Fraudulent Claim Detection

ADVANCED ML CASE STUDY

ANKIT DUBEY | UPGRADE | JAN 2026

Business Problem & Objective

- Manual fraud detection is slow and reactive.

•Global Insure processes **thousands of insurance claims annually**
•A significant portion are **fraudulent**
•Current fraud detection relies on **manual inspection**

Challenges
•Time-consuming investigations
•Fraud often detected **after claim settlement**
•Leads to:
- Financial loss
- Delayed genuine claims
- Inefficient resource allocation

Objective: Predict fraudulent vs legitimate claims early.

Other Objective

To answer the questions below through this presentation

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

Dataset Overview

Dataset Summary:-

Rows: 1,000 claims
Columns: 40 features
Target Variable: fraud_reported
Fraud Rate: ~24.7%

Feature Groups:-

Customer demographics
Policy details
Incident characteristics
Claim amount breakdown
Vehicle information

```
Numerical Features:
 Index(['months_as_customer', 'age', 'policy_number', 'policy_deductable',
       'policy_annual_premium', 'umbrella_limit', 'insured_zip',
       'capital-gains', 'capital-loss', 'incident_hour_of_the_day',
       'number_of_vehicles_involved', 'bodily_injuries', 'witnesses',
       'total_claim_amount', 'injury_claim', 'property_claim', 'vehicle_claim',
       'auto_year', '_c39'],
      dtype='object')

Categorical Features:
 Index(['policy_bind_date', 'policy_state', 'policy_csl', 'insured_sex',
       'insured_education_level', 'insured_occupation', 'insured_hobbies',
       'insured_relationship', 'incident_date', 'incident_type',
       'collision_type', 'incident_severity', 'authorities_contacted',
       'incident_state', 'incident_city', 'incident_location',
       'property_damage', 'police_report_available', 'auto_make', 'auto_model',
       'fraud_reported'],
      dtype='object')
```

Data Cleaning

Cleaning Steps

- Removed empty column _c39
- Handled missing categorical values using **mode**
- Removed high-cardinality identifiers:
  - Policy number
  - Zip code
  - Incident location
- Converted date columns to datetime format

Outcome

- Clean, consistent dataset
- No data leakage
- Ready for EDA and modeling

```
Out[166…  months_as_customer                int64
          age                               int64
          policy_bind_date         datetime64[ns]
          policy_state                     object
          policy_csl                       object
          policy_deductable                 int64
          policy_annual_premium           float64
          umbrella_limit                    int64
          insured_sex                      object
          insured_education_level          object
          insured_occupation               object
          insured_hobbies                  object
          insured_relationship             object
          capital-gains                     int64
          capital-loss                      int64
          incident_date            datetime64[ns]
          incident_type                    object
          collision_type                   object
          incident_severity                object
          authorities_contacted            object
          incident_state                   object
          incident_city                    object
          incident_hour_of_the_day          int64
          number_of_vehicles_involved       int64
          property_damage                  object
          bodily_injuries                   int64
          witnesses                         int64
          police_report_available          object
          total_claim_amount                int64
          injury_claim                      int64
          property_claim                    int64
          vehicle_claim                     int64
          auto_make                        object
          auto_model                       object
          auto_year                         int64
          fraud_reported                   object
          dtype: object
```

