Fraudulent Claim Detection Project Report

1. Problem Statement

Global Insure faces significant financial losses due to fraudulent insurance claims. Their current manual inspection process is slow and inefficient, often identifying fraud too late. The company aims to leverage data-driven insights to classify claims as fraudulent or legitimate early in the approval process, reducing financial losses and optimizing claims handling.

2. Business Objective

The primary business objective is to develop a predictive model that accurately classifies insurance claims as either fraudulent or legitimate. This model will be based on historical claim data and customer profiles, enabling the early detection of potentially fraudulent claims.

3. Methodology

The project followed a structured approach, encompassing several key stages:

- **Data Preparation**: The initial phase involved loading the dataset and conducting a preliminary examination. This included previewing the data, checking its dimensions, and understanding the column descriptions and data types.
- **Data Cleaning**: This step focused on refining the dataset to ensure its quality and suitability for analysis.
    - Handled missing values in various columns, including collision_type (filled with 'NA'), property_damage and police_report_available (filled with the mode), and umbrella_limit (filled with 0).
    - Corrected inconsistencies, such as replacing '?' in police_report_available with 'NA'.
    - Addressed an issue in the incident_date column where some dates were incorrectly formatted with the year 2018, changing these to 2017.
    - Removed the extraneous _c39 column.
- **Train-Validation Split**: The dataset was partitioned into training and validation sets, with a 70-30 split, to facilitate model training and evaluation.
- **Exploratory Data Analysis (EDA)**:
    - **EDA on Training Data**: A thorough exploration of the training data was conducted to understand the characteristics of the variables and their relationships with the target variable, fraud_reported. This involved:
        - Analyzing the distribution of the target variable.
        - Examining categorical variables, including high-cardinality ones.

        - Analyzing numerical variables.
        - Exploring relationships between variables, such as total_claim_amount and other claim-related amounts.
    - **EDA on Validation Data**: While marked as optional, a similar EDA process on the validation data would have been beneficial to ensure consistency with the training data and to assess how well the training data represents the validation data.
- **Feature Engineering**: New features were derived to potentially improve the model's predictive power. For example, new features were created, such as incident_month, incident_weekday, claim_amount_ratio, injury_vehicle_ratio, and property_vehicle_ratio.
- **Model Building**: Two classification models were developed:
    - **Logistic Regression**: Regularized logistic regression, with hyperparameter tuning.
    - **Random Forest**: An ensemble method known for handling non-linear relationships.
- **Prediction and Model Evaluation**: The trained models were used to predict fraud likelihood on the validation set. Model performance was evaluated using metrics such as the confusion matrix, accuracy, precision, recall, and F1-score.

4. Techniques Used

The project employed a variety of techniques:

- **Data Preprocessing**: Handling missing values, correcting inconsistencies, and formatting data.
- **Exploratory Data Analysis (EDA)**: Visualizations and statistical analysis to understand data characteristics and identify potential predictors of fraud.
- **Feature Engineering**: Creating new features from existing ones to improve model performance.
- **Model Selection**: Training and comparing Logistic Regression and Random Forest models.
- **Hyperparameter Tuning**: Optimizing model parameters to improve performance. For Logistic Regression, this included tuning the regularization parameter 'C'.
- **Model Evaluation**: Assessing model performance using appropriate metrics for imbalanced classification problems.
- **Feature Selection**: Recursive Feature Elimination with Cross-Validation (RFECV)

5. Key Insights

The analysis revealed several key insights:

- **Data Characteristics**: The dataset has 40 columns and 1000 rows. Several

columns had missing values that required imputation.

- **Target Variable Imbalance**: The fraud_reported class is imbalanced, with fewer cases of fraud, which is typical in fraud detection scenarios.
- **Predictive Features**: Certain features are more predictive of fraud than others. These include:
  - Claim-related amounts (total_claim_amount, injury_claim, property_claim, vehicle_claim)
  - Incident details (incident_type, incident_severity, incident_hour_of_the_day)
  - Customer behavior/history (months_as_customer)
- **Model Performance**:
  - The Logistic Regression model, after feature selection with RFECV and optimal cutoff tuning, achieved balanced performance with good sensitivity and specificity, making it interpretable and suitable for scenarios where understanding feature impacts is crucial.
  - The Random Forest model demonstrated strong predictive power, particularly in handling complex interactions, though it showed signs of overfitting.
  - Both models were evaluated on validation data, with the Logistic Regression generally outperforming Random Forest in accuracy and F1-score, suggesting it may be preferred for deployment.
- **Feature Importance**: Key features like claim amounts, incident severity, and customer tenure were highly predictive of fraud.

6. Recommendations

Based on the findings, the following recommendations are made:

- **Prioritize Key Features**: Global Insure should prioritize the identified key features (claim amounts, incident severity, customer tenure, etc.) in their fraud screening process.
- **Implement Predictive Model**: The Logistic Regression model, due to its performance and interpretability, can be implemented for initial screening of claims.
- **Continuous Monitoring**: The performance of the deployed model should be continuously monitored and updated as needed to adapt to evolving fraud patterns.
- **Further Investigation**: Claims flagged as potentially fraudulent by the model should be subject to further investigation.