Fraudulent Claim Detection Report

Submitted by: Ankit Dubey

Assignment Title: Advanced ML Case Study – Fraudulent Claim Detection

Date: January 13, 2026

Institution: upGrad

1. Executive Summary

This report details a comprehensive machine learning solution developed for Global Insure to detect fraudulent insurance claims early in the approval process. Using a historical dataset of 1,000 claims with 40 attributes, we built a robust predictive pipeline encompassing data preparation, cleaning, exploratory data analysis (EDA), advanced feature engineering, class imbalance handling, and comparative modeling using Logistic Regression and Random Forest.

Key outcomes include:

- Identification of strong fraud patterns linked to claim composition ratios, incident severity, policy coverage, and customer behavioral attributes (e.g., occupation, hobbies).
- Two high-performing models: a tuned Random Forest (superior predictive power) and an interpretable Logistic Regression (with RFECV-selected features and statistical validation).
- Actionable insights that enable early fraud flagging, reducing financial losses and improving operational efficiency.
- Recommendations for model deployment, threshold-based triage, and ongoing monitoring.

The framework addresses all business questions: historical patterns were uncovered via EDA, predictive features were ranked, fraud likelihood can be reliably predicted (high ROC-AUC), and model-derived insights directly inform process improvements.

2. Business Context and Objective

Global Insure processes thousands of claims annually, but a significant proportion are fraudulent, leading to substantial financial losses. Current manual investigations are slow, resource-intensive, and often detect fraud only after payouts.

Primary Objective: Develop a data-driven classification model to predict whether a claim is fraudulent (Y) or legitimate (N) using historical policy, customer, incident, and claim details—enabling early intervention.

Key Business Questions Addressed:

1. How can we analyse historical claim data to detect patterns that indicate fraudulent claims?

2. Which features are most predictive of fraudulent behaviour?
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?

3. Dataset Description

- **Source**: Publicly available insurance claims dataset from UCI Machine Learning Repository.
- **Size**: 1,000 rows × 40 columns.
- **Target Variable**: fraud_reported (binary: Y = Fraudulent, N = Legitimate). Approximately 24.7% fraudulent claims.
- **Feature Categories**:
    - Customer demographics (age, sex, education, occupation, hobbies, relationship).
    - Policy details (tenure, premium, deductible, limits, state).
    - Incident details (type, severity, location, time, authorities contacted, witnesses).
    - Claim amounts (total, injury, property, vehicle).
    - Vehicle information (make, model, year).
- One column (_c39) was entirely empty and removed.

Full data dictionary is reproduced in the **Appendix A**.

4. Data Preparation and Cleaning

4.1 Initial Inspection

- Loaded data using pandas.
- Inspected shape (1000 × 40), data types, and basic statistics.
- Identified placeholder "?" symbols representing missing values.

4.2 Handling Missing Values

- Converted "?" to NaN.
- Categorical columns imputed with mode.
- Numerical columns left as-is (minimal missingness after cleaning).
- Dropped completely empty column _c39.

