

Fraudulent Claim Detection



Rohit Vashishth

Saurabh Singh

Sandeep Santhosh

Rakesh Kumar Sahoo

Overview of assignment – Objective & Methodology

Objective

Global Insure (an insurance company) processes thousands of claims annually, with a significant percentage being fraudulent, leading to financial losses. The current manual fraud detection process is inefficient, often detecting fraud too late. The company seeks to improve fraud detection by using data-driven insights to classify claims as fraudulent or legitimate early in the approval process, minimizing financial losses and optimizing claims handling.

Objective is to build a model that will classify claims as either fraudulent or legitimate based on historical claims' and customers' details. This model should help the organization to predict which claims are likely to be fraudulent before they are approved.

Methodology

The **CRISP-DM** (Cross-Industry Standard Process for Data Mining) methodology was followed to build the model. The steps included:

1. **Business Understanding:** Our research in the insurance domain shows that recall is crucial for identifying fraud.
2. **Data Preparation:** Processed and prepped the data for analysis.
3. **Data Cleaning:** Handled missing values and fixed datatypes
4. **Data Splitting:** Divided the data into training and test (validation) sets.
5. **Exploratory Data Analysis (EDA):** Conducted thorough analysis on the training data to understand patterns and distributions.
6. **Feature Engineering:** Created new features to improve model performance.
7. **Model Building:** Developed two models — **Logistic Regression** and **Random Forest**.
8. **Feature Selection:** RFECV for Logistic Regression and **Feature Importance** For Random Forests
9. **Model Fine-Tuning:** Determined the optimal **cutoff** point for Logistic Regression and Tuned hyperparameters using **GridSearchCV** to improve both models.
10. **Model Evaluation:** Assessed model performance using key metrics, including **Accuracy**, **Recall**, and **F1-Score**.

Overview of assignment – Results & Findings

Results & Findings

Since the goal is to detect fraudulent claims, Recall becomes the most important evaluation metric for our models. Below is the comparison of both Logistic regression and Random Forests models:

Model	Accuracy (Train Data)	Recall (Train Data)	Accuracy (Validation Data)	Recall (Validation Data)
Logistic Regression	86.6%	87.3%	84%	90.5%
Random Forests	90.4%	93.3%	77.7%	75.7%

Logistic Regression outperformed **Random Forests** in terms of **Recall**, making it the better model for detecting fraudulent claims, with a Recall of 90.5% on the validation data.

Equation: $1.5584 + 4.2153 (\text{insured_hobbies_chess}) + 3.5445 (\text{insured_hobbies_cross-fit}) - 22.7954 (\text{insured_hobbies_dancing}) - 3.4093 (\text{incident_severity_Minor Damage}) - 3.2046 (\text{incident_severity_Total Loss})$

Insights:

- insured_hobbies_chess:** Claims with the insured person having a hobby of chess are positively associated with fraud. The coefficient (4.2153) suggests that if the insured person plays chess, it increases the log-odds of a fraudulent claim
- insured_hobbies_cross-fit:** Similarly, claims with insured individuals who have a hobby of cross-fit are also positively associated with fraud. The coefficient (3.5445) indicates that these claims are more likely to be fraudulent.
- incident_severity_Minor Damage:** Claims with minor damage as the incident severity are negatively associated with fraud. The coefficient (-3.4093) suggests that minor damage claims are less likely to be fraudulent compared to more severe damage claims.
- incident_severity_Total Loss:** Claims with total loss as the incident severity are also negatively associated with fraud. The negative coefficient (-3.2046) implies that total loss incidents are less likely to be fraudulent compared to minor damage or other less severe claims.

Overview of assignment – Recommendations

Recommendations

1. **Flag Claims with Chess and Cross-fit Hobbies:** Claims filed by individuals with chess or cross-fit as hobbies should be flagged for further verification. These claims should undergo additional scrutiny during the review process to minimize the risk of fraudulent payouts.
2. **Faster Processing for Minor Damage Claims:** Since these claims are less likely to be fraudulent, streamline the approval process for minor damage claims. This will help improve turnaround times and operational efficiency.
3. **Lower Fraud Risk for Total Loss Claims:** Given that total loss claims are less likely to be fraudulent, these claims could be processed with fewer checks or flagged as lower priority for manual review. Additionally, fraud detection efforts should be focused on claims that don't involve Total Loss and Minor damage as they will have higher likelihood of fraud

Questions

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

To analyze historical claim data for detecting fraudulent patterns, we begin with descriptive, statistical, temporal, and behavioral analysis as part of exploratory data analysis (EDA). Visualizations, such as plots and graphs, play a crucial role in enhancing our understanding of the data. For example, in this case study, we developed a function to identify categorical features that contribute minimally to explaining the target variable's variance. One key finding was that 63% of applications for **major damage** claims were likely fraudulent.

