



# Fraudulent Insurance Claim Detection

Amit Kumar Das  
Mukundan S



# Executive Summary

- Analysis of Insurance claims data was used to identify patterns for fraudulent claims.
- Develop a machine learning model to proactively classify claims as **fraudulent or legitimate** using historical claim data.
- Identify key indicators which skew the claim likelihood of it being fraud.
- Actionable insights and recommendations for business.



# Business Problem

- Global Insure faces significant financial losses due to fraudulent claims. This leads to increased financial leakage leading to annual losses.
- Real-World Impact
  - Unjustified payouts increase overall premiums and reduce profitability.
  - Manual investigations are time-consuming and prone to error.
  - Incorrect fraud handling can impact customer trust and regulatory compliance.
- Business Need
  - Identify high-risk claims early in the pipeline
  - Reduce manual workload on investigators
  - Enhance accuracy, efficiency, and customer satisfaction



# Dataset Overview

- Dataset Summary
  - Total Records : 1000 rows
  - Target variable: Fraud\_Reported
  - Class imbalance : 80% legitimate and 20% fraud
- Types of Data
  - Policy information like umbrella limits, policy dates, bind types
  - Incident details like incident date, severity, collision type
  - Customer Info like age, hobbies, gender
- Key Challenges
  - Target variable imbalance leads to biased predictions.
  - Missing values in various columns
  - Feature correction which required preprocessing