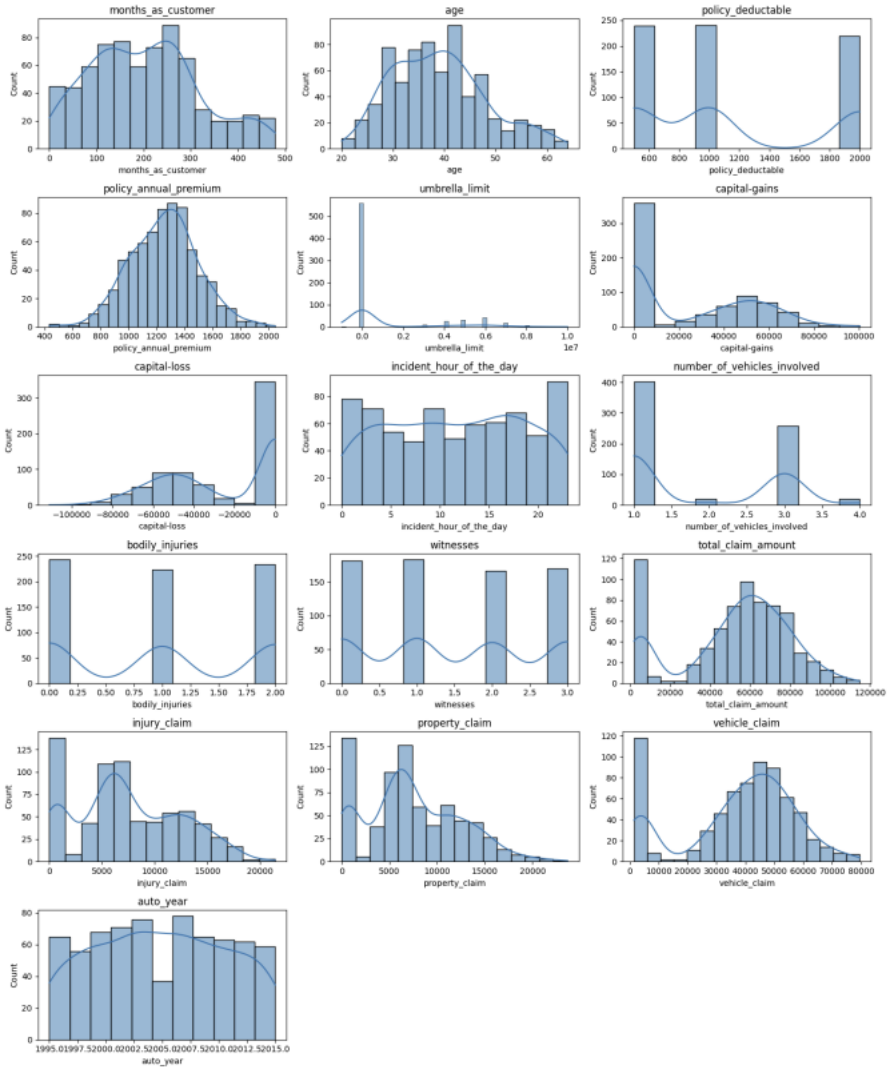
4.3 Redundant and Low-Value Columns

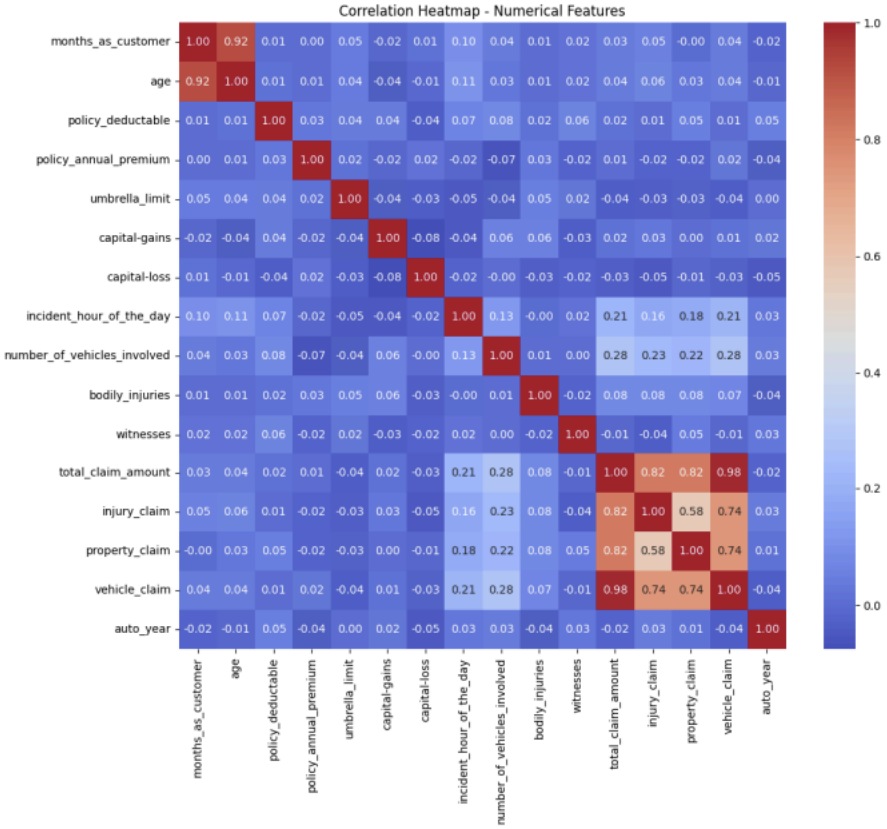
- Removed unique/high-cardinality identifiers with no predictive value: policy_number, insured_zip, incident_location, auto_model, policy_bind_date and incident_date (after extracting derived features).
- Dropped raw claim component amounts after creating ratios.