Feature	Category	Target_Likelihood	Sample_Size	Coeff_of_Variation	Low_Contribution
incident_severity	Major Damage	0.63	182	1.14	False
incident_severity	Trivial Damage	0.06	69	1.14	False
incident_severity	Total Loss	0.14	192	1.14	False
incident_severity	Minor Damage	0.11	256	1.14	False

To further substantiate these insights, we built supervised classification models to identify statistically significant features associated with fraud, using the provided labels. Interestingly, our random forest model contradicted the EDA finding that major damage was a significant influencer of fraud. Instead, the model revealed that applications involving **minor damage** and **total loss** were more likely to be fraudulent. Additionally, during behavioral analysis, we discovered that applicants with hobbies such as chess and CrossFit exhibited a higher likelihood of fraud. This observation was consistent across both EDA findings and our ensemble model's results.

To enhance the model's predictive power further, we can explore **unsupervised clustering models** to uncover novel patterns that may not be evident from historical data.

Questions

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

Top 5 features with corresponding coefficients are:

1. insured_hobbies_chess (4.2153)
2. insured_hobbies_cross-fit (3.5445)
3. insured_hobbies_dancing (- 22.7954)
4. incident_severity_Minor Damage (3.4093)
5. incident_severity_Total Loss (- 3.2046)

Questions

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

Yes, we can predict the likelihood of fraud for an incoming claim using historical data and machine learning models. By training models on past claims data, we can identify significant features that distinguish fraudulent from legitimate claims. For example, in one approach, we applied Recursive Feature Elimination with Cross-Validation (RFECV) alongside a Logistic Regression model to rank and select the most impactful features based on their contribution to predictive performance. Alternatively, using a Random Forest Classifier, we leveraged the built-in `feature_importances_` metric to determine key features driving fraud detection. Once the important features are identified and the model is trained, it can output a probability score for an incoming claim, indicating the likelihood of fraud based on patterns observed in the historical data.

```
# Make final predictions on the validation data using the optimal cutoff
y_validation_pred_final['predicted'] = y_validation_pred_final.Fraud_Prob.map(lambda x: 1 if x > 0.4 else 0)
y_validation_pred_final.head()
```

] ✓ 0.0s

	Fraud	Fraud_Prob	predicted
0	0	0.042658	0
1	0	0.072058	0
2	0	0.054242	0
3	0	0.090868	0
4	0	0.072058	0

Questions

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

Based on the model output, the following insights are drawn:

- **insured_hobbies_chess:** Claims with the insured person having a hobby of chess are positively associated with fraud. The coefficient (4.2153) suggests that if the insured person plays chess, it increases the log-odds of a fraudulent claim
- **insured_hobbies_cross-fit:** Similarly, claims with insured individuals who have a hobby of cross-fit are also positively associated with fraud. The coefficient (3.5445) indicates that these claims are more likely to be fraudulent.
- **insured_hobbies_dancing:** On the other hand, having dancing as a hobby is negatively associated with fraudulent claims. The large negative coefficient (-22.7954) means that if the insured person has dancing as a hobby, the log-odds of fraud decrease significantly, making it less likely for the claim to be fraudulent.
- **incident_severity_Minor Damage:** Claims with minor damage as the incident severity are negatively associated with fraud. The coefficient (-3.4093) suggests that minor damage claims are less likely to be fraudulent compared to more severe damage claims.
- **incident_severity_Total Loss:** Claims with total loss as the incident severity are also negatively associated with fraud. The negative coefficient (-3.2046) implies that total loss incidents are less likely to be fraudulent compared to minor damage or other less severe claims.

