**Fraudulent Insurance Claim Detection: Final Report**

**1. Problem Statement -** Insurance fraud is a significant issue in the industry, leading to substantial financial losses annually. The goal of this project is to develop a predictive model that can accurately detect potentially fraudulent insurance claims based on historical data. The challenge lies in identifying meaningful patterns that distinguish genuine claims from fraudulent ones while addressing class imbalance and ensuring model generalizability.

## 2. Methodology

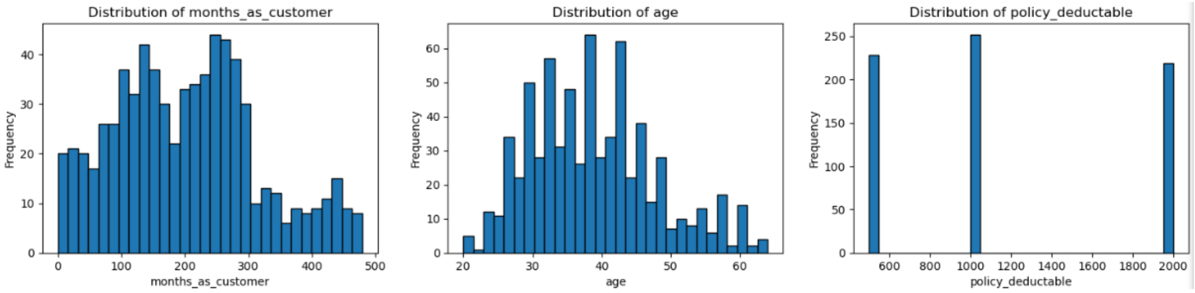
### Data Preprocessing

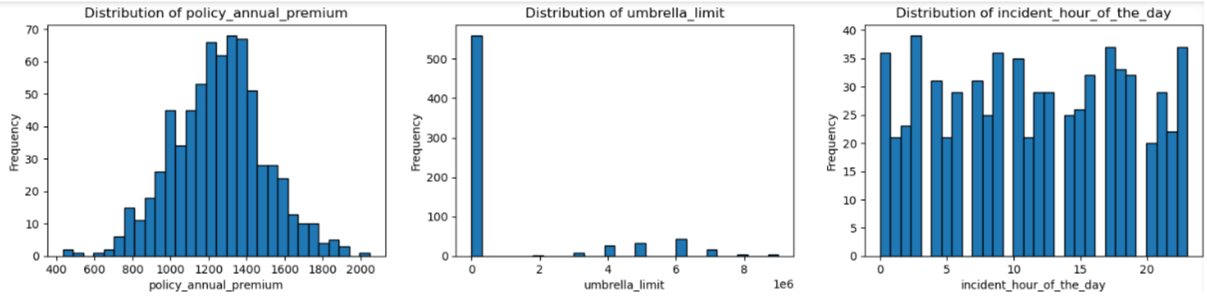
- Read and explored the insurance claims dataset.
- Identified and handled missing values, redundant, illogical, empty features along with the datatype fixing.

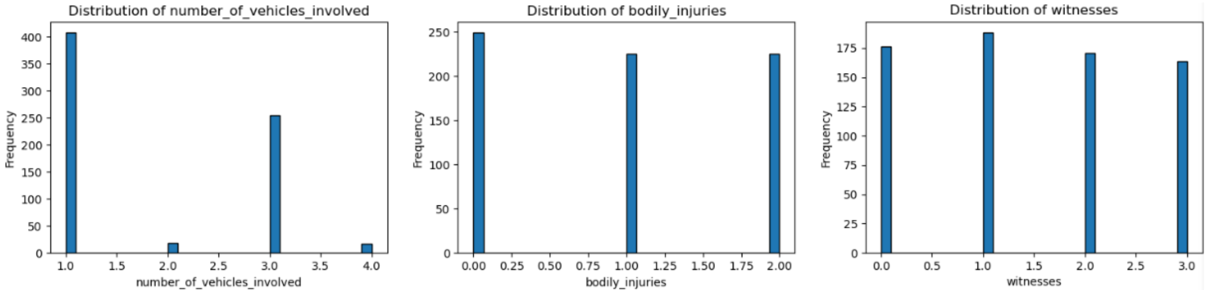
**Train and Test Data Split -** Split the dataset into 70% train and 30% test and use stratification on the target variable.

