Fraudulent Claim Case Study Report

Prepared for Global Insure

Prepared By: Adarsh Yadav and Saurabh Gupta

1. Problem Statement

Global Insure, a leading insurance provider, processes thousands of claims annually. However, a significant percentage of these claims are found to be fraudulent, resulting in substantial financial losses. Currently, the fraud detection process relies on manual inspection, which is both time consuming and inefficient. Often, fraudulent claims are identified only after payouts have been made.

To mitigate financial loss and streamline claims processing, Global Insure aims to adopt a data driven approach to detect and classify fraudulent claims early in the review process.

2. Business Objective

The goal is to develop a predictive model that classifies insurance claims as either fraudulent or legitimate, using historical claim data and customer profiles. Key features include:

- Claim amounts
- Customer demographics
- Claim types

The model will enable proactive fraud detection before claim approval, optimizing resource allocation and minimizing risks.

3. Key Questions Addressed

- How can historical data be leveraged to identify fraud patterns?
- Which features are most indicative of fraudulent behavior?
- Can we predict the likelihood of fraud in new claims?
- What actionable insights can the model offer to improve fraud detection?

4. Approach

1. **Data Preparation:**
    - Involves importing of the data and checking it
2. **Data Cleaning:**
    - Addressed missing values and inconsistencies.
3. **Train Validation Split 70-30:**
    - Standard Practice followed in model building
4. **EDA:**
    - Analyzed numerical and categorical variables.
    - Identified relationships with the target variable (fraud reported).
5. **Feature Engineering:**
    - Created and transformed variables for improved predictive power.
6. **Model Building:**
    - Built and tested Logistic Regression (LR) and Random Forest (RF) models.
7. **Predicting and Model Evaluation:**
    - Assessed model performance on test data.
    - Selected the bes tperforming model.
8. **Summary and Recommendations:**
    - Summarized the insights and provided Final Recommendations

We will emphasize on EDA, Feature engineering, Models and Recommendations in this report.

