Fraudulent Claim Detection - Project Summary

1. Problem Statement:

Insurance fraud is a significant issue impacting the financial stability of insurance companies. The objective of this project is to detect fraudulent claims using machine learning models, based on various features derived from insurance data.

2. Methodology:

The approach includes data cleaning, exploratory data analysis (EDA), feature engineering, model building (Logistic Regression and Random Forest), and model evaluation.

3. Techniques and Tools:

- Python libraries: Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, imbalanced-learn
- Preprocessing: Handling nulls, encoding categorical variables, feature scaling
- Models: Logistic Regression, Random Forest
- Evaluation: Confusion Matrix, Accuracy, Precision, Recall, F1 Score, ROC-AUC

4. Visualizations and Insights:

- Univariate distributions revealed outliers and imbalanced class labels.
- Bivariate plots showed relationships between fraud and variables like incident type, severity, and claim amount.
- Correlation matrix helped identify redundant features.

5. Key Findings:

- Logistic Regression achieved 79% accuracy with strong recall for fraud detection (0.84).
- Random Forest performed better with 80.5% accuracy and ROC-AUC of 0.84.

- Fraudulent claims often involve high claim amounts and certain incident types.

- The model shows high precision for legitimate claims but moderate precision for fraudulent ones.

6. Actionable Outcomes:

- Deploy Random Forest model for fraud prediction due to its robustness.

- Further enhance performance using ensemble methods like XGBoost.

- Deploy as a real-time fraud detection tool with user interface integration.