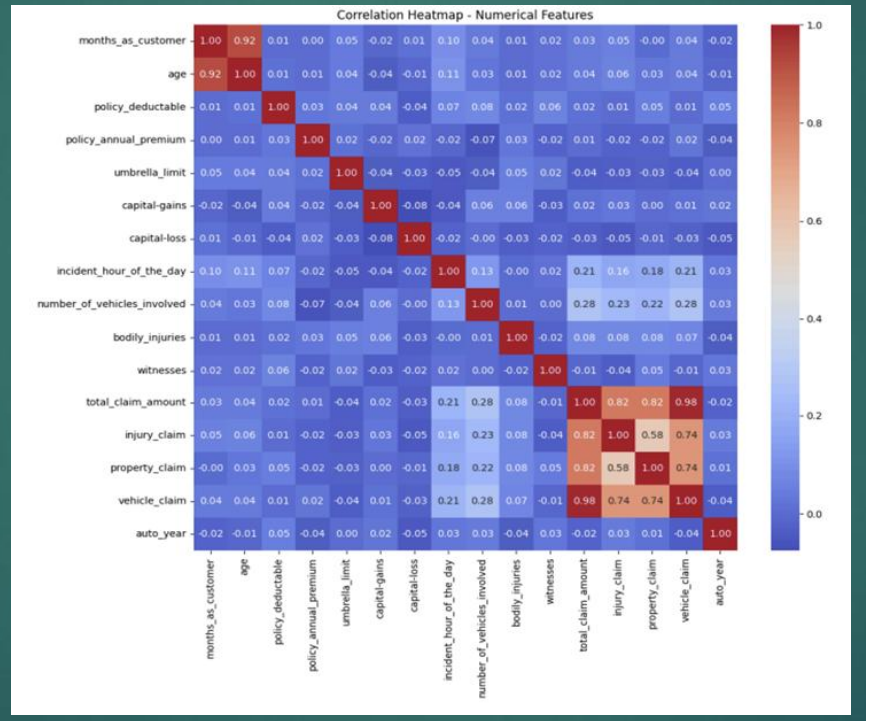
Univariate Analysis – Numerical Features

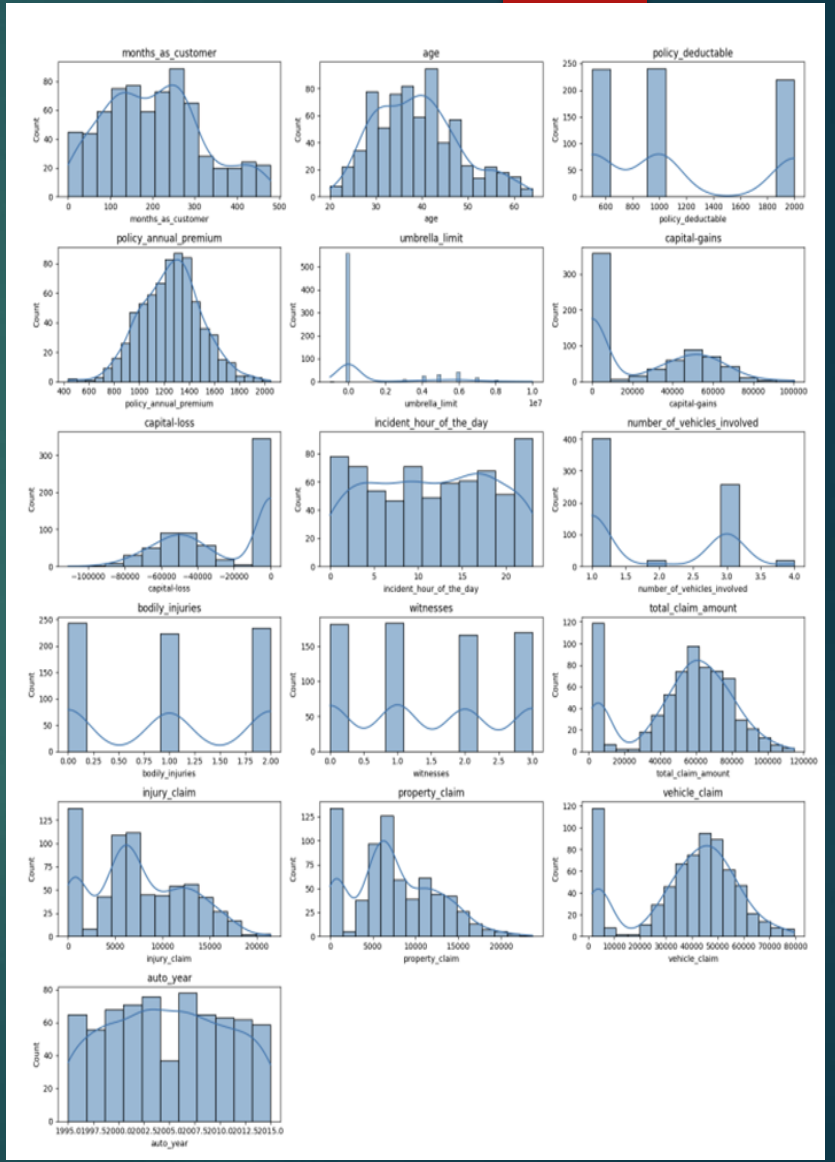
Why Univariate Analysis?