•Improvements:

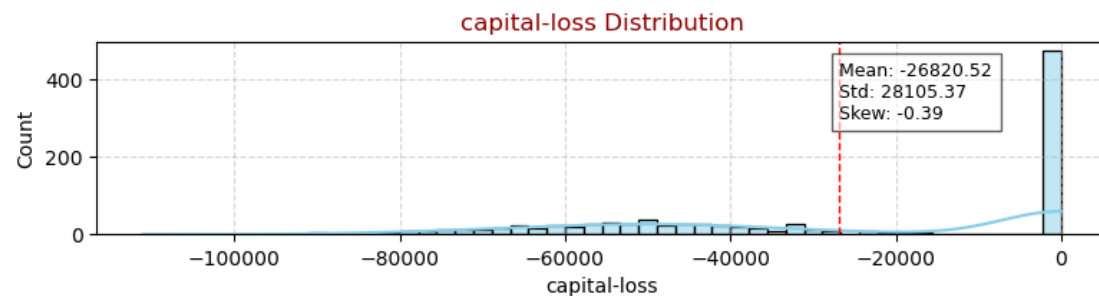
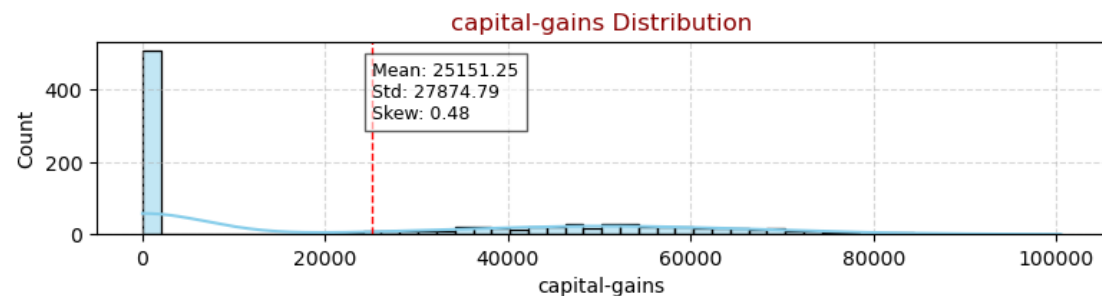
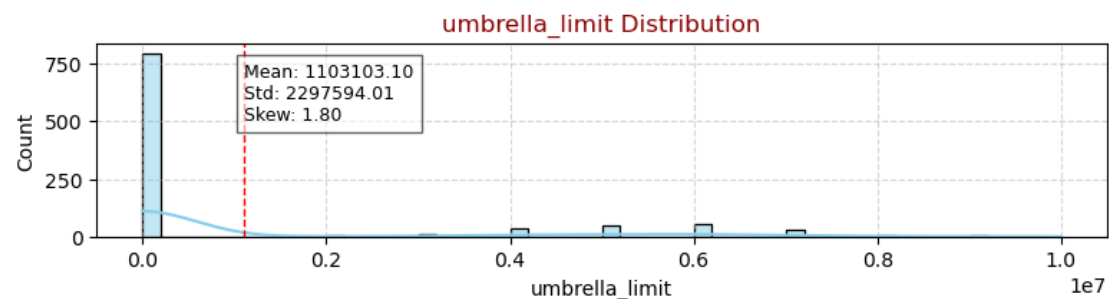
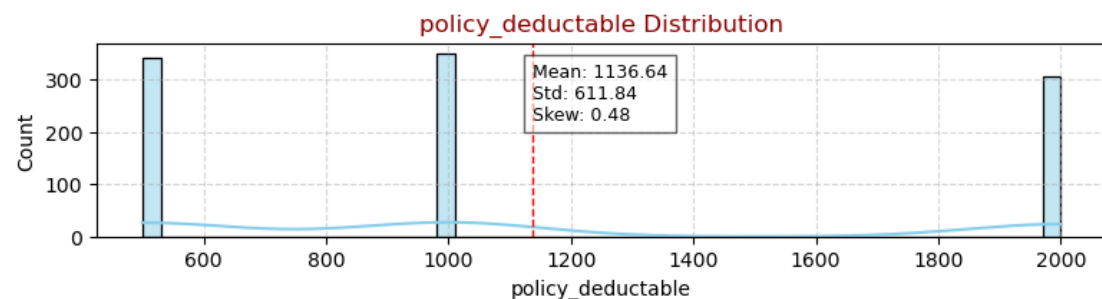
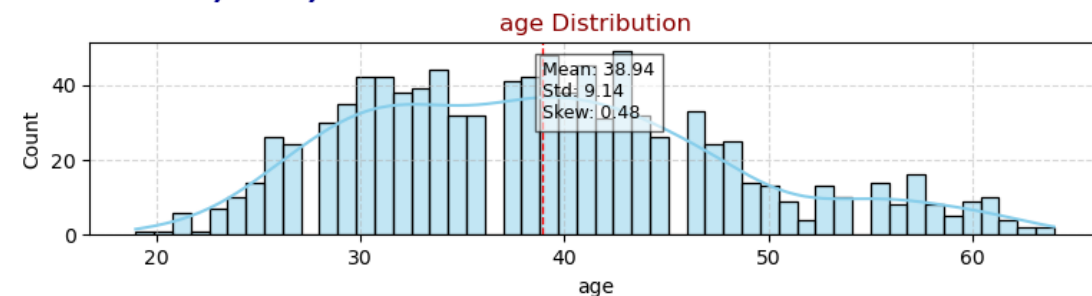
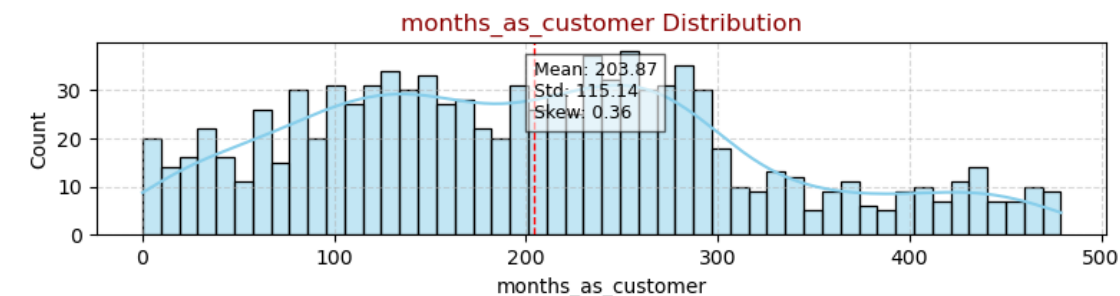
- **Flag Chess & Cross-fit Claims:** Claims with chess or cross-fit hobbies should be flagged for extra verification.
- **Faster Processing for Dancing Claims:** Claims from individuals with dancing as a hobby can be processed more quickly.
- **Streamline Minor Damage Claims:** Minor damage claims should be expedited due to lower fraud risk.
- **Lower Fraud Risk for Total Loss:** Total loss claims should undergo fewer checks, while fraud efforts should focus on claims without total loss or minor damage.