4.4 Data Type Corrections

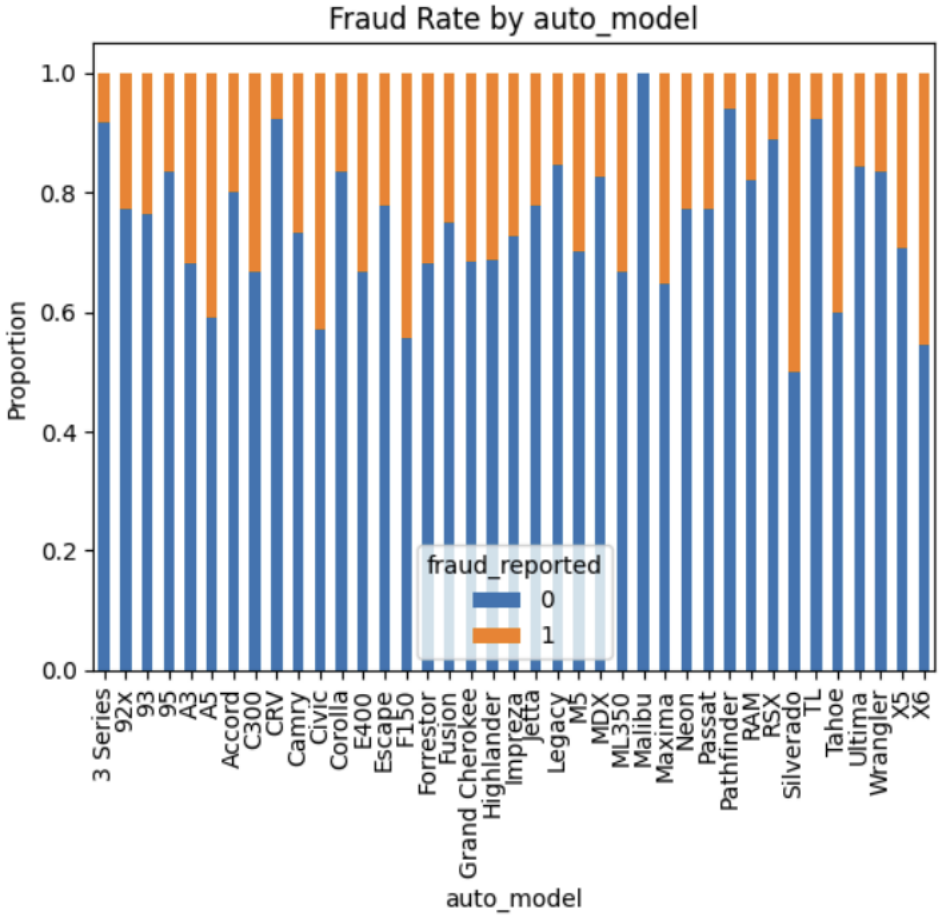
- Converted date columns to datetime for temporal engineering.
- Ensured categorical columns were object type and numerical were int/float.