- Understand data distribution
- Identify skewness and outliers
- Prepare for transformations if required

Key Observations

- Claim amounts are right-skewed
- Customer tenure varies widely
- Claim components show different scales





Correlation Analysis

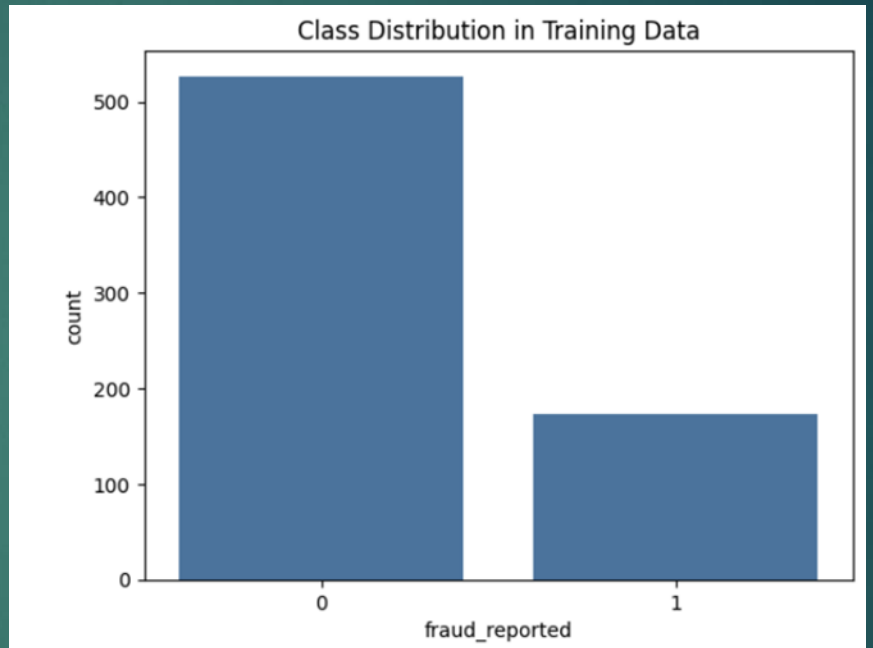
Objective

- Identify relationships between numerical variables
- Detect multicollinearity risks

Class Balance Analysis

Observation

Fraudulent claims ≈ 25%
Legitimate claims ≈ 75%




```
fraud_reported
0    0.752857
1    0.247143
Name: proportion, dtype: float64
```