5. Exploratory Data Analysis (EDA)

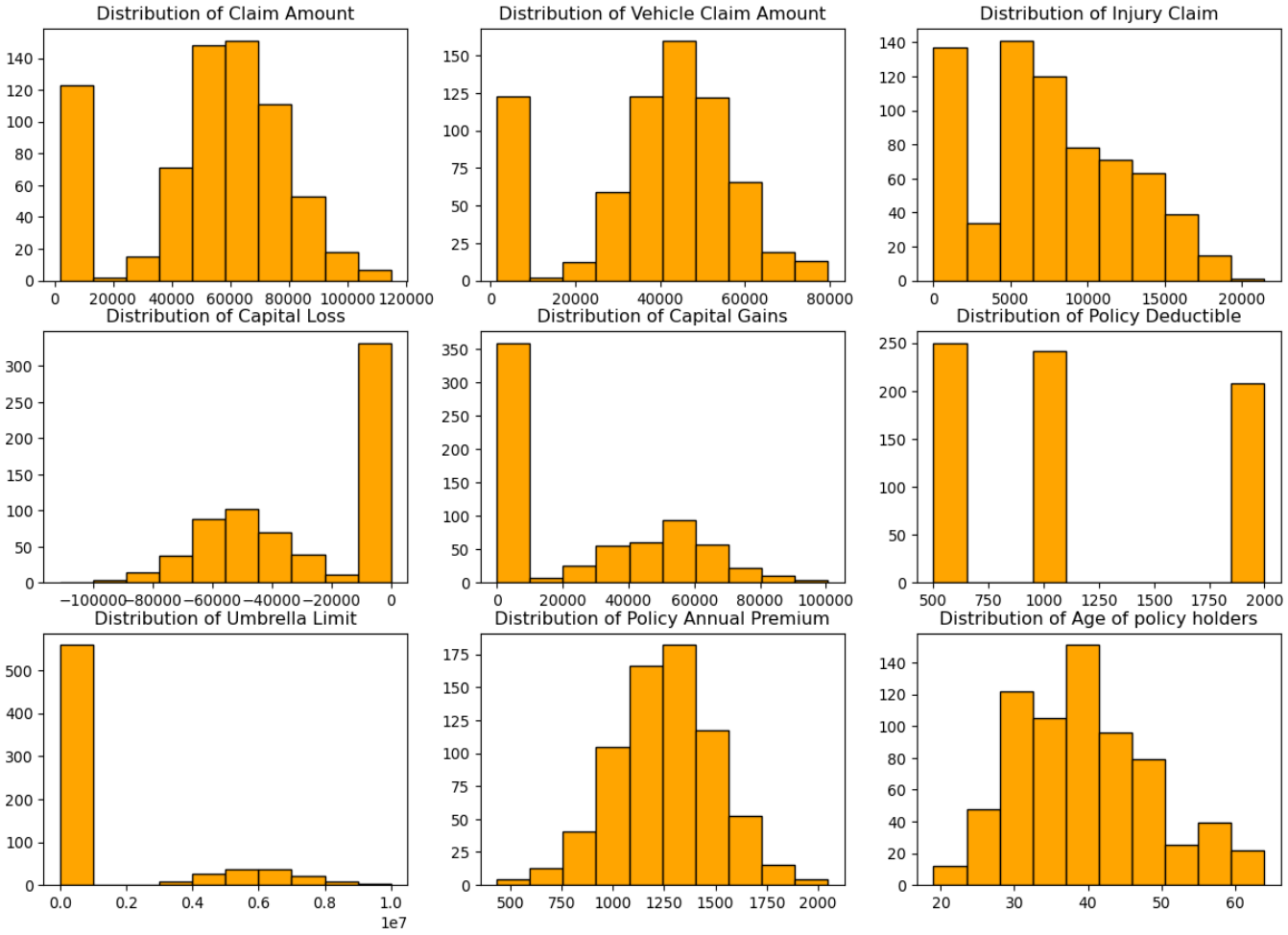
We have performed univariate, bi-variated and segmented EDA on the dataset

We can draw below insights from the distribution of the Histograms for various numerical features of the dataset:

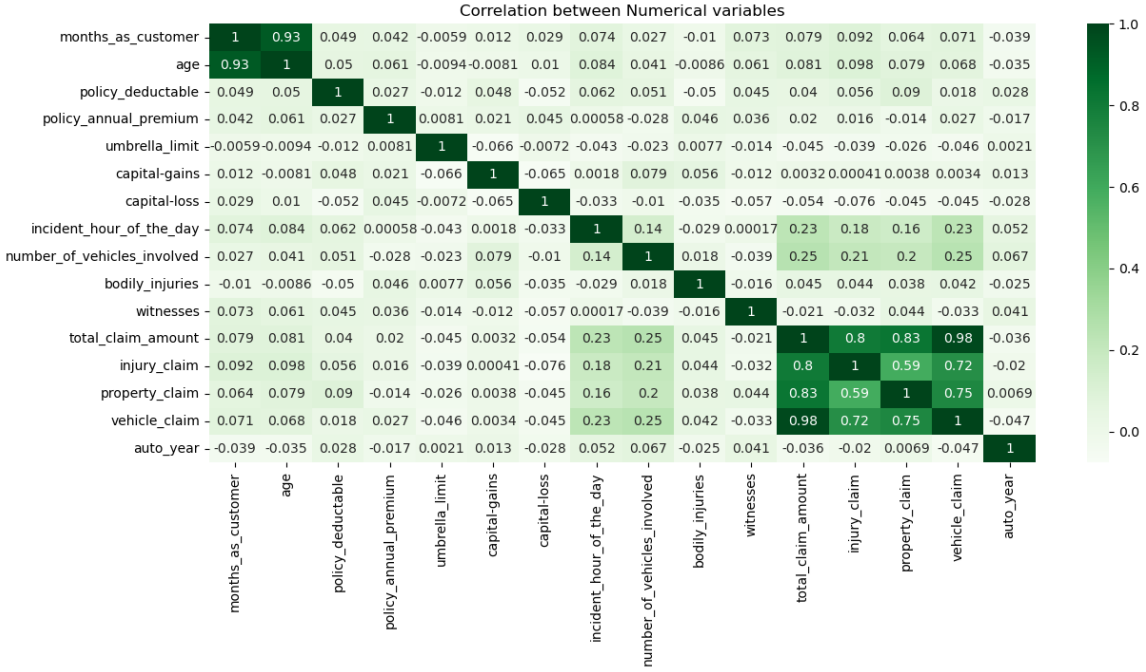
- Policy Holders are either claiming amount below 10K or mostly between 50K to 80K
- For Vehicle Claims mostly there's normal distribution and highest claim amounts are between 40K to 50K
- For injury claims the distribution is skewed towards left and most of the claim amount is within 10K
- A huge number of policy holders are not facing any capital loss

