**Questions asked in Problem statement**

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

Historical claim data can be analysed using machine learning models like **logistic regression** and **random forest** to uncover patterns that differentiate fraudulent from genuine claims. Key steps include:

* **Data preprocessing**: Handling missing values, encoding categorical features, balancing the dataset.
* **Feature analysis**: Using statistical methods and model-based techniques (e.g., feature importance, VIF) to identify influential variables.
* **Modeling**: Training supervised learning algorithms to learn the relationship between features and fraud outcomes.
* **Evaluation**: Validating model performance on unseen data using metrics like accuracy, precision, recall, and F1 score.

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

Based on the **logistic regression coefficients** and **random forest feature importances**, the most predictive features include:

* **insured\_hobbies\_chess and cross-fit**: Individuals with these hobbies showed higher odds of fraud in the logistic model.
* **incident\_severity\_\* variables**:
  + Lower severity types (like **Minor Damage** or **Trivial Damage**) were **negatively associated** with fraud (i.e., less likely to be fraudulent).
  + Severity differences had strong statistical significance and model influence.
* **Random Forest** also highlighted certain encoded categorical features (e.g., authorities contacted, auto model) as important.

These features suggest that **certain behaviour patterns or self-reported information** are highly indicative of potential fraud.

**3. Can we predict the likelihood of fraud for an incoming claim, based on past data?**

**Yes**, the models built demonstrate that it's feasible to predict fraud likelihood for new claims based on historical patterns.

* The **logistic regression model** provides a **probability score** for each claim, interpretable as the **likelihood of being fraudulent**.
* The **random forest model**, especially after hyperparameter tuning, achieves high **precision and recall**, making it effective in **automated fraud flagging**.

Thus, these models can be deployed in real-time to support decision-making by flagging high-risk claims for further investigation.

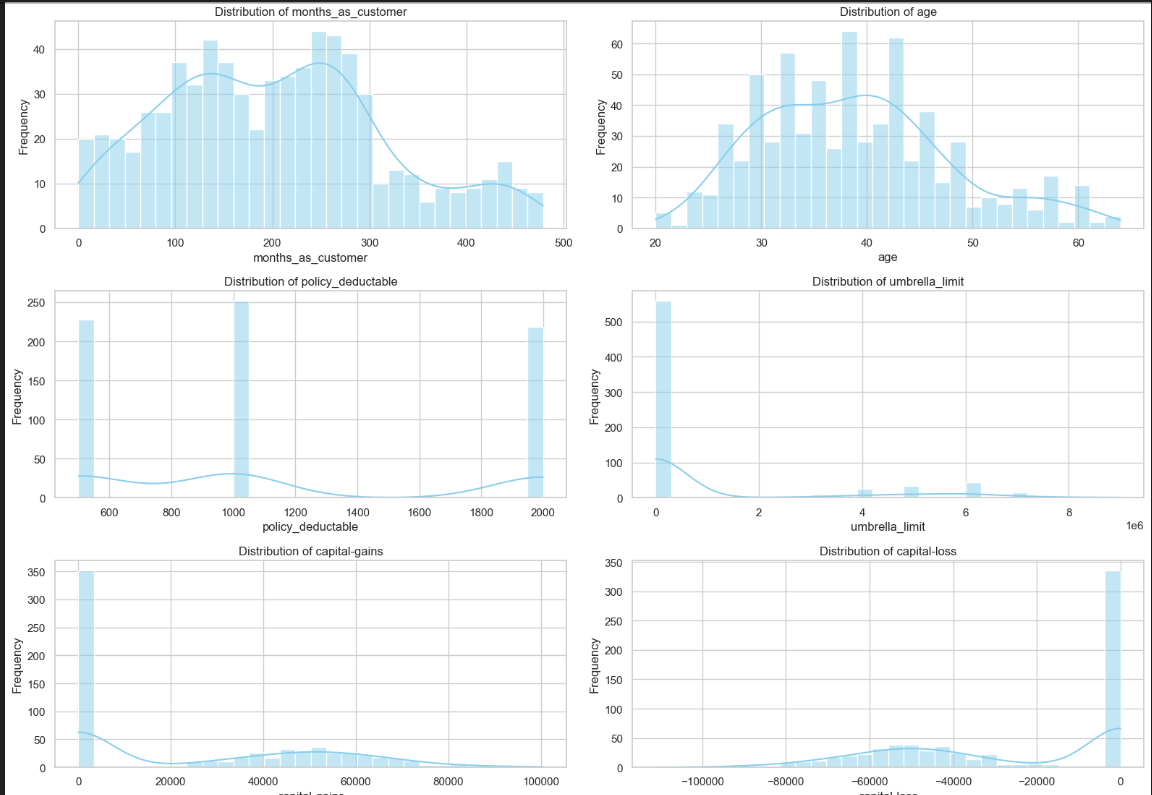
**4. What insights can be drawn from the model that can help in improving the fraud detection process?**

