DATA-DRIVEN INSURANCE FRAUD DETECTION USING MACHINE LEARNING

GROUP 1

Srushti Jagwani – B004
Amisha Aggarwal – B014
Ishika Rupareliya – B024
Ansh Kumar – B034
Prajay Singh – B044
Ashwin Gondi – B054
Vaibhav – B064

DATE: 22 AUG 2025

EXECUTIVE SUMMARY

- **Problem:** Insurance fraud causes major financial losses; manual reviews miss fraud and delay legitimate claims.
- **Approach:** Built predictive ML models using 50,000 historical claims to detect high-risk cases.
- **Key Findings:**
  - Identified a “Fraud Triangle” ➔ claims with High Amounts, No Witnesses, and No Police Report are the strongest fraud indicators.
  - Final Random Forest model ➔ detects 82% of fraud with 100% precision (no false positives).

PROJECT OBJECTIVE

Develop an automated fraud detection system that:

01 **Accurately flags fraudulent claims**

02 **Optimizes investigator time**

03 **Fast-tracks payouts for genuine customers**

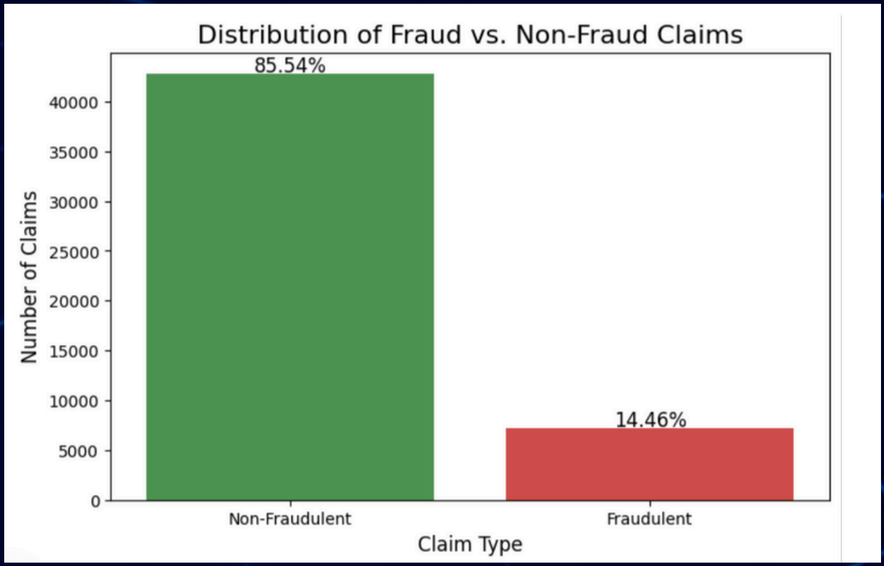
DATA UNDERSTANDING & VARIABLE EXPLORATION

Dataset Overview

- Insurance Claims Data.csv ➔ 50,000 anonymized claim records
- 29 attributes: identifiers, temporal fields, claim details, policyholder info, engineered flags
- Target variable: is_fraud (Fraud = 1, NoFraud = 0)

Key Variables & Data Types

- Numerical: claim_amount, policy_age_at_claim_months, time_to_report_days
- Categorical: loss_type, policyholder_occupation, location_zone, policy_type
- Flags (Binary): has_zero_witnesses, police_report_filed, is_high_claim_for_type, is_new_policy



Initial Observations from EDA

1.High Data Quality

- Dataset had no missing values across all 50,000 records ➔ clean, ready-to-use.

2. Severe Class Imbalance

- 85.5% legitimate claims vs. 14.5% fraudulent claims.
- This imbalance meant accuracy was misleading ➔ project focused on Precision, Recall, and ROC-AUC instead.

3.Fraud Correlated with Claim Circumstances, Not Demographics

- Strong indicators: High claim amount, no witnesses, no police report.
- Weak indicators: Occupation, location, and credit score showed almost no correlation with fraud ➔ broke common myths about profiling.

4.Distinct Numerical Distributions

- Fraudulent claims had higher average claim_amount (~$60,000 vs. ~$45,000 for legitimate).
- KDE plots showed clear distribution shift between fraud and non-fraud classes.