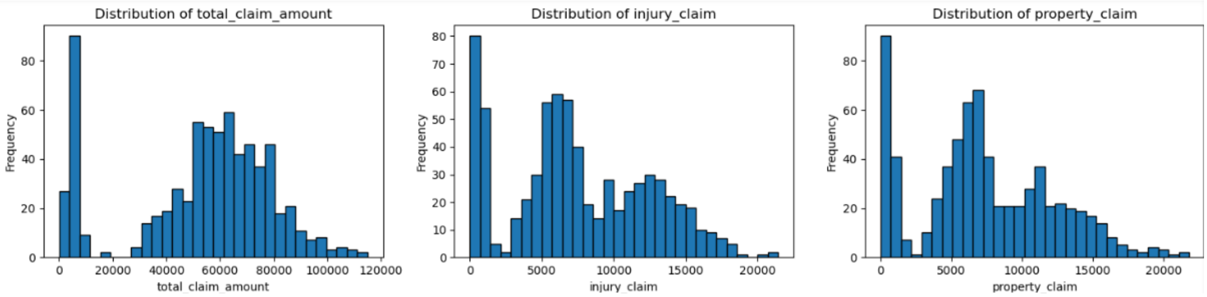
### Exploratory Data Analysis (EDA)

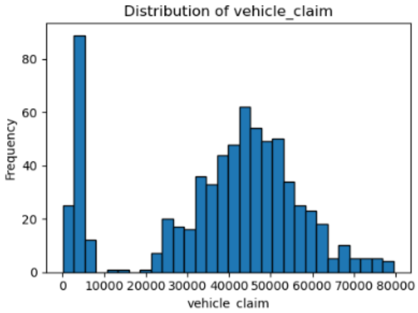
- Univariate and bivariate analysis on the numeric and categorical features highlights:
  - Multiple peaks in total_claim and injury_claim, property_claim, vehicle_claim could indicate subgroups in the data.
  - Normal (bell-shaped) distribution observed for policy premium - Most values cluster around the mean.
  - Data distribution for age feature is across various age groups between 20 to 70.
  - number_of_vehicles_involved are mostly 1 or 3 in the incident.
  - Umbrella_limit indicates majority data is around lower limit.
  - Frequency table are part of the python notebook which provide the insights on the categorical variables