- Similar to capital loss a huge number of Policy Holders are not having any capital gains
- Policy Deductible is at 600, 1K and 2K
- Many policy holders has almost 0 umbrella Limit
- Annual premium is normally distributed with most of the policy holders having 1200 to 1400 premium
- Most of the policy holders are within 30 to 40 years of age

Distribution of Numerical Variables on complete training data



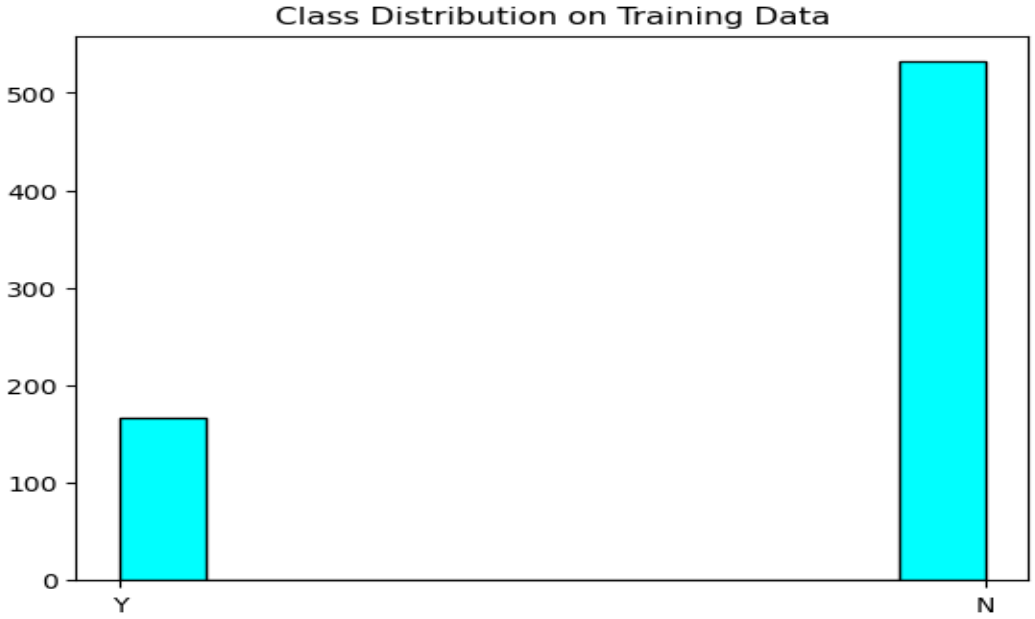
Correlation Analysis:



- Strong positive correlation between injury_claim, property_claim, vehicle_claim, and total_claim.
- Age is correlated with customer tenure.

Class Imbalance:

- The dataset is imbalanced, which may affect model performance. Resampling techniques were applied, to handle the same.