PROBLEM STATEMENT

- Fraudulent insurance claims cause financial losses and operational inefficiency.
- Current manual reviews are slow, costly, and unscalable.

Outcomes:

- Millions lost to undetected fraud.
- Legitimate claims delayed → customer dissatisfaction.
- Need scalable, accurate, data-driven fraud detection.

Key Question:

How can historical claims data be used to build an automated system that accurately distinguishes fraudulent vs. legitimate claims, focusing investigator effort on high-risk cases?

INITIAL HYPOTHESES

Claim Circumstances > Demographics

Witnesses, police reports, and claim amount are stronger fraud indicators than occupation, location, or credit score.

Time-Lapse Matters

Longer delays between incident and reporting (time_to_report_days) may indicate suspicious claims.

High-Value + New Policy = Higher Risk

Large claims filed soon after policy initiation are more likely fraudulent than similar claims from long-term policyholders.

DEPENDENT VARIABLES

Dependent Variable

- is_fraud ➜ Binary outcome
  - 1 = Fraudulent Claim
  - 0 = Legitimate Claim

- **Justification:**
  - Directly tackles financial leakage.
  - Gives a clear decision: Investigate vs. Fast-track.
  - Aligns with fraud management workflow.

WHY NOT OTHER VARIABLES?

- **dentifiers (IDs):** No generalizable predictive value ➜ would cause overfitting.
- **High-cardinality fields (agent_id, repair_shop_id):** Too many unique values ➜ model complexity with little added benefit.
- **Redundant Dates:** Already captured by time_to_report_days.
- **Text Narratives:** Required advanced NLP; tested, but no significant improvement ➜ excluded for efficiency.

INDEPENDENT VARIABLES

Independent Variables

- 16 features selected ➜ chosen for predictive power & business relevance.

Key Groups

1.Fraud Triangle:

- claim_amount, has_zero_witnesses, police_report_filed
- Strongest signals: financial motive & lack of verification.

2.**Policy & Claim Context:**

- policy_age_at_claim_months, time_to_report_days
- New policies + long reporting delays = higher fraud risk.

3.**Claim Descriptors:**

- loss_type, policy_type, location_zone, policyholder_occupation
- Add claim context & subtle fraud patterns.

4.**Pre-Engineered Flags:**

- is_high_claim_for_type, is_new_policy, has_vague_keywords
- Encode domain knowledge into simple fraud indicators.

MACHINE LEARNING ALGORITHM SELECTION

Evaluated 3 models → from simple baseline to advanced ensembles.

Goal: Benchmark performance, balance interpretability & accuracy, select the optimal solution.

LOGISTIC REGRESSION

- Simple, interpretable linear model.
- Used as a benchmark to test feature set predictive power.

BASELINE

XGBOOST

- Gradient boosting trees, captures complex non-linear patterns.
- Built-in handling of class imbalance.
- Widely used for structured tabular data like claims.

HIGH ACCURACY

RANDOM FOREST

- Bagging-based ensemble of decision trees.
- Resistant to overfitting, handles imbalance with class_weight.
- Provides robustness and strong out-of-the-box performance.

ROBUST CHAMPION

Why These Models?

Together, they cover the spectrum from interpretability → complexity.

Ensures the chosen champion model is credible, robust, and business-ready.

DATA PREPROCESSING

MISSING VALUES

01

None → dataset was complete across 50,000 records

TRAIN-TEST SPLIT

02

80/20 split with stratification to preserve 14.5% fraud ratio

CATEGORICAL ENCODING

03

One-Hot Encoding (e.g., loss_type, occupation) → avoids false order in nominal data

NUMERICAL SCALING

04

Standardization (claim_amount, policy_age, time_to_report_days) → mean = 0, std = 1 → ensures stable model training

FEATURE ENGINEERING

Created & Tested:

- Text-based: Sentiment, readability, and length from claim_narrative.
- Interaction: claim_amount ÷ policy_age → detects high-value claims on new policies.
- Date-based: Weekend incident flag, quarter of reporting.

Result:

No improvement over existing features → Random Forest already captured core fraud patterns.

Decision:

Final model trained on original 16 robust features → simpler, more efficient, equally accurate.