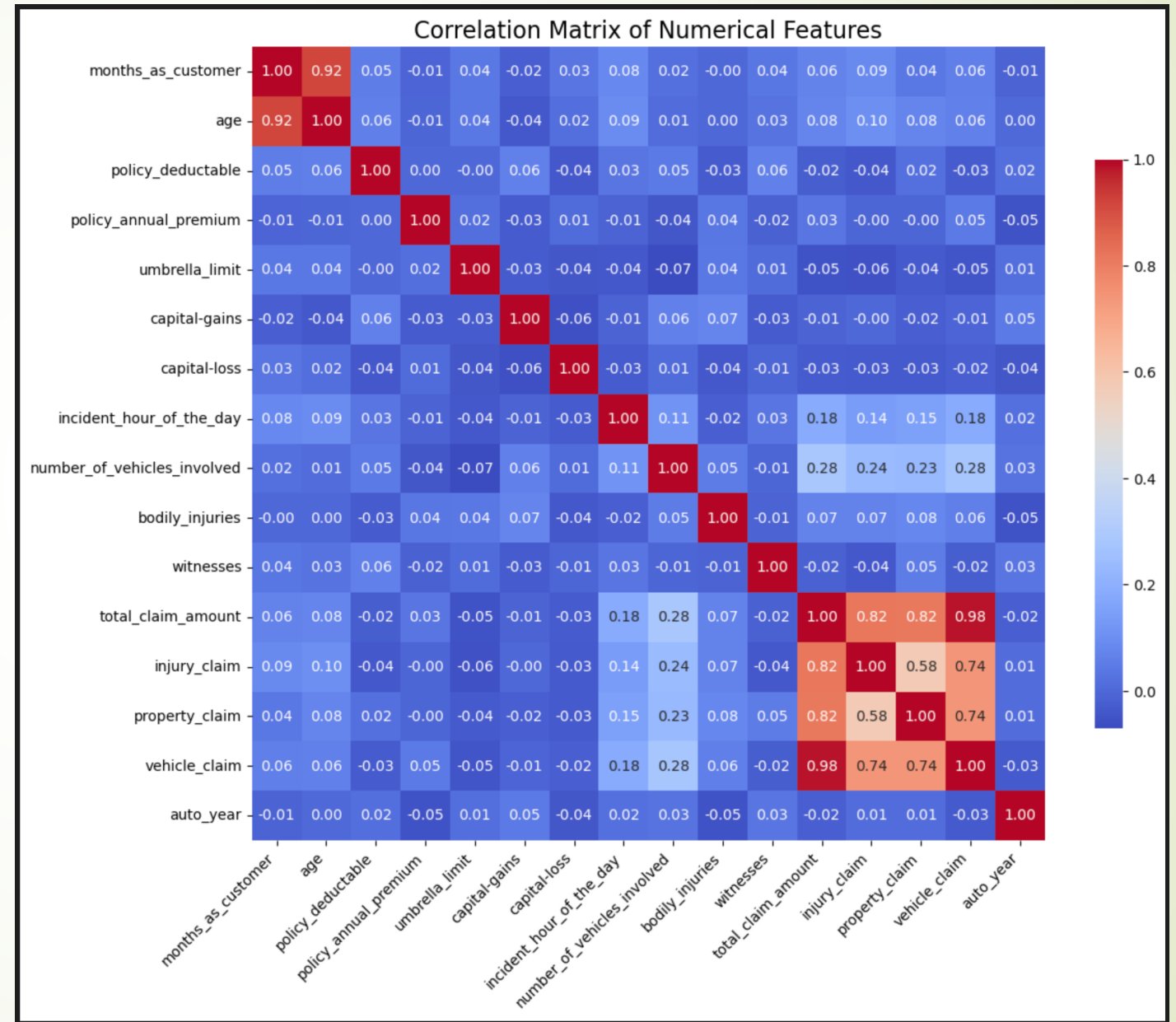


# Data Preprocessing

- Handling data issues
  - Columns with no values were not considered.
  - Data is imputed where data was missing like in 'authorities\_contacted'.
  - Illogical data is removed.
  - Fix data types
- Feature engineering
  - Identified columns which are mostly unique and have less predictive power like policy\_number, incident\_zip etc.
  - Converted categorical variables using one-hot encoding and ordinal encoding.
  - Created new variables from policy\_bind\_date, incident\_date, age\_group etc.
  - Standard scaling is performed to convert numerical variables to a standard scale.
- Class balancing
  - There is class imbalance and it's rectified by Random Sampling technique

# Exploratory Data Analysis

- Target likelihood shows certain car models, damage levels, collisions result in Fraud claims.
- Graphs used – histograms, boxplots, heatmap
- Anomalies and patterns detected which helped in finalizing the feature selection.





# Modelling Approach

- Algorithms
  - Logistic Regression – Efficient for Binary classification
  - Random Forest - Handles non-linearity
- Training Strategy
  - Training/Validation split
    - 70% Training and 30% Validation
    - Stratified approach to preserve target distribution
- Evaluation Metrics
  - Accuracy – Overall correctness
  - Sensitivity – Detect actual frauds
  - Specificity – Detect Legitimate claims
  - Precision – Accuracy of fraud detection
  - F1-score – Balance between Precision and Recall



# Model Performance

Model	Accuracy	Sensitivity	Specificity	Precision
Logistic Regression	83%	76%	85%	63%
Random Forests	80%	59%	87%	60%

- Logistic Regression outperformed Random Forest in overall balance across evaluation metrics, particularly in identifying fraudulent claims
- Confusion matrix, ROC curve, Precision-Recall curve were used to optimize performance
- Model accuracy is 83% which is a good indicator of model's capability.






# Model Selection

- Model Proposed
  - Logistic Regression
- Justification
  - Achieved better balance across key performance metrics
  - Provided higher **sensitivity (76%)** compared to Random Forest (55%), implying better fraud detection
  - Overall **Accuracy** of model(83%) is higher than Random Forest.
  - Logistic Regression provides a well-rounded, transparent, and business-aligned solution for fraud detection.



# Business Recommendations

- Integrate the Machine learning model in Claim registration workflow to detect Fraudulent claims
  - Improve data quality while claim registration to improve model performance.
  - Continuous model improvement to adapt with changing data and scenarios.
  - Focused Awareness campaigns for staff to improve data quality during registration.
- 



# Answers

■ Q - **How can we analyze historical claim data to detect patterns that indicate fraudulent claims?**

The historical data can be analyzed using EDA and machine learning models to identify patterns in dataset for predicting fraudulent claims. In EDA phase we saw that high claim amounts, long approval, hobbies like chess and cross-fit, incident severity can be used to identify fraud in the claim.

■ Q - **Which features are most predictive of fraudulent behaviour?**

**Incident Type** – Certain types of incidents (e.g., single-vehicle collisions) are more commonly associated with fraud.

**Collision Type** – Specific collision types (like rear-end or side-swipe) showed a stronger correlation with fraudulent behavior.

**Incident Severity** – High-severity incidents are often scrutinized more and were found to be key indicators.

**Police Report Availability** – Missing police reports were frequently observed in fraudulent claims.

**Insured Hobbies** – Certain hobbies had an unexpected association with fraudulent patterns.


**Auto Make** – Some car brands were statistically more common in fraudulent claims.

**Witnesses** – The absence of witnesses was a recurring trait in fraud cases.

**Authorities Contacted** – Fraudulent claims often skipped formal authority reporting.

**Number of Vehicles Involved** – Single-vehicle incidents had a higher fraud likelihood.

**Days Since Policy Inception** – Claims filed shortly after policy start were more suspicious.



# Answers – Contd.

- **Q - Can we predict the likelihood of fraud for an incoming claim, based on past data?**

We can definitely predict the likelihood of a fraud for an incoming claim using a well trained and optimized machine learning model. Using a machine learning model, like Logistic regression in this case, we can assign a probability score/value for the incoming claim. This will enable us in flagging the claim as Fraud by using the optimized threshold(0.6 in this case) early in the claim process.

- **What insights can be drawn from the model that can help in improving the fraud detection process?**

Using feature importance, variables such as policy\_annual\_premium, incident\_type, collision\_type have strong correlation with fraud. So more emphasis is required on these parameters during claim registration.

Claims without a police report had a higher probability of being fraud. So these can be flagged immediately.

Incident Severity and claim amount mismatch may hint towards a fraudulent case.

Data sanctity and sanity is required to improve model performance.



Thank You