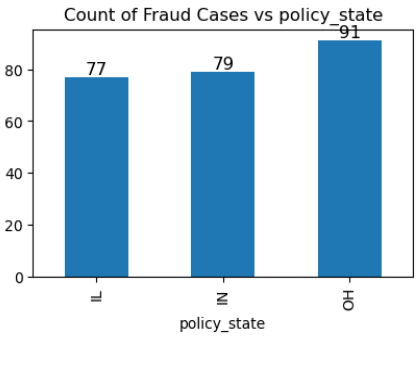


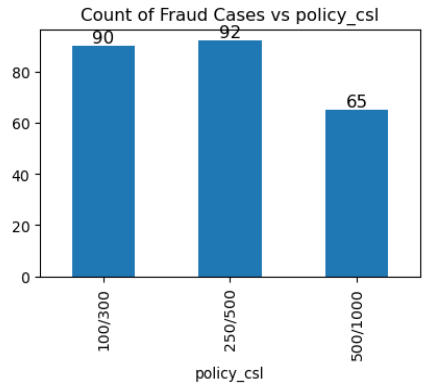
Likelihood of Fraud:

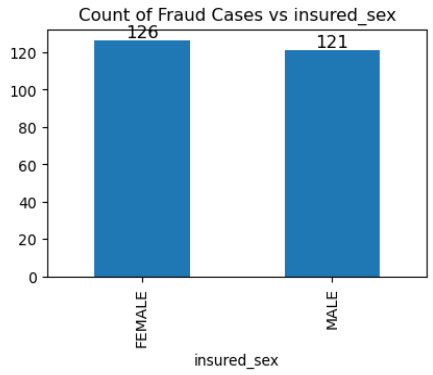
From above likelihood analysis we can see that below features are not contributing much to the solution of our problem:

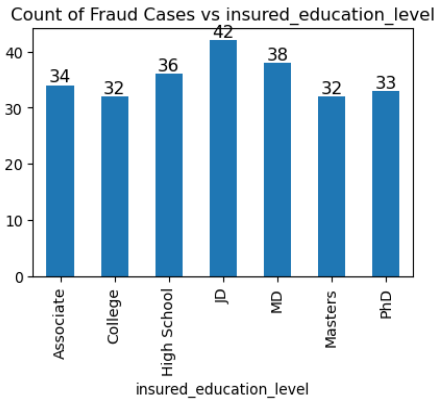
- policy_state
- insured_sex
- insured_relationship
- incident_city
- police_report_available
- insured_education_level
- policy_csl
- property_damage

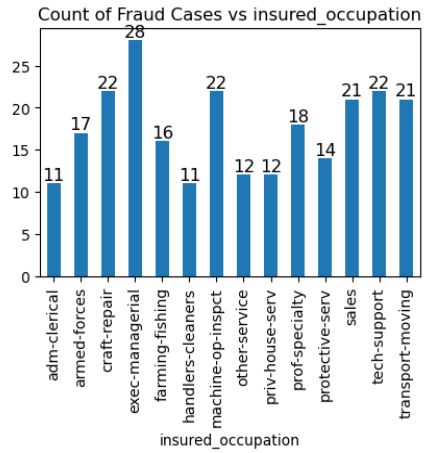
Bivariate analysis on Categorical features and Feature Engineering: We have plotted the count of the fraudulent claims as per the levels of categorical features, and upon plotting the bar chart we saw that there are many categorical features for which we can combine the levels and create new features to make our predictive models more interpretable.