MODEL IMPLEMENTATION & EXECUTION

TOOLS & LIBRARIES

- **Pandas / NumPy →** Data handling & feature engineering
- **Scikit-learn →** Preprocessing (Scaler, OneHotEncoder), models (Logistic, RF), evaluation (classification_report, confusion_matrix)
- **XGBoost →** Gradient boosting implementation
- **Matplotlib / Seaborn →** EDA & model performance visuals

TRAINING & TESTING PROTOCOL

- **Stratified Train-Test Split (80/20):** Preserved 14.5% fraud ratio
- Train: 40,000 claims
- Test: 10,000 claims (unseen until final evaluation)
- **Model Training:** .fit() on preprocessed training data
- **Prediction:** .predict() on test data → evaluated against true labels (y_test)

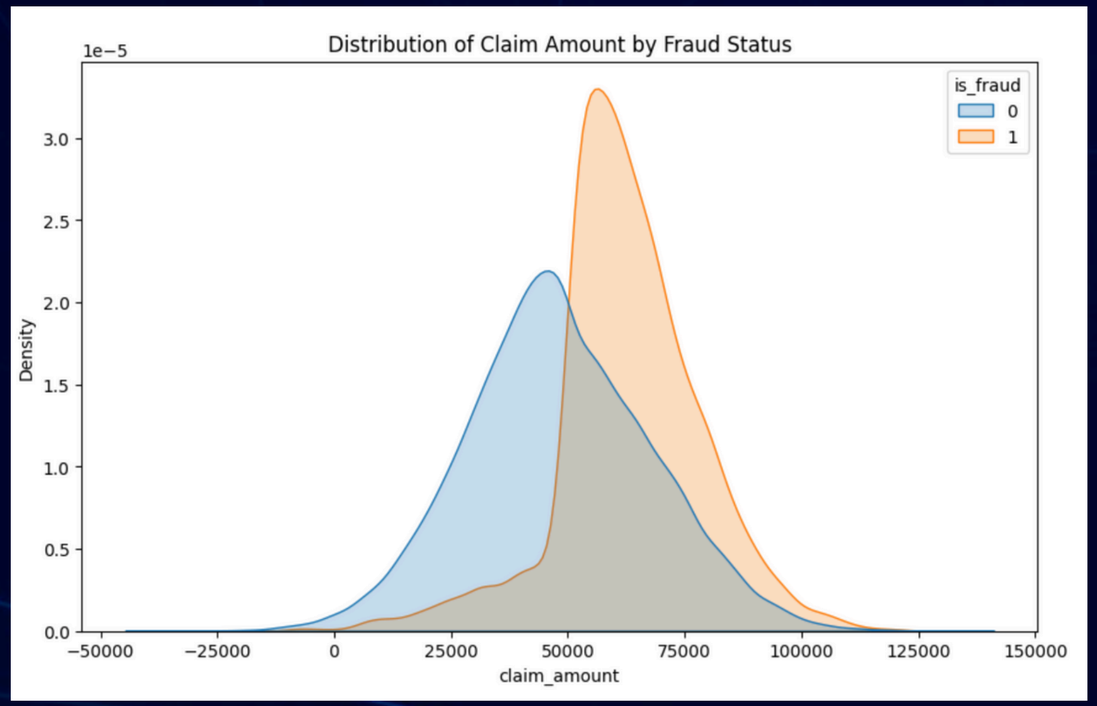
BEST PRACTICES APPLIED

- Strict separation of training & test sets → prevents leakage, ensures unbiased results
- Multiple algorithms tested under identical conditions → fair comparison
- Evaluation based on Precision, Recall, F1, ROC-AUC (not just accuracy, due to class imbalance)

ADDITIONAL IMPLEMENTATION DETAIL

- Tried **K-Fold Cross Validation** during model tuning → confirmed stability of results across different splits.
- Used **class weights (Random Forest)** and **scale_pos_weight (XGBoost)** to handle imbalance.