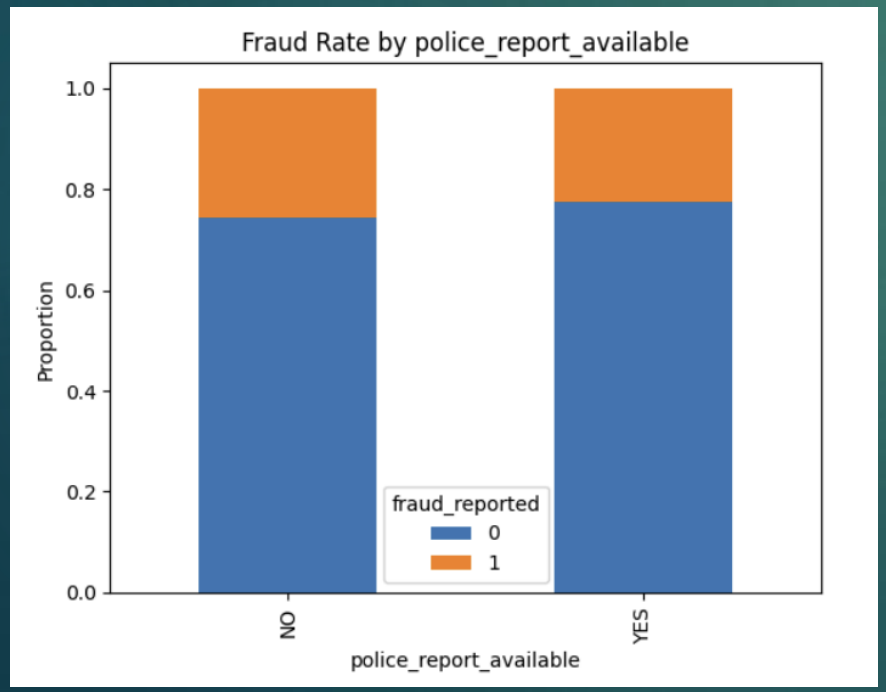
Bivariate Analysis – Categorical Variables

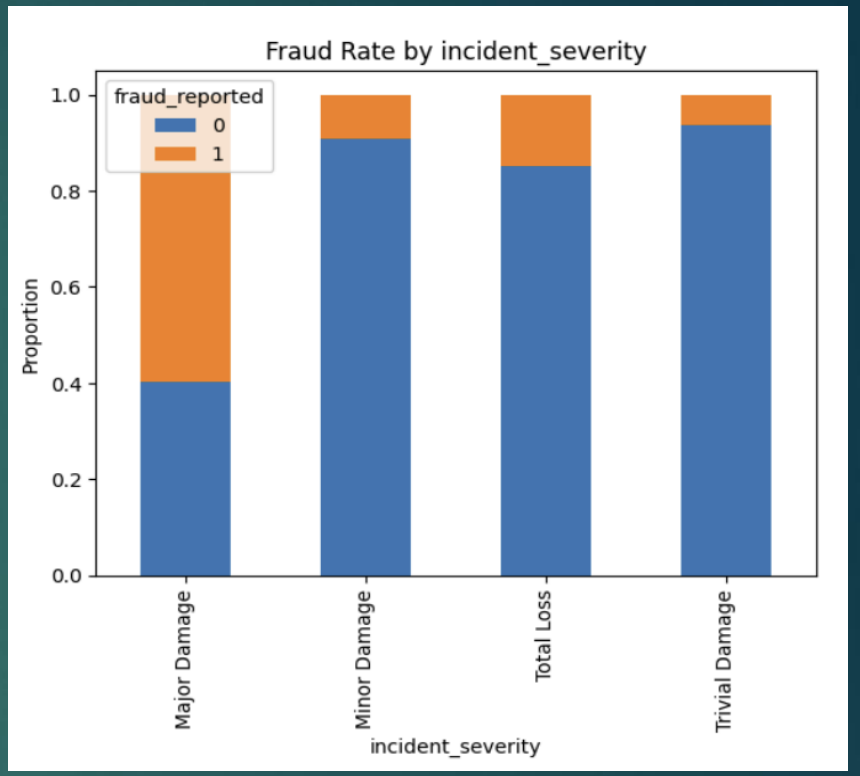
Method

Fraud rate comparison across categories

Key Insights

Major damage incidents show higher fraud rates
Claims without police reports are more suspicious
Certain incident types are more fraud-prone