Graphs & Insights

Graphs & Insights

Univariate Analysis (Numerical columns)

Distribution of Numerical Features with Mean, Std, and Skewness

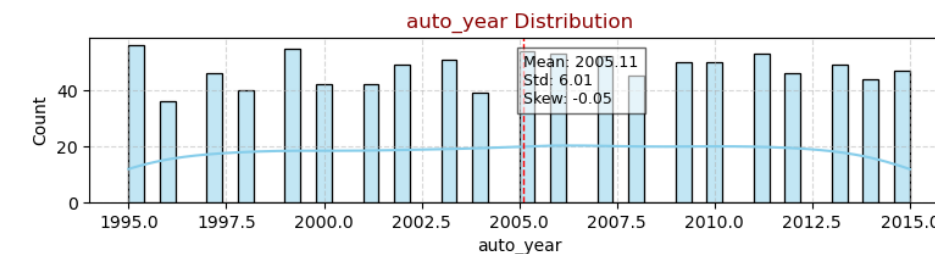
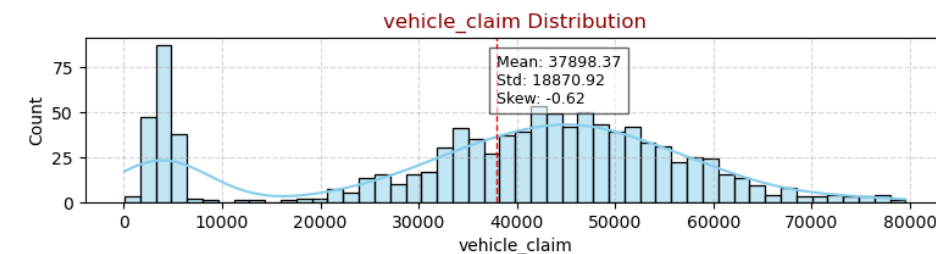
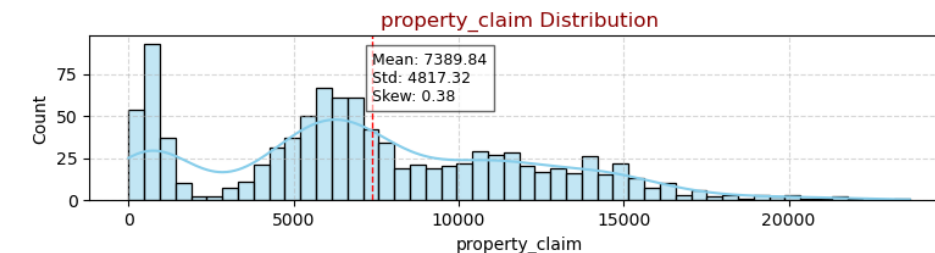
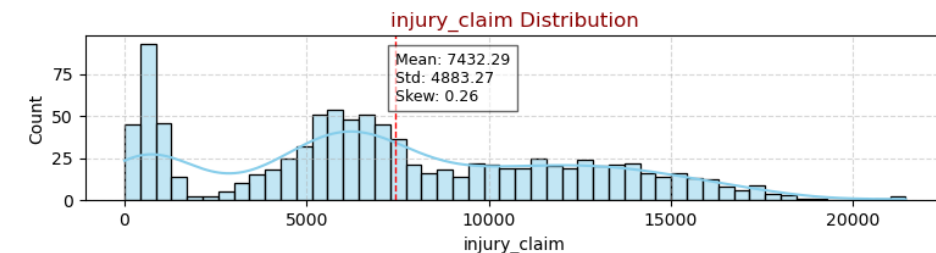
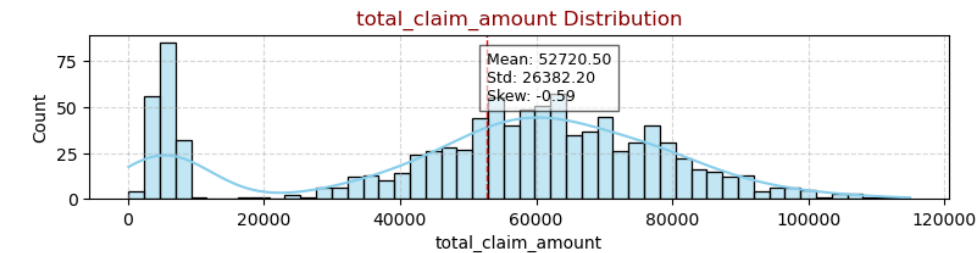
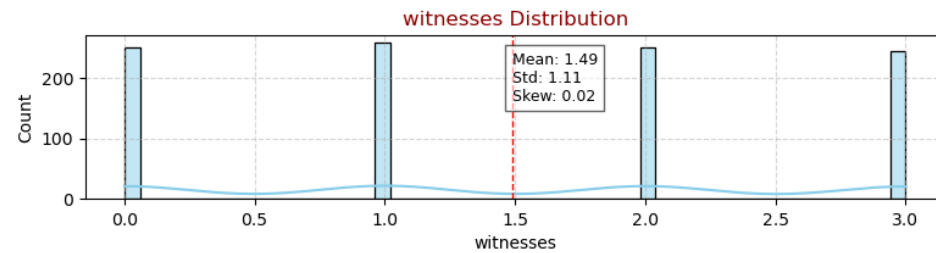
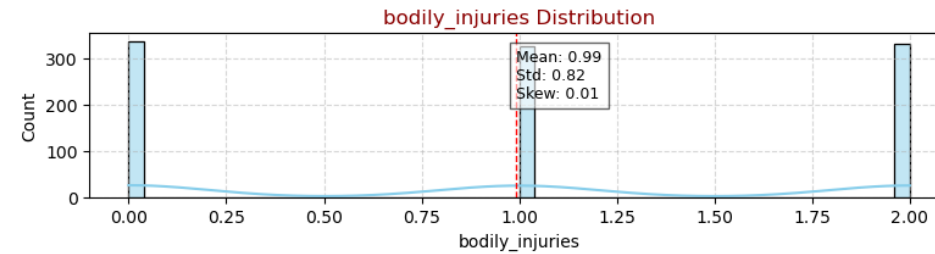
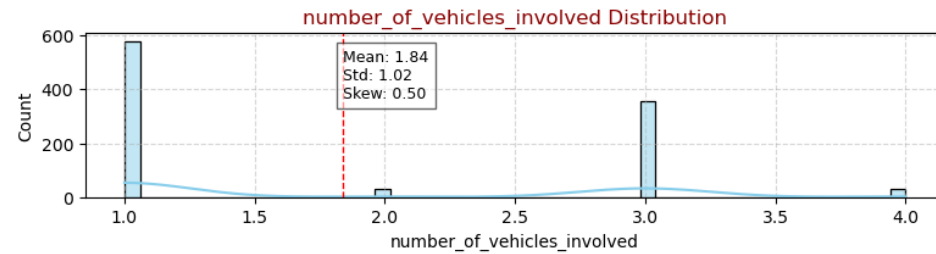