VISUALIZATIONS & EXPLORATORY DATA ANALYSIS (EDA)

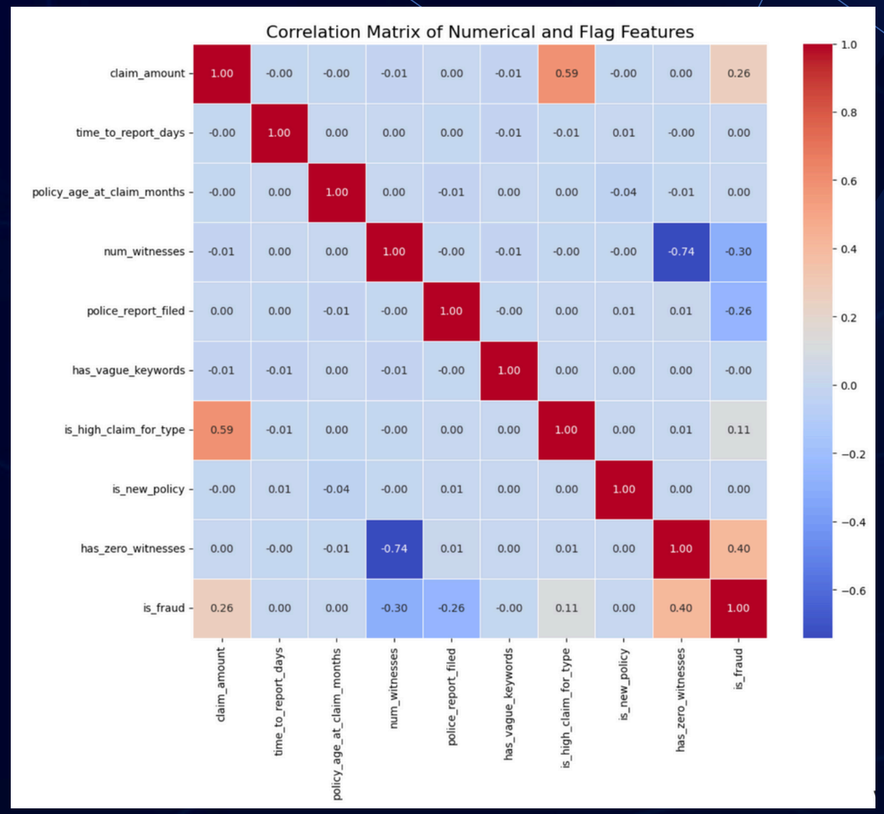


Fraudulent Claims = Higher Value

- Visualization: KDE plot of claim_amount
- Finding: Fraudulent claims peak near $60K vs. $45K for legitimate ones.
- Reason: Fraudsters inflate claim amounts.
- Implication: claim_amount is a top predictor and part of the Fraud Triangle.

Why It Matters

1. Validated the Fraud Triangle (high amount + no witnesses + no police report).
2. Guided feature prioritization for model building.
3. Boosted interpretability for investigators → fraud signals are clear and actionable.



Fraud Linked to Claim Circumstances

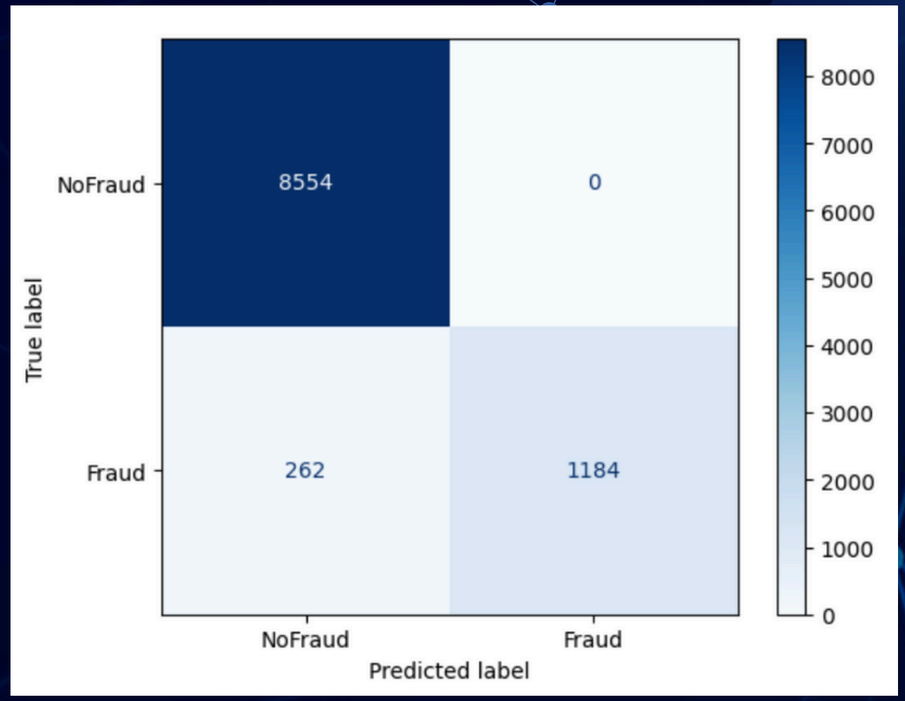
- Visualization: Correlation Matrix heatmap
- Finding:
  - has_zero_witnesses strongly correlates with fraud (+0.40).
  - police_report_filed shows a negative correlation (–0.26).
- Reason: Fraudulent claims lack verifiability (no witnesses, no police).
- Implication: Confirms that objective claim circumstances outweigh demographics (e.g., occupation, location, credit score) in detecting fraud.