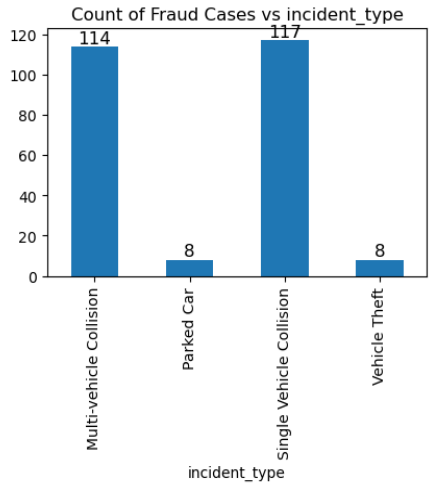


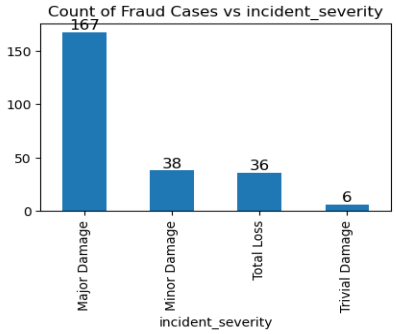


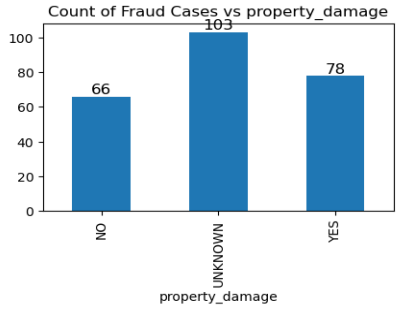


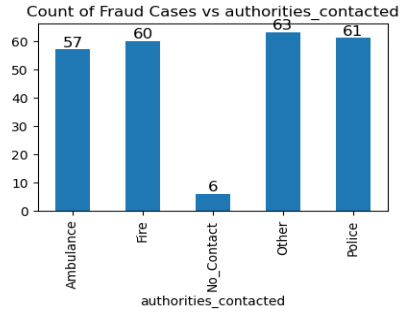


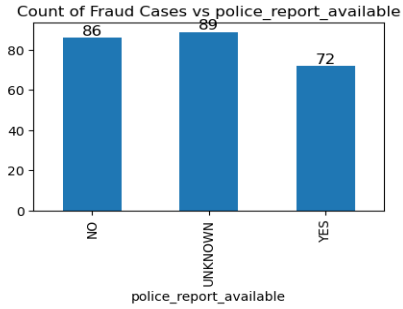


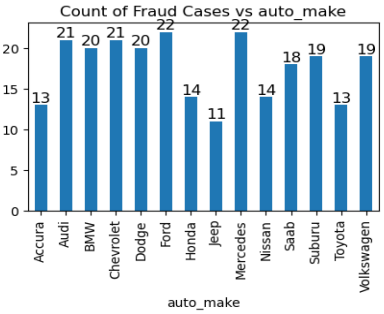


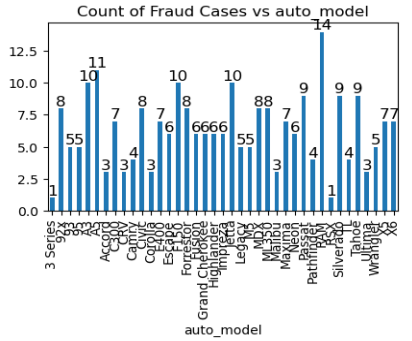












6. Feature Engineering

Key categorical variables were transformed to improve predictive power:

Education Level → `educational_group`
Occupation → `type_of_job`
Auto Make → `auto_make_country`
Hobbies → `hobby_grouped`
Auto Model → `auto_model_class`

These transformations helped reduce dimensionality and enhance interpretability.