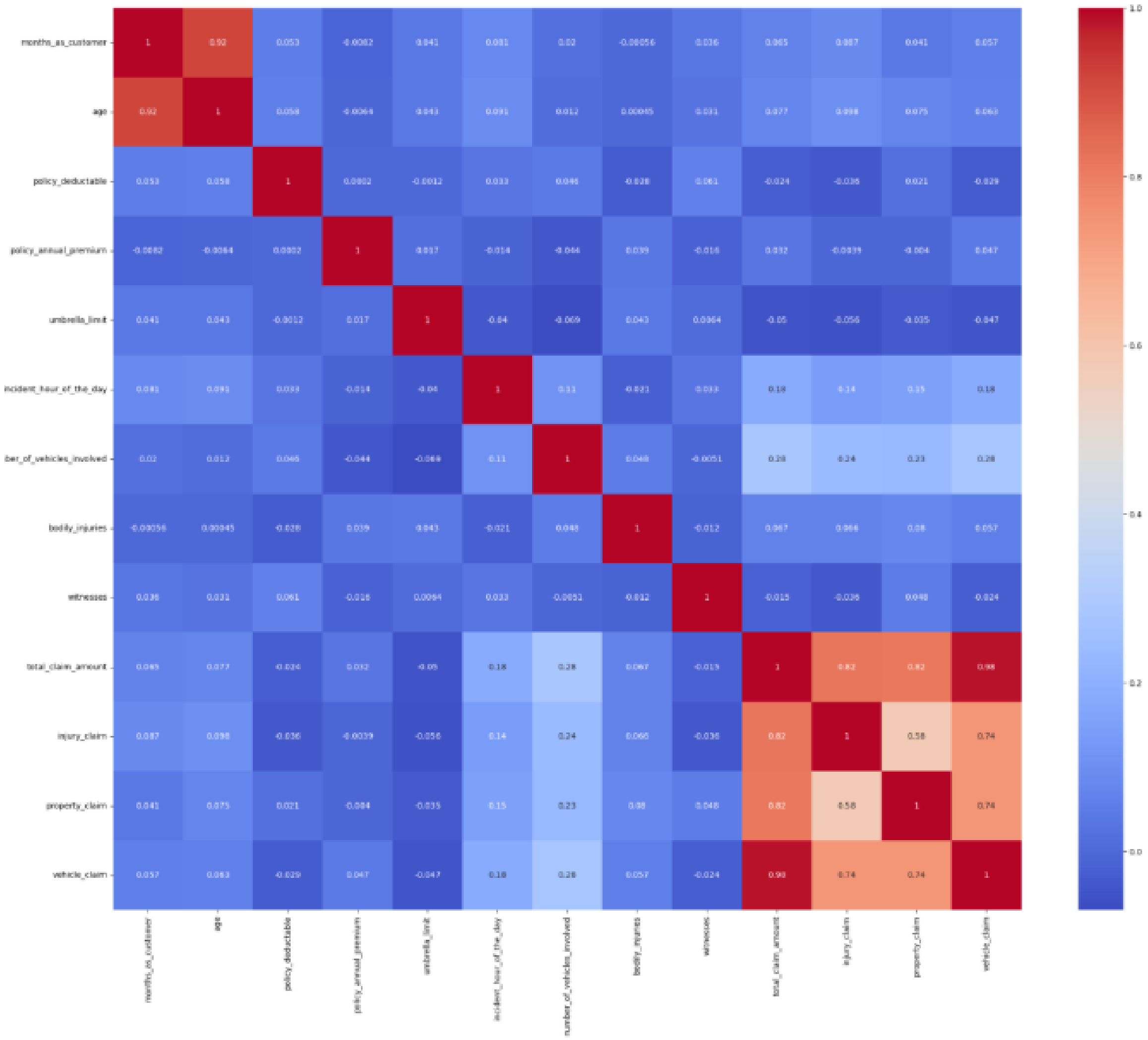






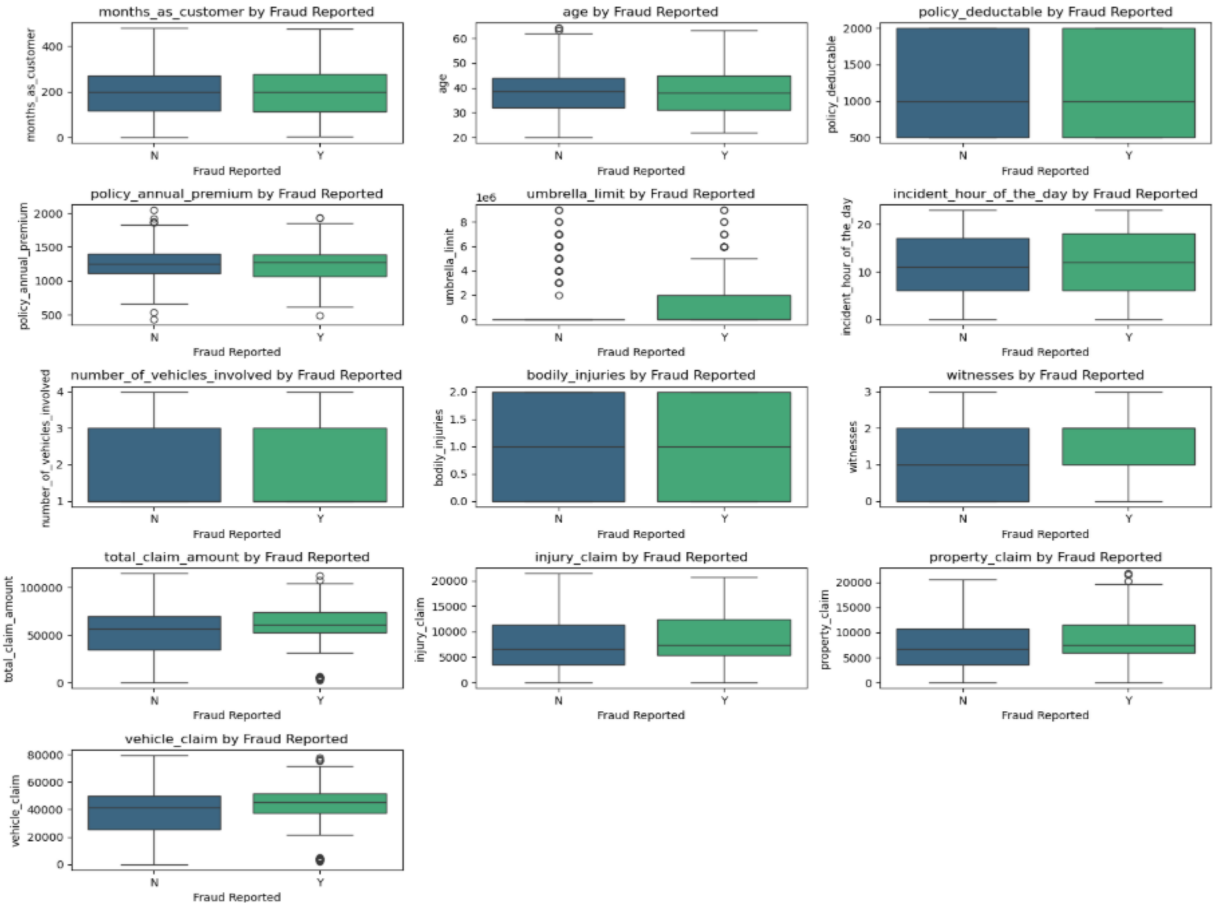
- Plotted correlation matrix - heatmaps to detect multicollinearity. The heatmap indicated collinearity between
    - age and month_as_customer