SOLUTION & MODEL PERFORMANCE

Champion Model: Random Forest Classifier

After testing Logistic Regression, XGBoost, and Random Forest, the Random Forest emerged as the champion model, balancing fraud detection with unmatched reliability.

Metric	Logistic Regression (Baseline)	XGBoost	Random Forest (Champion)
Recall (Fraud Detection)	0.81	**0.82**	**0.82**
Precision (Prediction Accuracy)	0.98	0.97	**1.00** ⭐
F1-Score	0.89	0.89	**0.90**
ROC-AUC Score	0.906	0.909	**0.911**
False Alarms (FP)	22	34	**0** ⭐



Key Insights from Results

- **82% Recall →** Model successfully captures 8 out of 10 fraudulent claims, preventing major financial leakage.
- **100% Precision →** Every claim flagged is truly fraud → no wasted investigations.
- **Balanced F1 = 0.90 →** Demonstrates reliability across detection and accuracy.
- **ROC-AUC = 0.911 →** Strong overall ability to separate fraud from legitimate claims.

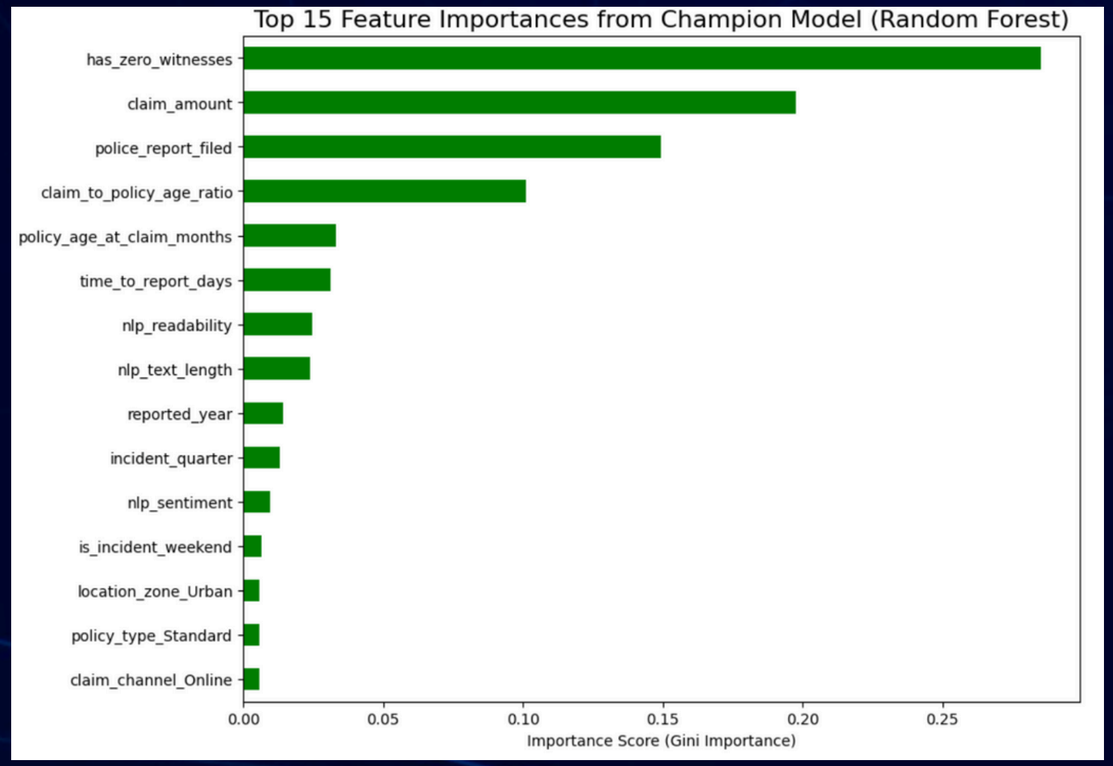
Why Random Forest Wins

- **Only model with perfect reliability (1.00 Precision).**
- **Guarantees investigator time is never wasted on false alarms.**
- **Matches XGBoost in recall while being more interpretable and stable.**

Model Limitations

- **Static Learning →** Needs periodic retraining to adapt to evolving fraud patterns.
- **Data Dependent →** Relies on accurate, consistent claim entry.
- **Individual Claim Focus →** Detects fraud at claim level, not yet at network level (fraud rings).

INSIGHTS & BUSINESS ACTIONS



Key Insights

1. **Fraud Driven by Circumstance, Not Demographics**
   - Model found demographics (occupation, location, credit score) have minimal predictive value.
   - Instead, the “Fraud Triangle” dominates:
     - High claim amount
     - Zero witnesses
     - No police report
   - Confirms fraudsters exploit unverifiable claim circumstances, not personal profiles.
2. **Feature Importance Confirms Focus Areas**
   - Top drivers: has_zero_witnesses, claim_amount, police_report_filed.
   - Demographic features ranked low → helps reduce investigator bias.

BUSINESS ACTIONS

01

SMART TRIAGE SYSTEM

- Use the model to auto-score claims.
- High-risk claims routed to Special Investigation Unit (SIU).
- Low-risk claims fast-tracked for payout.
