- Specificity: 97.25%
- F1 Score: 0.1951

Despite slightly better accuracy and high specificity, Random Forest fails to detect most fraud cases due to extremely poor recall.

Conclusion and Recommendation:
Random Forest Model vs Logistic Regression

While Random Forest has marginally better overall accuracy, the Logistic Regression model provides much better recall (82.35%) and F1 score (0.5957) — crucial for fraud detection where identifying fraudulent claims is the top priority.

Therefore, Logistic Regression is the preferred model for deployment.

Thank you