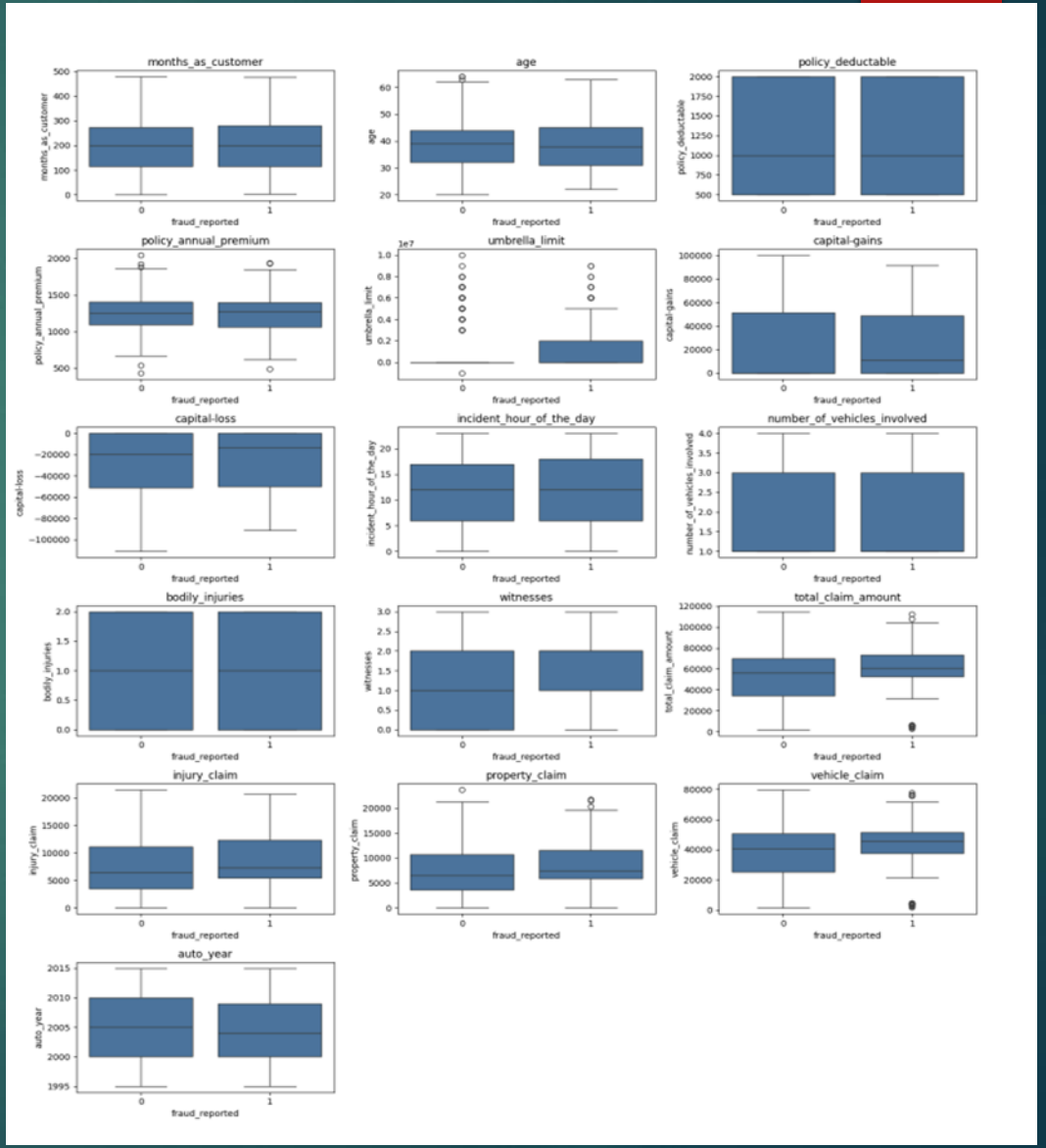
Bivariate Analysis – Numerical Variables