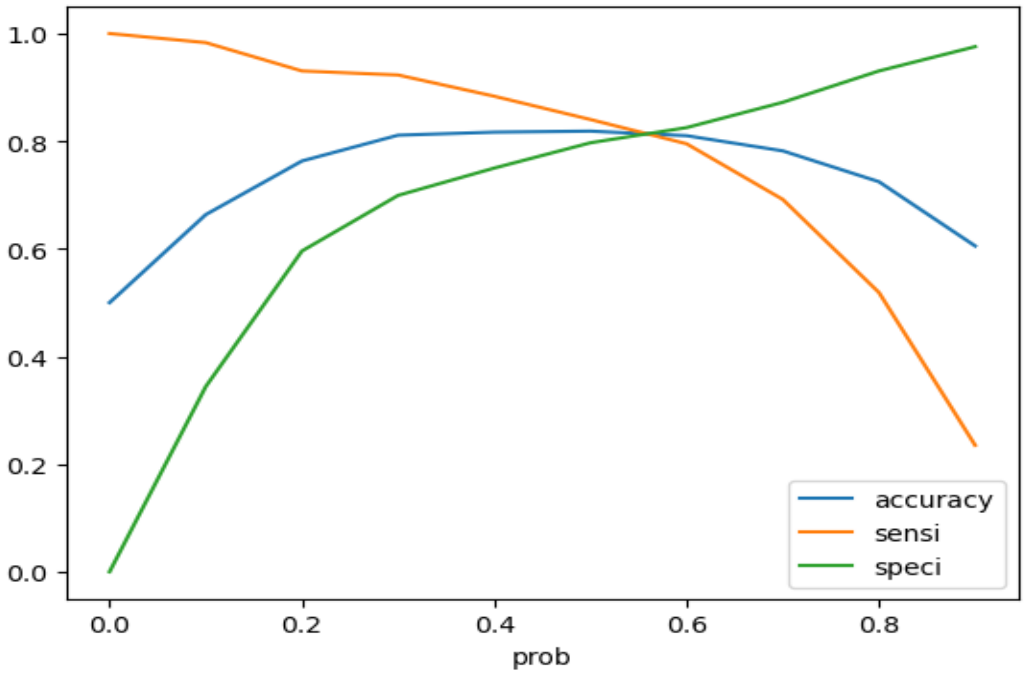
7. Model Development

Logistic regression Model:

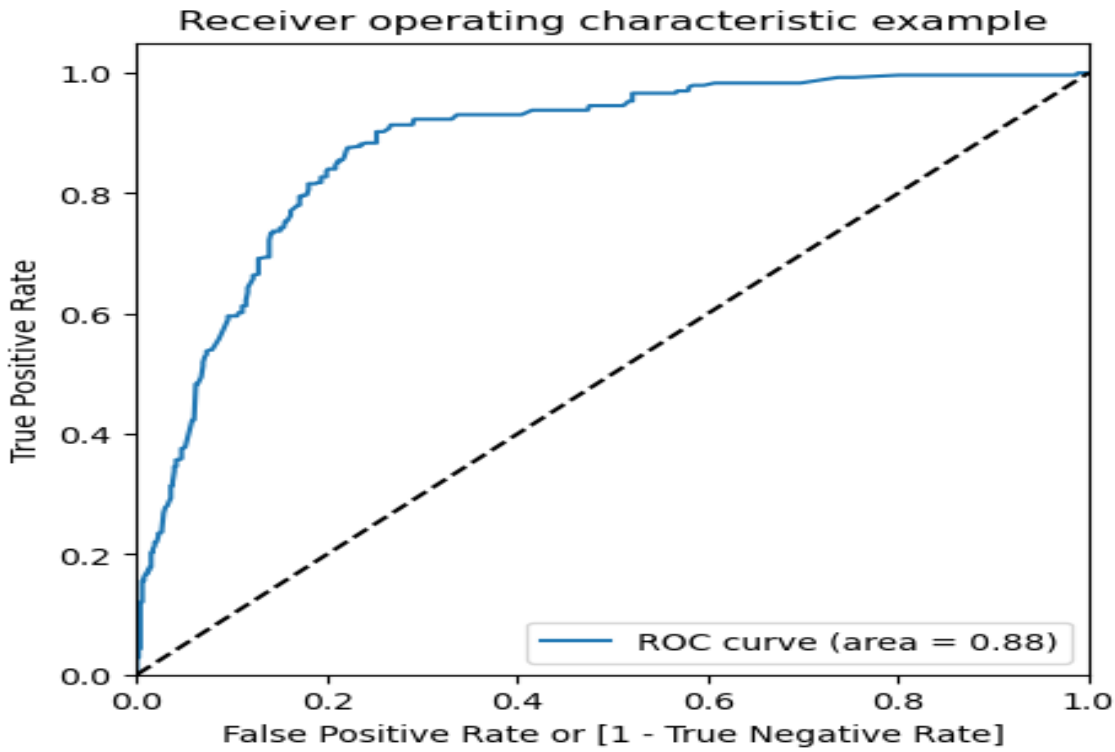
We have given more importance to specificity as we don't want to mark a genuine claim as fraudulent claim and 11 features were considered to build the LR Model using RFECV.