- Cuts fraud losses while improving customer satisfaction.

02

DYNAMIC INTAKE QUESTIONING

- If claim matches the Fraud Triangle, system prompts handlers with additional probing questions.
- Acts as a deterrent and captures richer evidence at the filing stage.

03

INVESTIGATOR TRAINING ON OBJECTIVE RISK FACTORS

- Shift focus from demographics to claim circumstances.
- Improves fairness, efficiency, and alignment with data-driven insights.

Beyond Prediction: Understanding Fraud Archetype

- Distinct fraud personas from clustering analysis

01

CLUSTER 1: HIGH-VALUE MAXIMIZER

- Claim Amt: $80K+
- No witnesses
- Premeditated, high-payout fraud
- Top priority for Special Investigation Unit

02

CLUSTER 2: STANDARD OPPORTUNIST

- Largest group
- Claim Amt: Moderately high
- Few witnesses
- Inflated real events
- Detected well by Random Forest
- Core of automated triage

03

CLUSTER 3: ATYPICAL FABRICATOR

- Claim Amt: Low
- Many witnesses
- Unusual pattern, possible collusion
- Model struggle → Manual review
- Focus: Witness validation, ongoing monitoring

CONCLUSION & FUTURE WORK

1. Developed a Random Forest model as the champion solution.
2. Achieved 82% fraud detection (Recall) with 100% Precision — no false positives
3. Demonstrated that fraud is driven by claim circumstances, not demographics.
4. Highlighted the Fraud Triangle: high claim amount, no witnesses, no police report.
5. Reduced financial leakage and improved operational efficiency.
6. Enhanced customer experience and trust through faster legitimate payouts.

FUTURE WORK

01

DEPLOYMENT & A/B TESTING

- Deploy in shadow mode on live claims.
- Run A/B testing to measure real-world fraud savings and customer satisfaction uplift.
- Set up continuous performance monitoring & retraining.

02

ADVANCED ANALYTICS – FRAUD NETWORKS

- Move from single-claim detection to network analysis.
- Link claims across shared agents, repair shops, or policyholders to uncover fraud rings invisible at claim level.

03

HYPERPARAMETER TUNING & MODEL ENHANCEMENTS

- Explore advanced tuning to raise Recall above 82% while maintaining perfect Precision.
- Consider ensemble blending with XGBoost for marginal gains.

THANK YOU