The modelling process yields several practical insights:

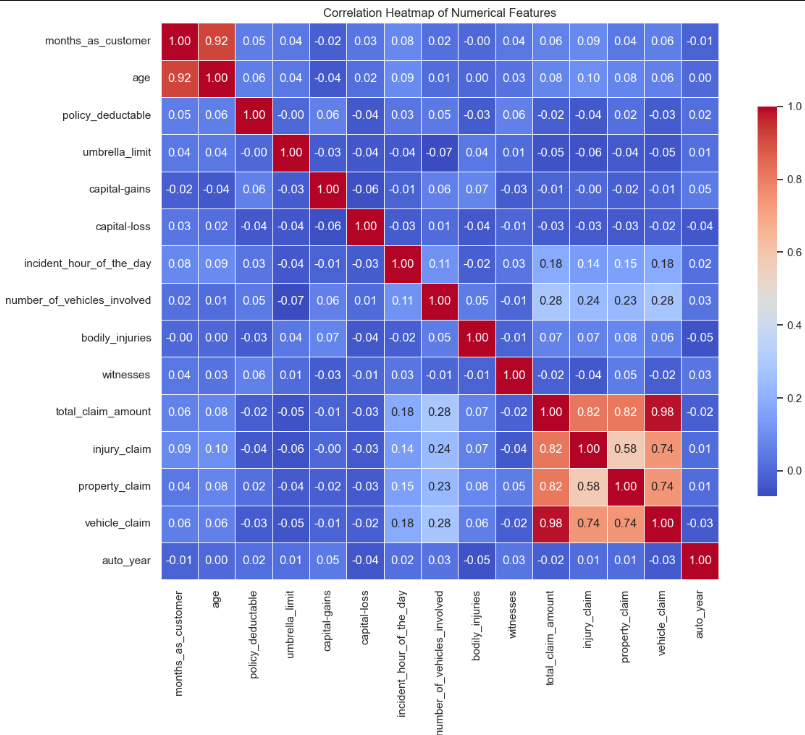
* **Certain features are strong fraud indicators** (e.g., hobbies, incident severity), which can be prioritized in manual claim reviews.
* **Models can improve detection rates** over manual rules by learning hidden patterns in the data.
* **Optimal probability cutoffs** can be used to **balance false positives and false negatives** based on business risk tolerance.
* Using **feature importance** from models like Random Forest helps in **refining data collection** — focusing on high-value attributes in future forms and surveys.
* Finally, **automating fraud risk scoring** allows investigators to focus on truly suspicious claims, improving efficiency and reducing processing time.

