Fraudulent Detection Report

Analysis approach-

- Importing required libraries and Input data loading
- Inspect the Input data, Exploratory Data Analysis (EDA) and Visualization
- Train and Test data split
- Data Preparation
- Feature Scaling
- Looking at correlations
- Feature Selection using RFECV
- Model Building
- Model Evaluation and Prediction

Theory Questions

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

To analyze historical claims for fraud detection, the following approach was applied:

- **Data Cleaning and Preprocessing**: Removed irrelevant, redundant, and low-variance features that do not contribute meaningful signals.
- **Class Imbalance Handling**: Used oversampling techniques like `RandomOverSampler` to address the rarity of fraud cases.
- **Feature Engineering**: Created new features such as `days_between_policy_and_incident`, `age_group`, etc., which revealed latent behavioral patterns.
- **Categorical Grouping**: Rare categories in categorical variables were grouped to improve model robustness.
- **Statistical Analysis**: Explored categorical vs. target relationships and examined variable distributions using univariate and bivariate analysis.
- **Model-Based Evaluation**: Applied machine learning models (Logistic Regression, Random Forest) to detect non-linear patterns in the data.

These steps allowed for the detection of relationships and behaviors indicative of fraud, such as higher claim amounts or specific incident severity levels.

Theory Questions

2. Which features are the most predictive of fraudulent behaviour?

Based on feature selection and importance techniques such as `RFECV` and Random Forest importance scores, the following features were found to be the most predictive:

- **incident_severity** (Minor Damage, Total Loss)
- **total_claim_amount**
- **policy_annual_premium**
- **days_between_policy_and_incident**

These features either capture the scale or suspicious timing of an incident, or represent contextual indicators that correlate with fraud.

Theory Questions

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

Yes. A predictive model using Random Forest was developed with below performance:

- **Sensitivity (Recall)**: ~55%
- **Precision**: ~57%
- **F1 Score**: ~56%
- **Specificity**: ~87%

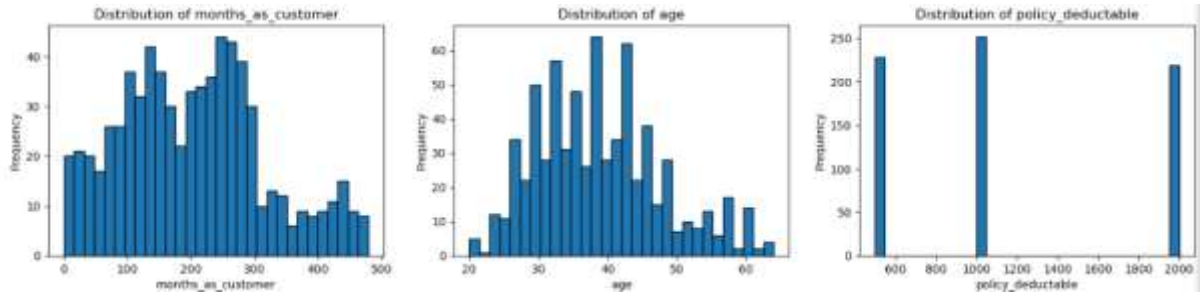
The model effectively identifies a significant portion of fraudulent claims with balanced precision, and can output probabilities to assess the fraud likelihood of incoming claims. Threshold tuning further refines the decision boundary depending on business needs.

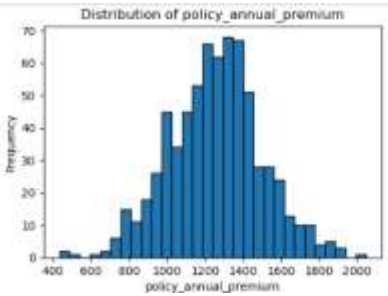
Theory Questions

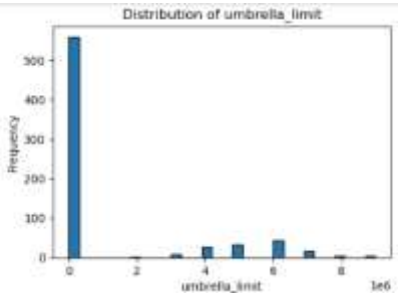
4. What insights can be drawn from the model that can help in improving the fraud detection process?

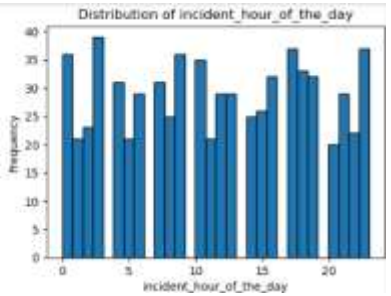
- **Top predictors** like `incident_severity` and `total_claim_amount` can help in triaging claims for manual review.
- **Model explains risk** through feature importance, supporting interpretability.
- **Threshold tuning** based on precision-recall tradeoff allows better control over false positives 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.