Method

Boxplots: Numerical features vs fraud outcome

Insights

Fraudulent claims show higher median values
Greater variability indicates inflated claims



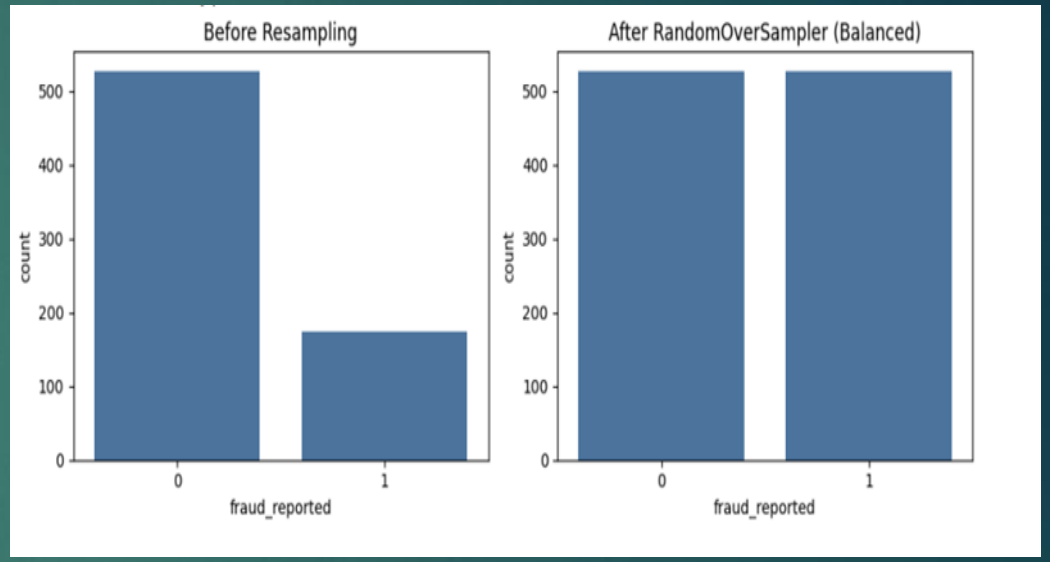
Feature Engineering

Why Feature Engineering?

Raw features may not capture fraud behavior
Derived features improve signal strength

Engineered Features

Policy duration (risk maturity)
Claim per month (claim aggressiveness)
Injury claim ratio (inflation indicator)
Total capital (financial stability proxy)
Age group segmentation



Handling Class Imbalance

Technique Used:- RandomOverSampler on training data only

Outcome

Balanced 1:1 fraud ratio
Improved recall for minority class

Feature Selection (Logistic Regression)

Method

•RFECV with 5-fold stratified cross-validation

Results

•Optimal subset: **26 features**
•Reduced noise and overfitting
•Improved interpretability

Model Performance Comparison

Key Results

- Random Forest:
  - Best accuracy and ROC-AUC (0.93)
- Logistic Regression:
  - Strong baseline
  - Clear interpretability

Model Building

Models Used

1.Logistic Regression

1. Interpretable
2. Statistical significance via p-values

2.Random Forest

1. Non-linear
2. High predictive power

- Algorithm: Logistic Regression (Statsmodels – MLE)
- Training samples: **1,054**
- Selected features (RFECV): **26**
- Pseudo R^2: **0.54**
- Model convergence: **Successful**

Statistically Significant Predictors (p < 0.05)