  - total_claim and injury_claim, property_claim, vehicle_claim exhibited highest collinearity



- Explored the relationships between numerical features and the target variable to understand their impact on the target outcome using appropriate visualisation techniques to identify trends and potential interactions.

  - The boxplot below indicates the fraudulent claims tend to have higher/lower claim amounts, more variability, or more outliers compared to non-fraud claims.



### Feature Engineering

- Created new features (e.g., age group bins, days_between_policy_and_incident, incident) from existing ones to enhance the model's ability to capture patterns in the data.
- Removed the redundant features which has minimal contribution towards predicting the target variable.
- Combined rare categories to reduce sparsity.
- Performed one-hot encoding on categorical variables.
- Applied feature scaling to numerical variables using StandardScaler to prevent features with larger values from dominating the model.

**Class Imbalance Handling -** on the training data by applying **RandomOverSampler** resampling technique to balance the data.This helps prevent the model from being biased toward the majority class and improves its ability to predict the minority class more accurately.

---

## 3. Modeling Techniques Used

### Logistic Regression

- Applied RFECV with logistic regression to select the most important features.
- Built baseline logistic regression model.

- Calculated VIF and p-values for feature significance and multicollinearity assessment.
- Made predictions and evaluated metrics: accuracy, precision, recall, specificity, F1-score.
- Adjusted probability cutoff to enhance sensitivity and balance performance.

### Random Forest Classifier

- Trained a random forest model using default parameters.
- Extracted feature importance and trained the model.
- Performed hyperparameter tuning using GridSearchCV (parameters such as max_depth, min_samples_leaf, class_weight) to enhance the performance of the model.
- Re-trained the model using best parameters.

## 4. Evaluation Metrics

- **Accuracy**: Overall correctness of the model.
- **Precision**: Percentage of predicted frauds that were actually fraudulent.
- **Recall (Sensitivity)**: Ability to detect actual frauds (important in minimizing false negatives).
- **Specificity**: Correctly identifying non-fraudulent claims.
- **F1 Score**: Harmonic mean of precision and recall.

---

## 5. Key Insights and Outcome

- **Top predictors** - Features such as incident_severity, total_claim_amount, policy_annual_premium, and days_between_policy_and_incident were highly predictive.
- **Class Imbalance** - Addressing class imbalance significantly improved the model's ability to detect fraudulent cases.
- **Threshold tuning** - Logistic regression with an adjusted probability cutoff can improve recall and balance the cost of misclassification.
- **Hyperparameter tuning** - Random forest - in terms of sensitivity and F1 score metrics are improved after hyperparameter tuning.