months_as_customer: Fairly uniform distribution with mild right skew (**Skew: 0.36**). Indicates a balanced customer base in terms of tenure. **policy_deductable & capital-gains:** Strongly right-skewed and discrete. Most customers lie in limited deductible brackets.

age: Near-normal distribution with slight skew (**Skew: 0.48**). Age cluster lies between 30–40. Ideal for targeted age-specific insurance products. **umbrella_limit:** Highly right-skewed (**Skew: 1.80**) with extreme values. A few customers have very large cover limits, potentially outliers. **capital-loss:** Left-skewed (**Skew: -0.39**) and mostly negative.

Graphs & Insights

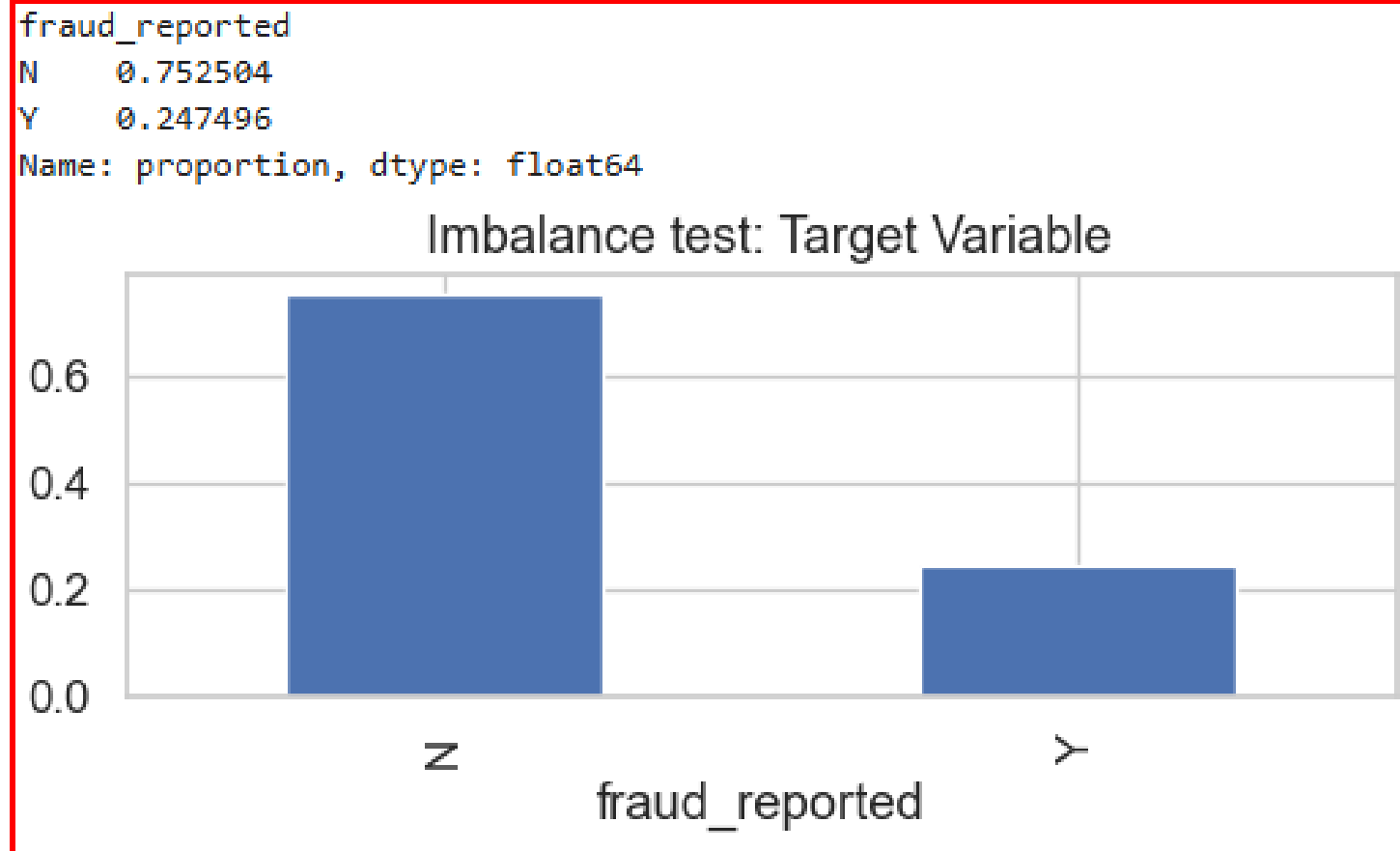
Univariate Analysis (Numerical columns)