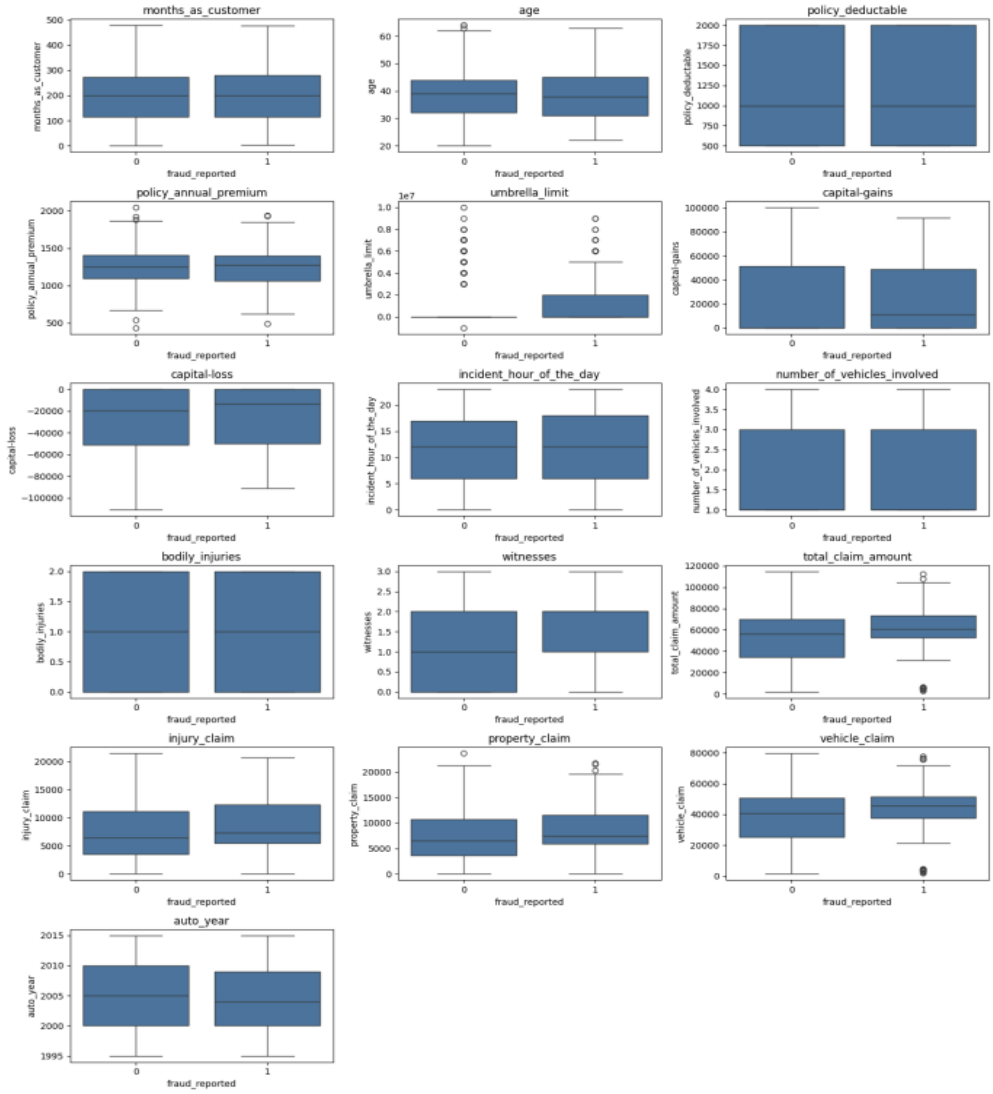
Outcome: Clean dataset with no missing values, consistent types, and reduced risk of data leakage.