Feature	Coefficient Direction	Business Interpretation
policy_csl_250/500	Positive	Medium policy coverage shows higher fraud risk
insured_occupation_execmanagerial	Positive	Executive roles show increased fraud tendency
insured_occupation_transportmoving	Positive	Transport-related occupations linked to fraud
insured_occupation_handlerscleaners	Negative	Lower fraud likelihood for this occupation
insured_hobbies_chess	Positive	High-risk hobby cluster
insured_hobbies_cross-fit	Positive	Strong behavioral fraud indicator
insured_hobbies_dancing	Negative	Lower observed fraud risk
insured_hobbies_exercise	Negative	Fitness-oriented hobbies correlate with legitimacy
insured_hobbies_kayaking	Negative	Outdoor hobbies show lower fraud
insured_hobbies_movies	Negative	Entertainment hobbies less fraud-prone
insured_hobbies_sleeping	Negative	Lower behavioral risk
incident_severity_Minor Damage	Negative	Less severe incidents less likely to be fraud
incident_severity_Total Loss	Negative	Controlled severity reduces fraud probability
incident_severity_Trivial Damage	Negative	Lowest fraud likelihood
incident_state_PA	Negative	Lower fraud incidence in this state
incident_state_VA	Positive	Elevated fraud likelihood in this state
incident_state_WV	Negative	Lower fraud risk region
auto_make_Audi	Positive	Certain premium vehicle brands linked to fraud
auto_make_Honda	Positive	Higher fraud occurrence observed

Key Takeaways

- Claim context variables (severity, policy coverage) have the strongest influence.
- Behavioral attributes (occupation, hobbies) capture fraud intent effectively.
- Logistic Regression provides transparent and explainable fraud drivers, suitable for audit and regulatory use

Threshold Optimization

Why Optimize Cutoff?

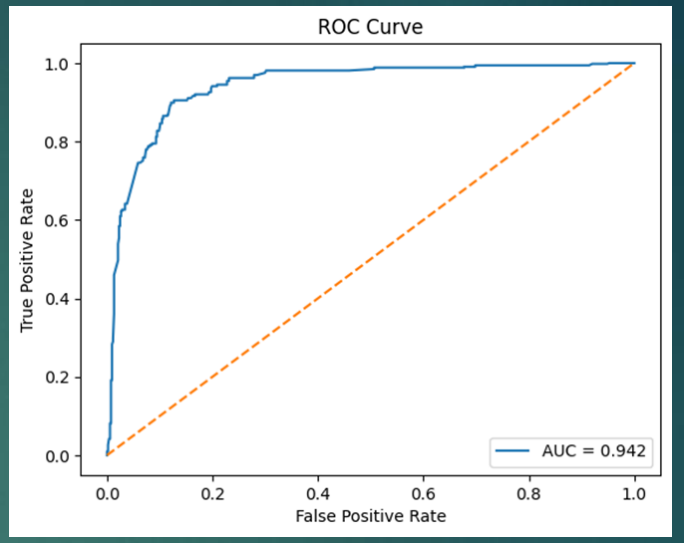
Business cost of false negatives is high