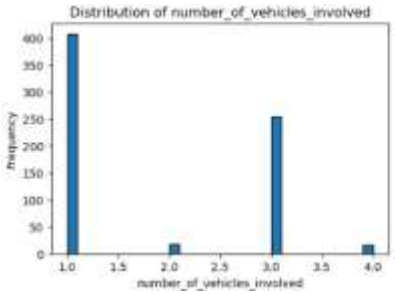
Univariate Analysis : analyze the single variable

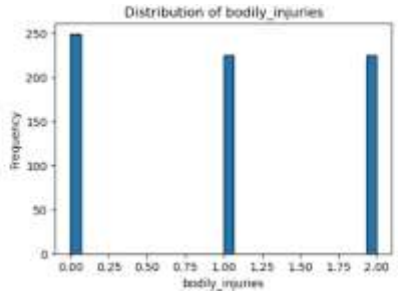


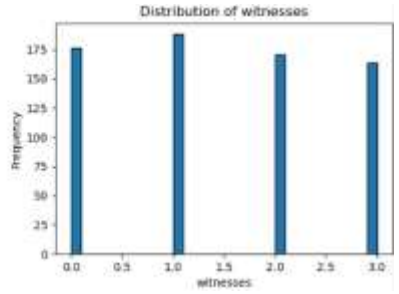


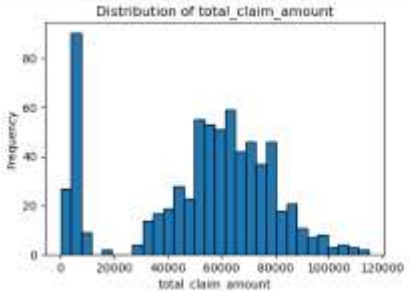


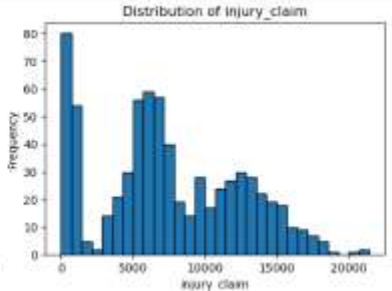


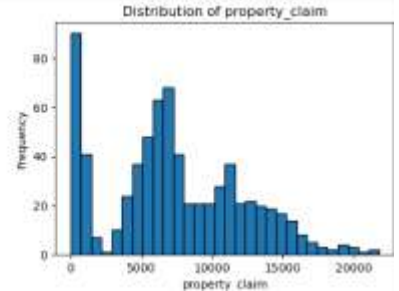


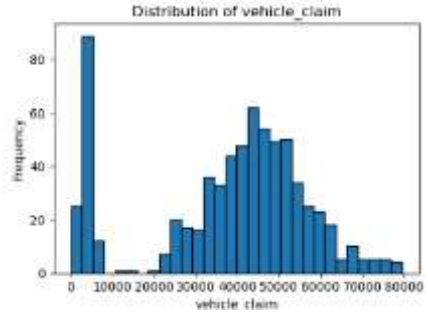








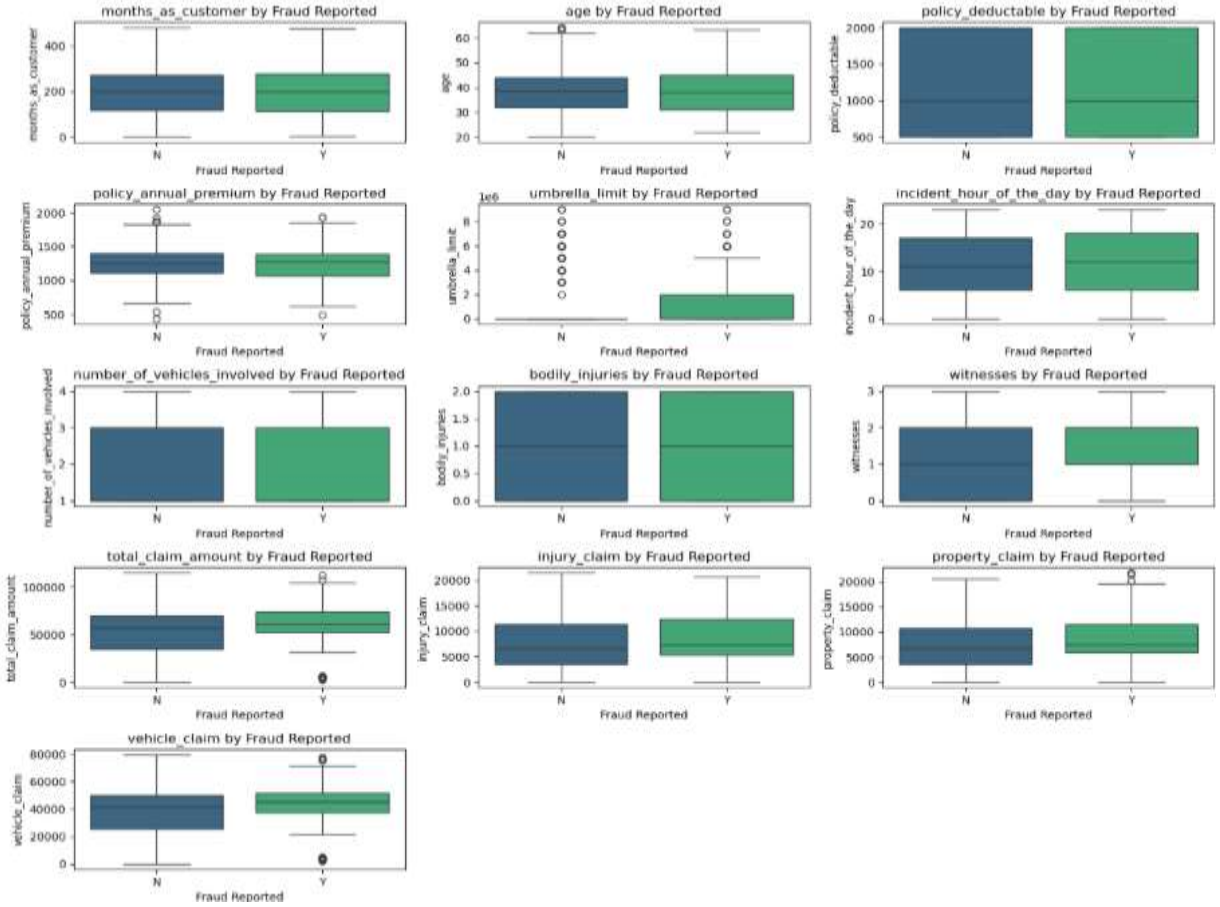


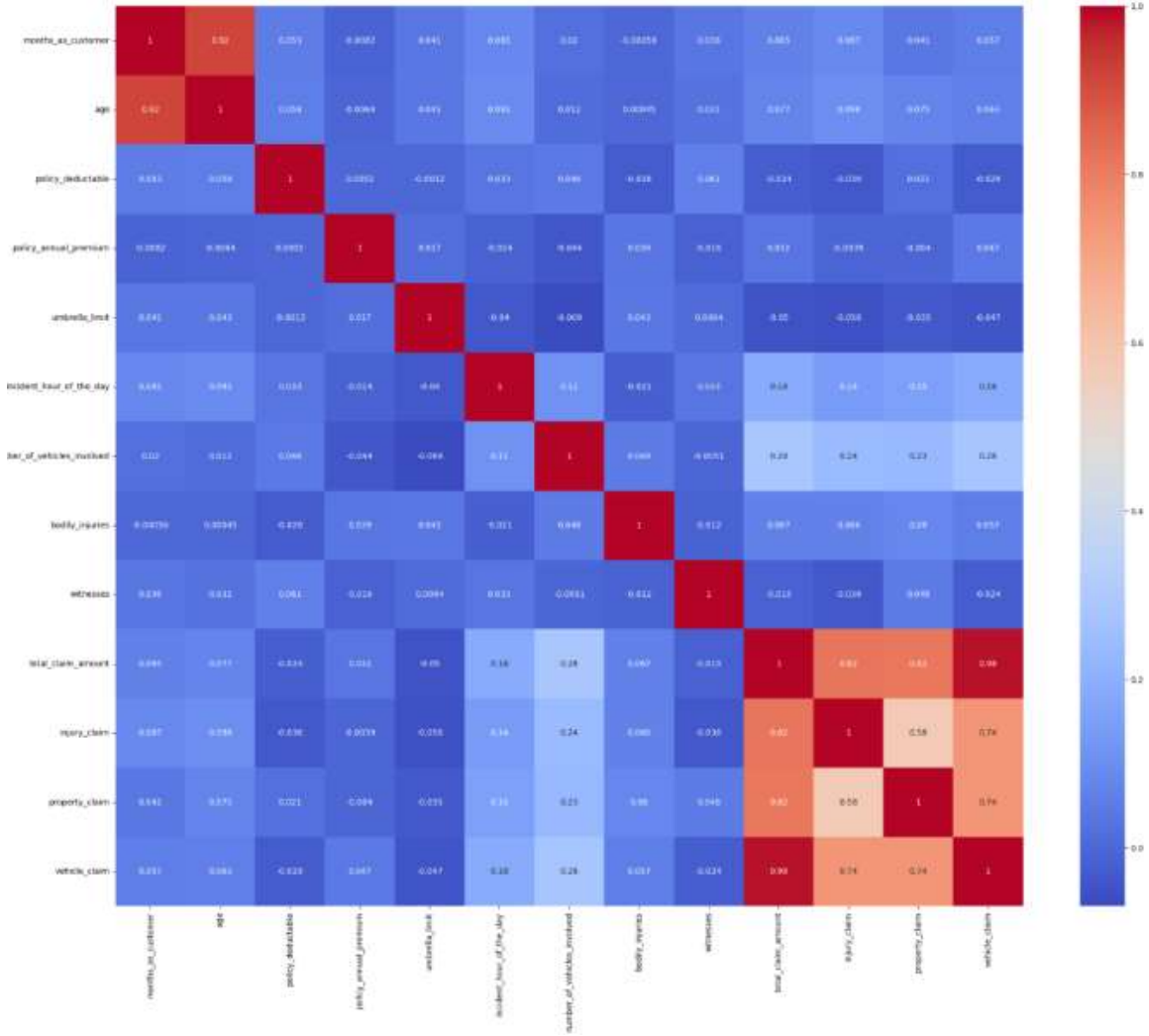


Multivariate Analysis : analyze the correlations

Based on the assessment and visualization -

- age and month_as_customer
- total_claim and injury_claim, property_claim, vehicle_claim exhibited highest collinearity
- The boxplot below indicates the fraudulent claims tend to have higher/lower claim amounts, more variability, or more outliers compared to non-fraud claims.





Summary

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 and total_claim_amount are **highly predictive** of fraud.