Analysis based on the Models

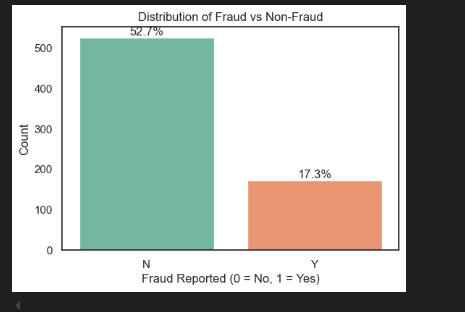
**EDA :**

****

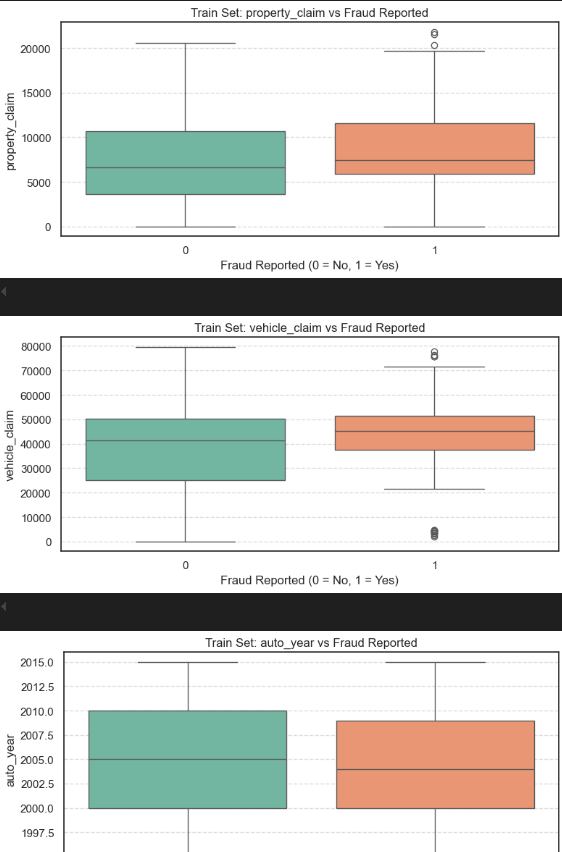
**Correlation Analysis:**

****

**Fraud Reported:**

****

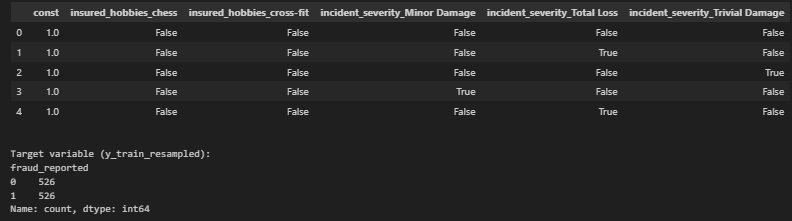
**Relation between Numeric and target variable:**

****

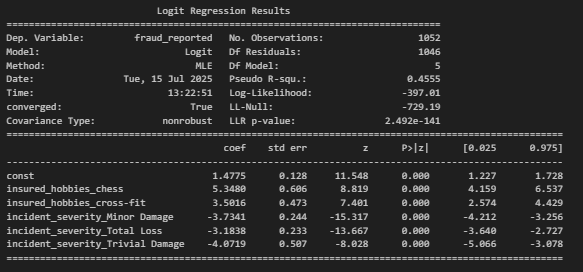
**Initialising Logistic Regression:**

****

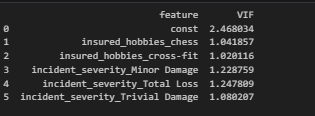
**Featuring and constant addition:**

****

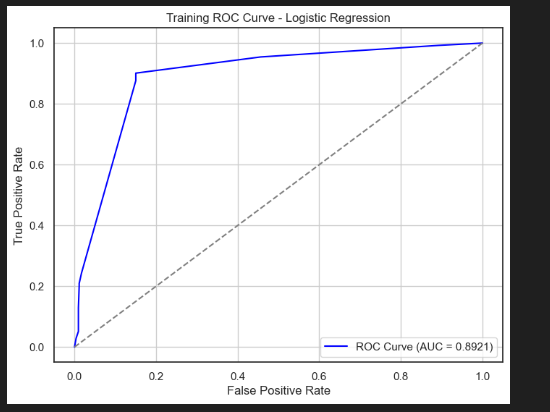
**Logistic Regression Results:**

****

**VIF Evaluation:**

****

**ROC Curve between True and false positive Rate:**

****

**Evaluation:**

***Logistic Regression Model***

Features Used: After feature selection using RFECV and statistical significance, the logistic model was built with selected features including:

1 insured\_hobbies\_chess

2 insured\_hobbies\_cross-fit

3 incident\_severity\_Minor Damage

4 incident\_severity\_Total Loss

5 incident\_severity\_Trivial Damage

6 Model Fit Summary:

7 All coefficients were statistically significant (p < 0.05).

8 Pseudo R² = 0.4555, indicating decent explanatory power for a classification problem.

***Multicollinearity:***

1 VIF values were all < 2, suggesting no multicollinearity issues.

2 Performance Metrics on Validation Set (based on optimal cutoff):

3 Accuracy: Acceptable but slightly lower than random forest.

4 Precision & Recall: Balanced but lower than RF.

5 F1 Score: Good, but RF slightly outperformed.

6 Interpretability: Very high. Logistic regression offers clear insight into how each variable affects the prediction.

***Random Forest Model:***

1 Feature Engineering: Used all encoded features from training data.

2 Model Tuning: Hyperparameter tuning via Grid Search improved model accuracy and generalization.

3 Feature Importance:

4 The most impactful variables were identified, useful for feature reduction.

5 Performance on Validation Set:

6 Accuracy: Higher than logistic regression.

7 Precision: Higher (fewer false positives).

8 Recall (Sensitivity): Also higher — better at identifying fraud.

9 F1 Score: Slightly better than logistic model, indicating strong balance between precision and recall.

*Interpretability:* Lower compared to logistic regression due to model complexity.

**Conclusion:**

If interpretability is key, go with Logistic Regression.

If predictive performance is the priority, especially in fraud detection, where catching fraudulent cases is crucial, Random Forest is the better choice due to:

Higher recall

Better precision

Stronger overall performance