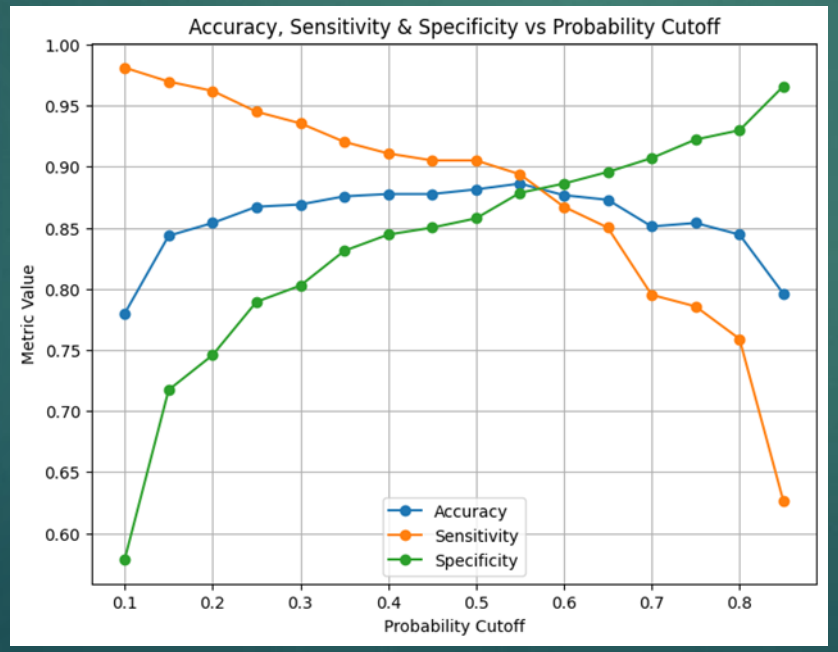
Default 0.5 cutoff is sub-optimal

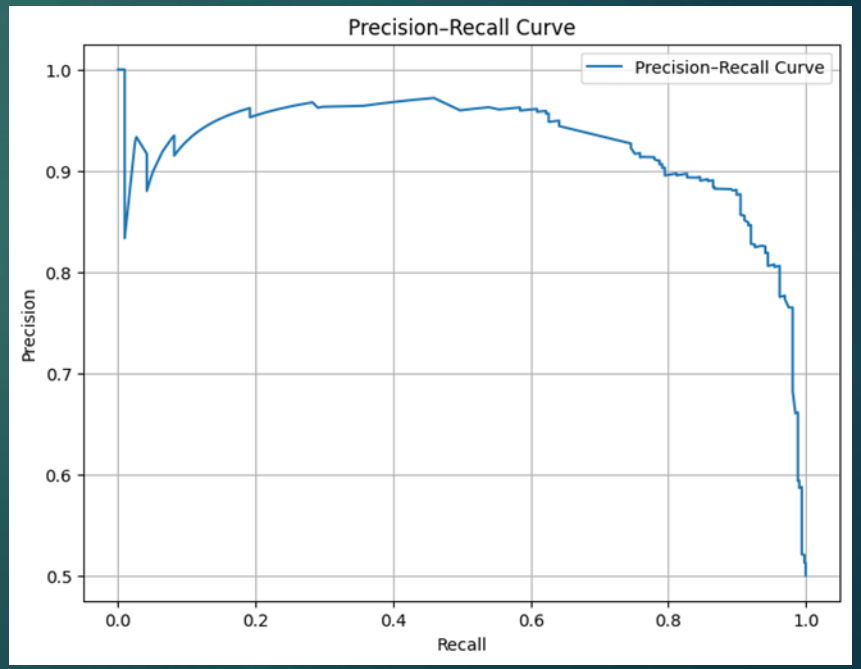
Insight

Optimal cutoff ≈ **0.6**

Better fraud capture without excessive false alarms







Answers to Business Questions & Business Recommendations

Key Answers

1.Fraud patterns:
    1. Inflated claim ratios
    2. Severe incidents
    3. Missing police reports
2.Predictive features:
    1. Claim ratios, severity, witnesses
3.Fraud prediction:
    1. Yes, high confidence (ROC-AUC 0.93)
4.Business insights:
    1. Risk-based triaging

Business Recommendations

Recommendations

- Deploy Random Forest as scoring engine
- Use Logistic Regression for audit explanations
- Introduce tiered claim processing:
    - Low risk → auto approval
    - High risk → manual review
- Retrain model quarterly

Conclusion

This project shows how a structured, data-driven machine learning approach can be effectively used to detect fraudulent insurance claims at an early stage. By carefully cleaning the data, exploring historical claim patterns, and creating meaningful features from policy, incident, and claim details, clear indicators of fraud were identified. The Logistic Regression model helped in understanding which factors are statistically important and why certain claims are more likely to be fraudulent, while the Random Forest model achieved better overall prediction performance and was able to estimate fraud risk with high accuracy. Together, these models provide a good balance between interpretability and performance. The insights obtained from this analysis can help Global Insure prioritise high-risk claims for investigation, reduce unnecessary manual effort, and speed up the approval of genuine claims. Overall, the proposed solution is practical, scalable, and suitable for real-world implementation in insurance fraud detection.