Cut-off value:

Based on the accuracy, sensitivity, and specificity curve we come at the cut-off value of 0.58:



ROC Curve: Using ROC Curve to check the variation covered by the LR Model:



Final LR Model Features and Evaluations:

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4617	0.246	1.875	0.061	-0.021	0.944
umbrella_limit	2.2355	0.355	6.291	0.000	1.539	2.932
incident_state_WV	-0.5973	0.218	-2.745	0.006	-1.024	-0.171
incident_type_Single Vehicle Collision	0.3648	0.171	2.130	0.033	0.029	0.700
incident_severity_Minor or Trivial Damage	-3.6331	0.232	-15.678	0.000	-4.087	-3.179
incident_severity_Total Loss	-2.9827	0.224	-13.332	0.000	-3.421	-2.544
authorities_contacted_Fire	-0.6274	0.211	-2.979	0.003	-1.040	-0.215
property_damage_UNKNOWN	0.7450	0.213	3.505	0.000	0.328	1.162
property_damage_YES	0.6793	0.224	3.033	0.002	0.240	1.118
auto_make_country_German	0.8412	0.182	4.624	0.000	0.485	1.198
hobby_grouped_Fitness & Competitive Sports	0.8320	0.224	3.713	0.000	0.393	1.271
hobby_grouped_Indoor Games	2.1959	0.237	9.284	0.000	1.732	2.660

Performance on Train Data:

- Accuracy = 0.817
- Sensitivity = 0.81
- Specificity = 0.81
- Precision = 0.81
- Recall = 0.81
- F1 Score = 0.81

- **Performance on Test Data:**
- Accuracy = 0.76
- Sensitivity = 0.66
- Specificity = 0.80
- Precision = 0.55
- Recall = 0.66
- F1 Score = 0.60

Random Forest Model (RF):

Tuned using GridSearchCV and selected features with importance > 0.01.

Below are the hyperparameters used:

```
RandomForestClassifier(max_depth=20, max_features=10, min_samples_leaf=5,
                       n_jobs=-1, random_state=42)
```

RF Model Features:

Based on the above Hyperparameters and a feature importance greater than 0.01 below 20 features were used to build the RF Model:

Varname	Imp
incident_severity_Minor or Trivial Damage	0.149470
incident_severity_Total Loss	0.092041
vehicle_claim	0.060857
property_claim	0.060618
total_claim_amount	0.051316
policy_annual_premium	0.050556
injury_claim	0.050358
days_to_incident	0.044470
months_as_customer	0.034748
incident_hour_of_the_day	0.030992
umbrella_limit	0.028966
age	0.027947
hobby_grouped_Indoor Games	0.027755
capital-loss	0.022079
capital-gains	0.021467
witnesses	0.014804
bodily_injuries	0.011710
policy_deductable	0.011586
police_report_available_UNKNOWN	0.010195
property_damage_UNKNOWN	0.010171

RF Model Performance:

Performance on Training Data:

- Accuracy = 0.98
- Sensitivity = 0.98
- Specificity = 0.98
- Precision = 0.98
- Recall = 0.98
- F1 Score = 0.98

Performance on test data:

- Accuracy = 0.75
- Sensitivity = 0.55
- Specificity = 0.83
- Precision = 0.54
- Recall = 0.55
- F1 Score = 0.54

We can see from the model performance on test and train data the RF model tends to overfit and not performing up to the mark on unseen data on comparing with training data.

