Fraudulent Insurance Claim Detection

Final Project Report

Vijay Kumar & Murali Krishna Kondapally

Analysis Approach

- **Data Loading and Preprocessing**
- **Handling Missing Values and Redundant Columns**
- **Exploratory Data Analysis (EDA) and Visualization**
- **Train and Validation Data Split** (70:30 with Stratification)
- **Feature Engineering** (Dummy variables, Scaling, SMOTE)
- **Feature Selection** (RFECV, Random Forest Importance)
- **Model Building** (Logistic Regression, Random Forest)
- **Model Evaluation** (Metrics, Cutoff Adjustment)

Theory Question 1

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

- Data Cleaning and Preprocessing: Removed irrelevant features, fixed datatypes.
- EDA: Identified unusual distributions, outliers, and correlations.
- Class Imbalance: Treated with SMOTE to identify minority class (frauds).
- Feature Engineering: Created new feature such as claim_to_premium_ratio, incident_severity_score.
- Categorical Grouping: Rare categories in categorical variables were grouped to improve model robustness.
- Modeling: Employed machine learning models Logistic Regression and Random Forest to identify non linear pattern in data.

Theory Question 2

Which features are the most predictive of fraudulent behaviour?

Top features identified:

- incident_severity_score
- insured_hobbies_chess
- incident_severity_Minor Damage
- incident_severity_Total Loss
- witnesses

These features captured severity, financial risk, and behavioral patterns linked to fraud.

Theory Question 3

Based on past data, can we predict the likelihood of fraud for an incoming claim?

- Yes, we can predict the likelihood of frauds using machine learning models like Random Forest and Logistic Regression.
  - Sensitivity (Recall): 1.0000
  - Specificity: 0.9962
  - Precision: 0.9962
  - F1 Score: 0.9981
- Random Forest achieved ~85% accuracy and ROC–AUC of 0.88 on validation data.

Theory Question 4

What insights can be drawn from the model to improve fraud detection?

- Top predictors like *incident_severity_score*, *insured_hobbies_chess*, *incident_severity_Minor* can help in triaging claims for manual review.
- Model explains risk through feature importance, supporting interpretability.
- Threshold tuning based on precision-recall tradeoff allows better control over false positives and negatives.
- Dynamic updating: Fraud patterns may shift over time; hence the model should be retrained periodically.
- Operational deployment: Integrating this model into claims systems can automate the initial fraud risk scoring, accelerating response time.

Summary of Findings

Summary of Analysis and Findings

- Data Imbalance Identified:
  - The dataset showed class imbalance with significantly fewer fraudulent claims (fraud_reported = Y) than non-fraudulent ones. This can impact model performance if unaddressed.
- Feature Insights:
  - incident_severity_score, insured_hobbies_chess, incident_severity_Minor, witnesses are highly predictive of fraud.
- Model Performance:
  - Logistic Regression and Random Forest showed decent performance.
  - After hyperparameter tuning, Random Forest improved Recall from 0.92 to 0.98 and F1 Score is .98, balancing between capturing frauds and avoiding false alarms.
- Optimal Cutoff Threshold:
  - Fixed threshold (0.45) was suboptimal.

Business Implications

Business Implications

- Fraud Loss Reduction: Early detection of fraud helps in preventing financial losses, especially by flagging high-value suspicious claims (total_claim_amount).
- Operational Efficiency: Automating the detection using models will help reduce the burden on human investigators by prioritizing high-risk claims.
- Better Resource Allocation: High-risk cases can be fast-tracked to special investigation teams, while low-risk claims can be approved faster—improving customer satisfaction.
- Regulatory Compliance: Having a well-documented and explainable model helps demonstrate due diligence to insurance regulators and audit teams

THANK YOU!