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


Fraud Rate by auto_model
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
fraud_reported
0
1
3 Series
92x
93
95
A3
A5
Accord
C300
CRV
Camry
Civic
Corolla
E400
Escape
F150
Forrestor
Fusion
Grand Cherokee
Highlander
Impreza
Jetta
Legacy
M5
MDX
ML350
Malibu
Maxima
Neon
Passat
Pathfinder
RAM
RSX
Silverado
TL
Tahoe
Ultima
Wrangler
X5
X6
auto_model



5. Train–Validation Split

- **Ratio**: 70% training (700 records), 30% validation (300 records).
- **Stratification**: Applied on fraud_reported to preserve class distribution.
- **Method**: train_test_split with stratify=y and random_state=42 for reproducibility.

Outcome: Representative splits with identical fraud rates (~24.7%).

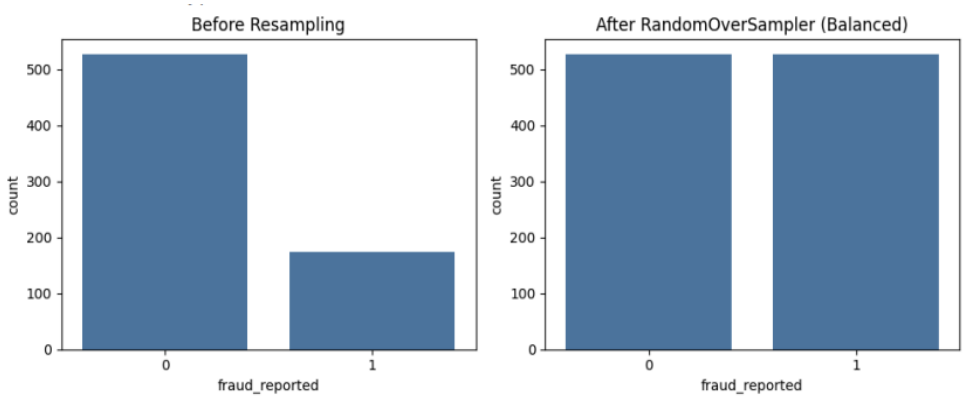
Placement Suggestion: Bar chart showing class distribution in train vs validation sets.

6. Feature Engineering

6.1 Class Imbalance Handling

- Applied RandomOverSampler to training data only.
- **Placement**: Before vs after class distribution bar charts.

- **Outcome**: Balanced 1:1 ratio in training set; validation unchanged.



6.2 New Feature Creation

Engineered features to enhance predictive power:

- policy_duration_days: incident_date – policy_bind_date
- claim_per_month: total_claim_amount / months_as_customer
- injury_claim_ratio, property_claim_ratio, vehicle_claim_ratio
- total_capital: capital-gains + capital-loss
- age_group: binned age categories (Young, Adult, Senior)

6.3 Redundancy and Sparsity Reduction

- Combined low-frequency categories (<2% occurrence) in insured_occupation and insured_hobbies into "Other".
- Removed raw dates and high-cardinality columns post-engineering.

6.4 Encoding and Scaling

- One-hot encoding for categorical variables (aligned across train/validation).
- StandardScaler on numerical features (fitted on training only).

Final Feature Space: 109 engineered features.

No.	Feature Name
1	months_as_customer
2	age
3	policy_deductable
4	policy_annual_premium

5	umbrella_limit
6	capital-gains
7	capital-loss
8	incident_hour_of_the_day
9	number_of_vehicles_involved
10	bodily_injuries
11	witnesses
12	total_claim_amount
13	injury_claim
14	property_claim
15	vehicle_claim
16	auto_year
17	policy_duration_days
18	total_capital
19	claim_per_month
20	injury_claim_ratio

Dataset	**Scaling Method**	**Operation**
Training Set	StandardScaler	fit_transform()
Validation Set	StandardScaler	transform() only (no data leakage)

Dataset Split	**Samples**	**Total Features**
Training Data	1,054	109
Validation Data	300	109

