Fraudulent Claim Detection Case Study Report

🔍 Key Findings:

- The dataset exhibits **clear class imbalance**, reflecting real-world insurance fraud scenarios. Approximately **80% of claims are legitimate**, while **20% are fraudulent**.
- To ensure robust model performance, the data was **split into training and validation sets**. The training set was used to build models, and the validation set helped assess their generalization.
- Extensive **data preprocessing** was conducted, including:
    - Handling missing values
    - Dropping redundant columns
    - Data Type conversions
    - Feature scaling
    - Balancing techniques (e.g., random sampling)
    - Label encoding for categorical variables
- A thorough **Exploratory Data Analysis (EDA)** was performed to uncover relationships between input features and fraudulent outcomes. This step highlighted the importance of domain understanding in minimizing fraud risks.

🧠 Methodology and Models Applied:

Two machine learning classification models were evaluated:

1. **Logistic Regression**
2. **Random Forest Classifier**

Models were trained on the prepared training data and validated on the holdout set. Performance was assessed using **multiple metrics** beyond accuracy — namely **Sensitivity (Recall), Specificity, and Precision** — due to the importance of minimizing both false positives and false negatives in fraud detection.

📊 Model Performance Summary

Model	Accuracy	Sensitivity/Recall	Specificity	Precision
Logistic Regression	83%	76%	85%	63%
Random Forest	80%	55%	88%	61%

☑ Conclusion:

- Logistic Regression outperformed Random Forest in overall balance across evaluation metrics, particularly in identifying fraudulent claims (Recall = 76%).
- While Random Forest showed slightly higher specificity, its lower recall (55%) suggests more missed fraudulent claims, which could be risky in a real-world setting.
- This exercise demonstrates that careful preprocessing, model tuning, and metric-driven evaluation are key to building effective fraud detection models.

- In practical applications, ongoing retraining and refinement using fresh, high-quality data is essential to maintain model effectiveness over time.

📋 Recommendations for Improved Fraud Prevention

- **Improve Data Quality**
  Improve claim data intake processes by enforcing mandatory fields and validating inputs. Include data points known to be predictive of fraud, such as prior claim history, policy bind time, and vehicle profile.

- **Real-Time Scoring during Claim registration**
  Integrate the fraud detection model into the live claims registration system to score each claim at submission. This helps proactively identify suspicious activity and reduces investigation delays.

- **Risk-Based claim Scrutiny**
  Fraud cases are relatively rare, so investigating every high-risk score manually is not cost-effective. Implement a tiered approach, this ensures efficient use of fraud investigation resources while minimizing exposure to risk.
  - Fully investigate claims above a high-risk threshold
  - Use automated checks or light human review for mid-range scores
  - Allow low-risk claims to pass through with minimal intervention

- **Continuous Model Improvement**
  Regularly retrain models with fresh data to capture evolving fraud tactics. Incorporate investigator feedback to refine prediction accuracy and align with real-world fraud trends.

- **Drive Awareness Campaigns**
  Use model insights to develop focused education and awareness campaigns for customers, agents, and internal staff. Preventive communication can deter fraudulent behaviour and promote compliance.