- Features like policy_annual_premium, days_between_policy_and_incident, and incident_hour_of_the_day also show strong importance in the Random Forest model.

· **Model Performance**:

- **Logistic Regression** and Random Forest showed decent performance.
- After hyperparameter tuning, **Random Forest improved Recall from 0.27 to 0.55** and F1 Score from 0.30 to 0.56, balancing between capturing frauds and avoiding false alarms.

· **Optimal Cutoff Threshold**:

- Fixed threshold (0.5) was suboptimal.
- Precision-recall tradeoff analysis suggests lowering threshold may be beneficial for better recall (detecting more frauds).

Business Implications :

1.Fraud Loss Reduction: Early detection of fraud helps in **preventing financial losses**, especially by flagging high-value suspicious claims (total_claim_amount).

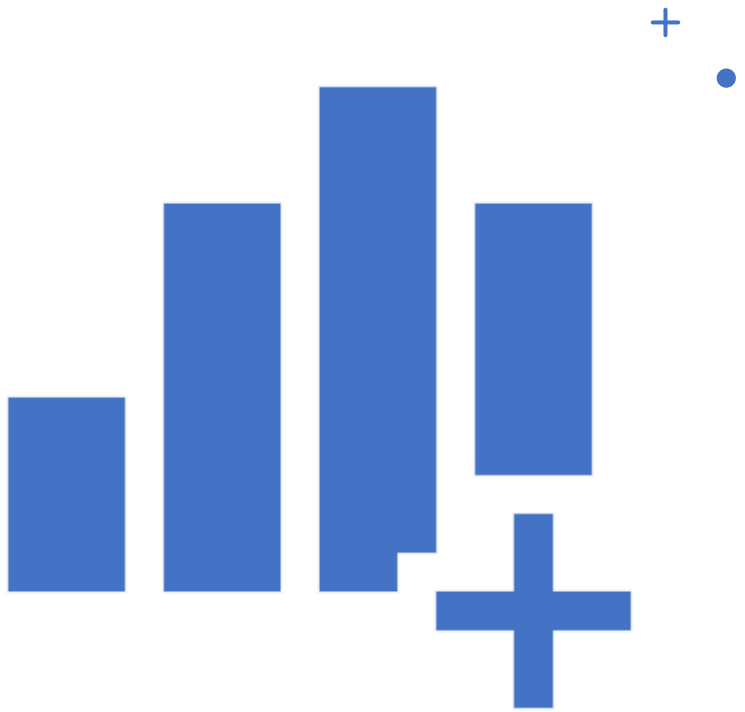
2.Operational Efficiency: Automating the detection using models will help reduce the burden on human investigators by **prioritizing high-risk claims**.

3.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**.

4.Regulatory Compliance: Having a well-documented and explainable model helps demonstrate due diligence to **insurance regulators** and audit teams

Recommendations :

- 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 improves the model's ability to detect fraudulent cases.
- Threshold tuning - adjusted probability cutoff can improve recall and balance the cost of misclassification.
- Hyperparameter tuning can improve the model performance significantly.



By leveraging the insights from this model, insurance firms can significantly improve fraud detection accuracy, reduce financial losses, and optimize resource allocation in claim investigations.