# Fraud Detection System – Case Study Report

**1. Problem Statement**

Insurance companies face significant challenges due to fraudulent claims, which result in financial losses and reputational risks. Fraudulent activities often blend in with legitimate claims, making it difficult to detect using traditional rule-based systems.

**Objective**: Build a **Fraud Detection System** that:

* Identifies fraudulent claims with high accuracy.
* Improves recall (catching more fraud) while minimizing false alarms.
* Simulates real-time fraud detection to support investigation teams.

**2. Dataset Description**

* **Rows:** 1000
* **Columns:** 40 (after cleaning → 145 features after encoding)
* **Target:** fraud\_reported (1 = Fraud, 0 = Legitimate)
* **Class Balance:** ~25% Fraud vs 75% Non-Fraud (imbalanced).

**Key Features:**

* Customer info: age, occupation, education, relationship.
* Policy info: policy deductible, premium, umbrella limits.
* Incident details: type, severity, collision type, date/time.
* Claim info: total, injury, property, vehicle claim amounts.
* Vehicle info: make, model, year.

**3. Data Preprocessing**

* Removed useless columns (policy\_number, insured\_zip, incident\_location, \_c39).
* Converted date columns into derived features (year, month, day\_of\_week).
* Replaced ? with Unknown and filled missing categorical values.
* One-Hot Encoded categorical variables (final dataset: 145 columns).
* Target (fraud\_reported) mapped to numeric (1/0).

**4. Exploratory Data Analysis (EDA)**

* **Fraud Distribution:** Only 25% of claims were fraudulent → clear imbalance.
* **Claim Amounts:** Fraud claims often had **higher variability** in total claim amounts.
* **Incident Types:** Fraud more common in *Single Vehicle* and *Vehicle Theft* cases.
* **Auto Makes:** Certain brands (e.g., *Chevrolet, Dodge*) appeared more frequently in fraud cases.
* **Property Damage:** Fraud often associated with “Unknown” property damage responses.

📊 **Visuals created:**

* Fraud vs Non-Fraud distribution.
* Claim amount distributions by fraud status.
* Incident type vs fraud.
* Auto make vs fraud.
* Property damage vs fraud.
* Confusion Matrix, ROC Curve, Feature Importance.
* Fraud Probability Distribution.

**5. Model Development**

We experimented with three approaches:

1. **Logistic Regression (Baseline)**
   * Accuracy: 75%
   * ROC-AUC: 0.60
   * Recall for fraud: 0 → failed to detect fraud.
2. **Random Forest**
   * Accuracy: 73%
   * ROC-AUC: 0.82
   * Fraud recall: 0.10 → improved, but still low.
3. **Random Forest + SMOTE (Final Model)**
   * Accuracy: 76%
   * ROC-AUC: 0.77
   * Fraud recall: **0.27** → caught more fraud cases.
   * Fraud precision: 0.52 → about half of flagged frauds were true frauds.

📌 **Tradeoff**: Fraud recall improved significantly, even though AUC dropped slightly — acceptable in fraud detection (better to flag more frauds than miss them).

**6. Real-Time Fraud Monitoring Simulation**

We built a function to simulate live transaction monitoring:

* Each claim is scored with fraud probability.
* If fraud probability ≥ 0.5 → ⚠️ *ALERT triggered*.
* Otherwise → ✅ *Legitimate Transaction*.

**Example Output:**

Transaction 498: ✅ Legitimate Transaction (fraud risk 0.46)

Transaction 130: ✅ Legitimate Transaction (fraud risk 0.24)

Transaction 758: ⚠️ ALERT: Fraud risk 0.72 → Review Needed

This demonstrates how the model can integrate into a live fraud monitoring pipeline.

**7. Insights & Business Recommendations**

* **Imbalanced Data:** Fraud cases are rare but critical. Balancing techniques like SMOTE or cost-sensitive learning improve detection.
* **Key Predictors:** Claim amounts, incident severity, property damage, and auto make/model strongly influence fraud risk.
* **High-Risk Patterns:**
  + Unusual claim amounts relative to incident type.
  + Missing/“Unknown” values in property damage or police reports.
  + Certain auto makes involved disproportionately.
* **Business Actions:**
  + Prioritize high-risk claims for manual review.
  + Flag suspicious patterns (e.g., multiple single-vehicle incidents with high claims).
  + Integrate model outputs into investigation dashboards for real-time monitoring.

**8. Deliverables Completed**

✔ Cleaned & processed dataset  
✔ Comprehensive EDA report with visuals  
✔ Multiple fraud detection models (baseline + advanced)  
✔ Real-time fraud detection simulation function  
✔ Fraud investigation dashboard visuals  
✔ Final report with insights & recommendations

**✅ Conclusion**

This project successfully built a **Fraud Detection System** that combines:

* **Data-driven insights** (EDA).
* **Machine learning models** (Random Forest + SMOTE).
* **Real-time monitoring simulation**.

Although the model is not perfect (fraud detection recall = 27%), it significantly outperforms a baseline system. With further tuning (e.g., XGBoost, deeper feature engineering, business rules), detection rates can be improved further.

The project is **ready for presentation** as a professional fraud detection case study.