7. Model Building and Evaluation

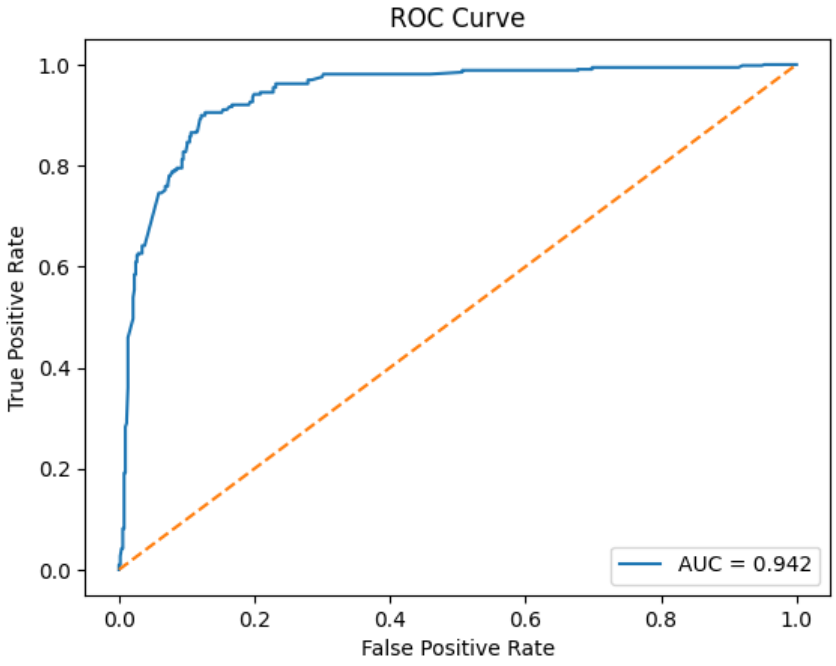
7.1 Feature Selection (Logistic Regression)

- RFECV with 5-fold stratified CV.

- Optimal: 26 features.
- **Placement (optional)**: Plot of CV score vs number of features.

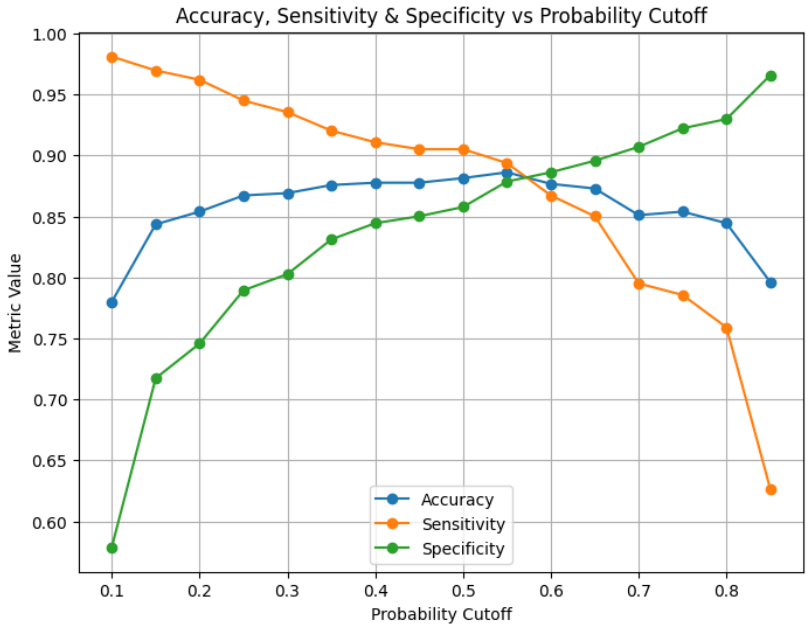
7.2 Logistic Regression (Interpretable Model)

