**Exploratory Data Analysis of Insurance Claims Data**

**1. Introduction**

- **Problem Statement:** Global Insure faces significant financial losses due to fraudulent insurance claims. The current manual inspection process is inefficient and results in delayed fraud detection. The goal is to develop a data-driven model to classify claims as fraudulent or legitimate early in the process, minimizing losses and optimizing claims handling.
- **Business Objectives:**
    - Build a predictive model to classify insurance claims as fraudulent or legitimate.
    - Identify key features that indicate fraudulent behavior.
    - Enable early detection of fraudulent claims to reduce financial losses.
    - Improve the efficiency of the claims processing workflow.
- **Data Description:**
    - The dataset contains 1000 rows and 40 columns.
    - Key columns include:
        - fraud_reported (target variable)
        - Numerical features (e.g., age, total_claim_amount)
        - Categorical features (e.g., policy_state, incident_type)
        - Date/time features (e.g., incident_date)
    - The data dictionary (provided earlier) gives description of each column.

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

- **2.1. Data Loading and Initial Inspection**
    - The dataset was loaded into a Pandas DataFrame.
    - The first few rows were inspected to understand the data structure.
    - Column data types were checked to identify numerical, categorical, and date columns.

- Missing values were checked. Column '_c39' has all missing values and can be dropped.

- **2.2. Target Variable Analysis (fraud_reported)**
  - The target variable fraud_reported is categorical, with values 'Y' (fraudulent) and 'N' (legitimate).
  - The distribution of the target variable shows an imbalance:
    - Legitimate claims (N): 753
    - Fraudulent claims (Y): 247
  - The class imbalance suggests that the majority of claims are legitimate, which is typical in fraud detection scenarios.
  - This imbalance needs to be addressed during model training to avoid biased predictions.
  - **[Include Chart: Count plot of fraud_reported]**
- **2.3. Numerical Feature Analysis**
  - Distributions of numerical features were visualized using histograms and box plots.
  - Relationships between numerical features and the target variable were explored using box plots.
  - Key numerical features analysed: months_as_customer, age, policy_annual_premium, total_claim_amount, vehicle_claim, injury_claim, property_claim.
  - Observations:
    - total_claim_amount, vehicle_claim, injury_claim, and property_claim show right-skewed distributions.
    - There are potential outliers in policy_annual_premium and claim amount-related features.
    - Some numerical features show differences in distribution between fraudulent and legitimate claims.
  - **[Include Charts: Histograms and box plots for numerical features]**

- **2.4. Categorical Feature Analysis**
    - Distributions of categorical features were visualized using count plots.
    - Relationships between categorical features and the target variable were explored using stacked bar plots.
    - Key categorical features analysed: policy_state, insured_sex, insured_education_level, insured_occupation, insured_hobbies, insured_relationship, incident_type, collision_type, incident_severity, authorities_contacted, property_damage, police_report_available.
    - Observations:
        - Some categorical features have imbalanced distributions (e.g., insured_sex).
        - Certain categories within features show higher proportions of fraudulent claims. For example, 'Theft' in incident_type has higher number of fraud cases.
        - incident_severity appears to be a strong indicator of fraud.
    - **[Include Charts: Count plots and stacked bar plots for categorical features]**
- **2.5. Date/Time Feature Analysis**
    - Date features policy_bind_date and incident_date were converted to datetime objects.
    - Incident month and year were extracted from incident_date.
    - Distributions of incidents by month and year were plotted.
    - Observations:
        - The dataset contains incident dates for the year 2015.
        - There is some variation in the number of incidents across months.
    - **[Include Charts: Count plots for incident month and year]**
- **2.6. Feature Interactions**
    - Interactions between age and total_claim_amount were explored using a scatter plot, with points colored by fraud_reported.

- Observations:
    - No clear pattern of fraud based on the interaction between age and total claim amount is observed.
- **[Include Chart: Scatter plot of age vs. total_claim_amount by fraud_reported]**

- **2.7. Summary of EDA Findings**
    - The target variable fraud_reported is imbalanced.
    - Numerical features show varying distributions and potential outliers.
    - Categorical features exhibit different distributions and relationships with fraud.
    - Date features provide insights into incident timelines.
    - Feature interactions may reveal more complex patterns.
    - Further data preprocessing and feature engineering are necessary before building a predictive model.

**4. Next Steps**

- Data preprocessing:
    - Handle missing values and outliers.

      Handled missing values by imputing with mode and outliers by capping
    - Encode categorical variables.

      Created dummies of categorical variables
    - Scale numerical features.

      Scalar function is used to standardize the numerical features
    - Address class imbalance.

      Performed Random Oversampling to address class imbalance
    - Fixed datatypes
- Feature engineering:

- Create new features (e.g., ratios, aggregations).

  Created new features like claim ratio, claim_per_age by other variables

- Identified unique values and grouped wherever possible
- Select relevant features.

  Selected relevant features by dropping high correlated to avoid multicollinearity – We used RFECV, High P value and VIF techniques for selecting the features

- Model building:
  - Train and evaluate classification models by Logistic Regression and Random Forest
  - Tune model parameters.
- Model interpretation:
  - Analyse feature importance.
- Understand model predictions. Predictions on validation data