8. Model Insights and Final Recommendations

Based on the evaluation of model performance, the **Logistic Regression (LR) model** is currently considered more suitable for predicting fraudulent claims. This preference is influenced by two key factors:

1. The relatively small sample size, which tends to favor simpler, interpretable models like Logistic Regression over more complex models such as Random Forest (RF).
2. Evidence of **overfitting** in the RF model during evaluation, reducing its reliability at this stage.

Data Considerations:

- **Data quality** was generally acceptable; however, several features had missing values. It is recommended that efforts be made to ensure more consistent and complete data recording in future.
- Features identified as significant by **both the LR and RF models** reinforce their importance in identifying fraud. These include:
    1. incident_severity_Minor or Trivial Damage
    2. incident_severity_Total Loss
    3. umbrella_limit
    4. property_damage_UNKNOWN
    5. hobby_grouped_Indoor Games

Model Insights and Policy Recommendations

As the LR model has been selected for its interpretability and predictive performance, the following key features and their associated insights can help inform underwriting decisions and claims review processes:

- **umbrella_limit (Coef = 2.2355)**
  A higher umbrella limit is strongly associated with increased odds of fraudulent claims. Each unit increase raises the odds of fraud by approximately **9.35 times**. While not suggesting a reduction in umbrella limits, this feature should be treated as a **risk flag** during claim assessments.
- **incident_state_WV (Coef = -0.5973)**
  Claims originating from **West Virginia** are approximately **45% less likely** to be fraudulent compared to other states. This geographic insight may inform risk-based regional analysis.
- **incident_type_Single Vehicle Collision (Coef = 0.3648)**
  Single vehicle collisions are **44% more likely** to be associated with fraud. These should warrant closer scrutiny during claim review.
- **incident_severity_Minor or Trivial Damage (Coef = -3.6331)**
  Surprisingly, claims reported with **minor or trivial damage** are significantly **less likely to be fraudulent**. This feature is a **very strong predictor** and suggests that fraud may be more common in less obvious or exaggerated claims rather than low-value ones.
- **incident_severity_Total Loss (Coef = -2.9827)**
  Total loss claims also show a strong **negative correlation with fraud**, indicating they are

more likely to be genuine. This is another **very strong predictor** supporting claim validity.

- **authorities_contacted_Fire (Coef = -0.6274)**
  When fire authorities are involved, the likelihood of fraud **decreases significantly**. This could indicate higher legitimacy due to third-party involvement.
- **property_damage_YES (Coef = 0.6793)**
  The presence of property damage **nearly doubles** the odds of fraud, indicating a potential area for deeper validation.
- **auto_make_country_German (Coef = 0.8412)**
  Claims involving **German-made vehicles** are over **twice as likely** to be fraudulent. While this may reflect certain usage patterns or repair costs, it remains a notable risk factor.
- **hobby_grouped_Indoor Games (Coef = 2.1959)**
  This feature stands out as one of the **strongest predictors** of fraud. Individuals listing hobbies such as video games, chess, or board games have significantly higher odds of filing fraudulent claims — almost **9 times more likely**. While this insight is unexpected, it is consistently supported by the data and may serve as a **behavioural risk indicator**.

Final Recommendations

1. **Prioritize Logistic Regression** for current deployment due to its transparency and performance with limited data.
2. **Accumulate more samples** to allow robust re-evaluation of both LR and RF models. This may enable more advanced ensemble methods in the future.
3. **Enhance data collection practices**, especially for fields with missing values, to improve model accuracy.
4. **Use high-risk features (e.g., umbrella_limit, hobby_grouped_Indoor Games)** as part of a **risk scoring or flagging system**, not as standalone criteria for rejection or pricing.
5. **Monitor patterns over time** to validate these findings with more extensive datasets and continually refine the fraud detection strategy.