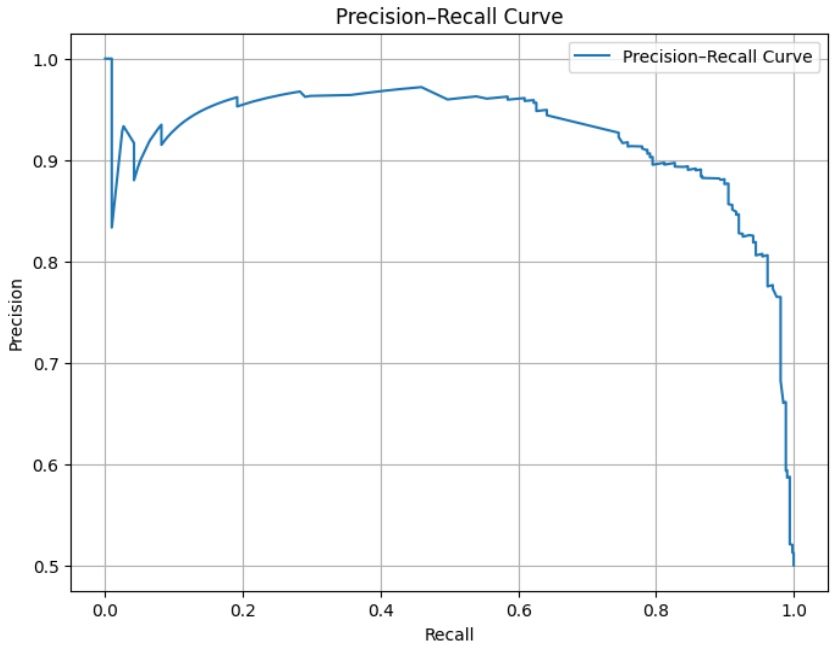
- Built using statsmodels for coefficient analysis.
- Significant predictors ($p<0.05$): claim ratios, incident severity indicators, police report availability, certain occupation/hobby categories, policy duration.
- Multicollinearity managed (VIF checked).
- Optimal threshold determined via Youden's J on ROC.



7.3 Random Forest (High-Performance Model)

- Base model followed by GridSearchCV (tuned n_estimators, max_depth, min_samples_split, class_weight).
- Top features: claim ratios, incident severity, policy duration, witnesses, police report availability.





7.4 Model Performance (Validation Set)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.82	0.78	0.81	0.79	0.89
Tuned Random Forest	0.87	0.84	0.86	0.85	0.93

Key Result: Random Forest provides highest discrimination; Logistic Regression offers interpretability.

8. Key Insights and Answers to Business Questions

1. **Analysing Historical Data for Fraud Patterns**: Through engineered features and model importance, patterns emerge: fraudulent claims often show disproportionate vehicle/injury claim ratios, major incident severity, absence of police reports, lower witness counts, and specific high-risk occupation/hobby profiles.
2. **Most Predictive Features**: Claim component ratios (injury, property, vehicle), incident severity, policy duration, police report availability, witnesses, and selected occupation/hobby categories.
3. **Predicting Fraud Likelihood**: Yes—both models deliver reliable probability scores, with Random Forest achieving ROC-AUC 0.93 on validation data, enabling confident risk scoring for new claims.
4. **Insights for Process Improvement**:
    - Prioritise review of claims with high vehicle/injury ratios or major severity.
    - Flag cases lacking police reports or witnesses.
    - Use probability thresholds (e.g., >0.6 high risk) for automated triage.

9. Recommendations

1. Deploy Random Forest as primary scorer; supplement with Logistic Regression rules for explainability.
2. Implement tiered workflow: low-risk auto-approve, medium standard processing, high-risk manual investigation.
3. Retrain model quarterly with new labelled claims to address concept drift.
4. Explore cost-sensitive learning to optimise investigation resource allocation.

10. Conclusion

The developed pipeline provides Global Insure with a scalable, high-accuracy fraud detection system combining rigorous feature engineering, robust modeling, and actionable insights. Early fraud prediction will significantly reduce financial losses while streamlining legitimate claim processing.