number_of_vehicles_involved & bodily_injuries: Discrete and low-range values. Most cases involve a single vehicle and single injury. **witnesses:** Balanced distribution with low skew (**Skew: 0.02**). Majority of claims have 0–2 witnesses. **total_claim_amount:** Slight negative skew (**Skew: -0.59**). Indicates a small number of very high claims. **injury_claim, property_claim, vehicle_claim:** All show positive or slight negative skew (e.g., vehicle_claim **Skew: -0.62**). **auto_year:** Uniformly distributed from 1995 to 2015. Suggests no bias toward vehicle age. Can be directly used or categorized into "new", "mid", and "old" segments.

Graphs & Insights

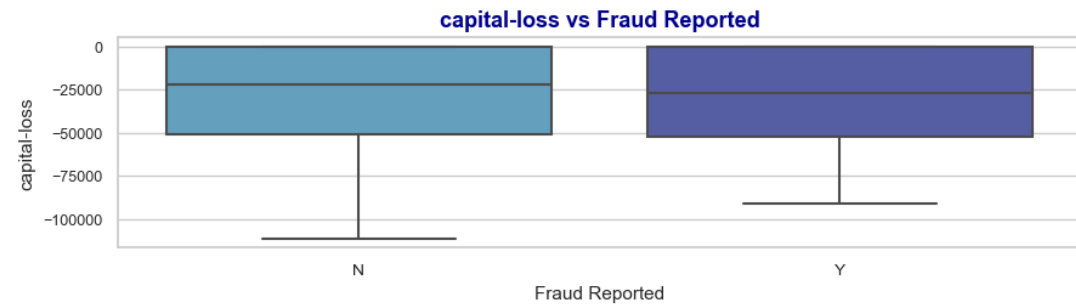
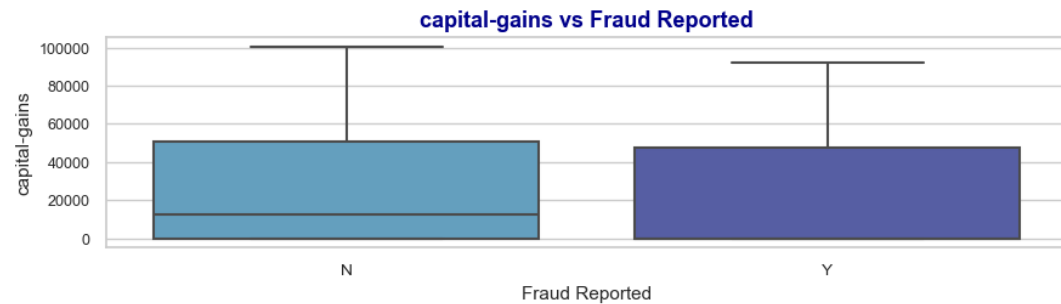
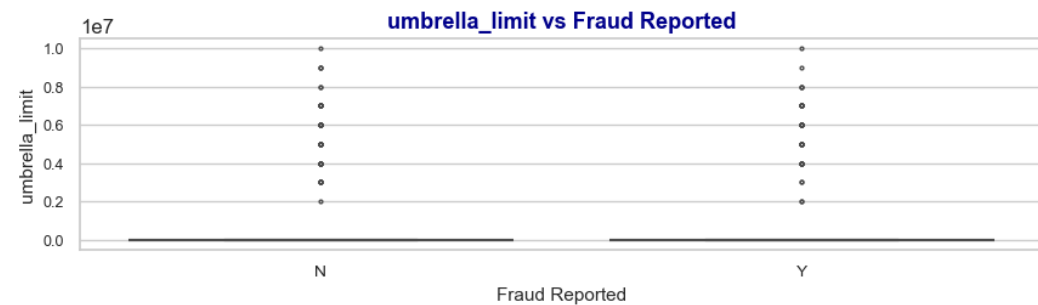
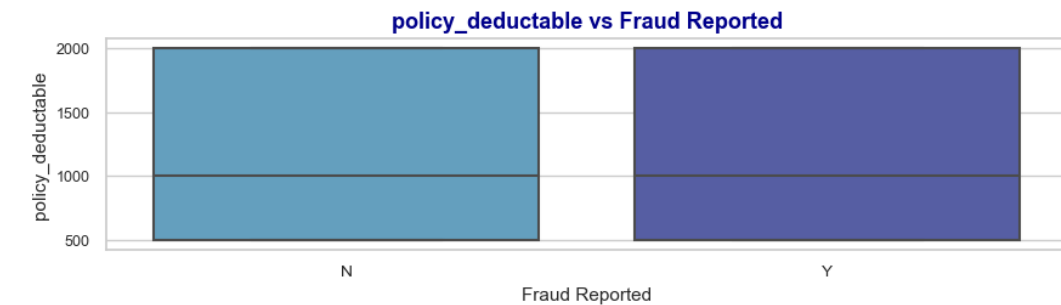
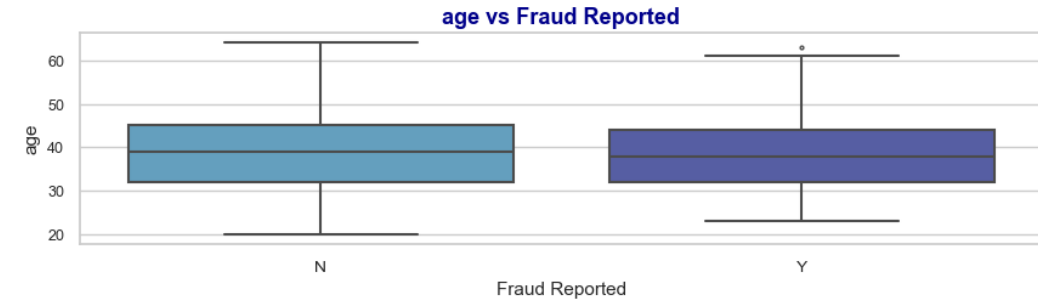
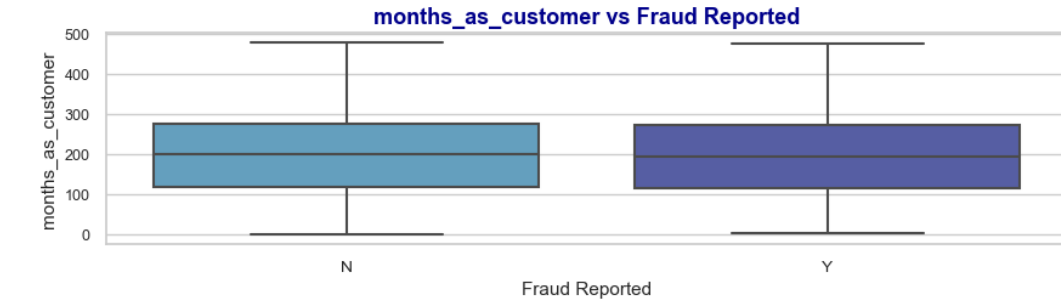
Class Imbalance



Dataset Imbalance (fraud_reported: N = 0.7527, Y = 0.2473): The significant class imbalance, with non-fraudulent claims (75.27%) outnumbering fraudulent ones (24.73%)

Graphs & Insights

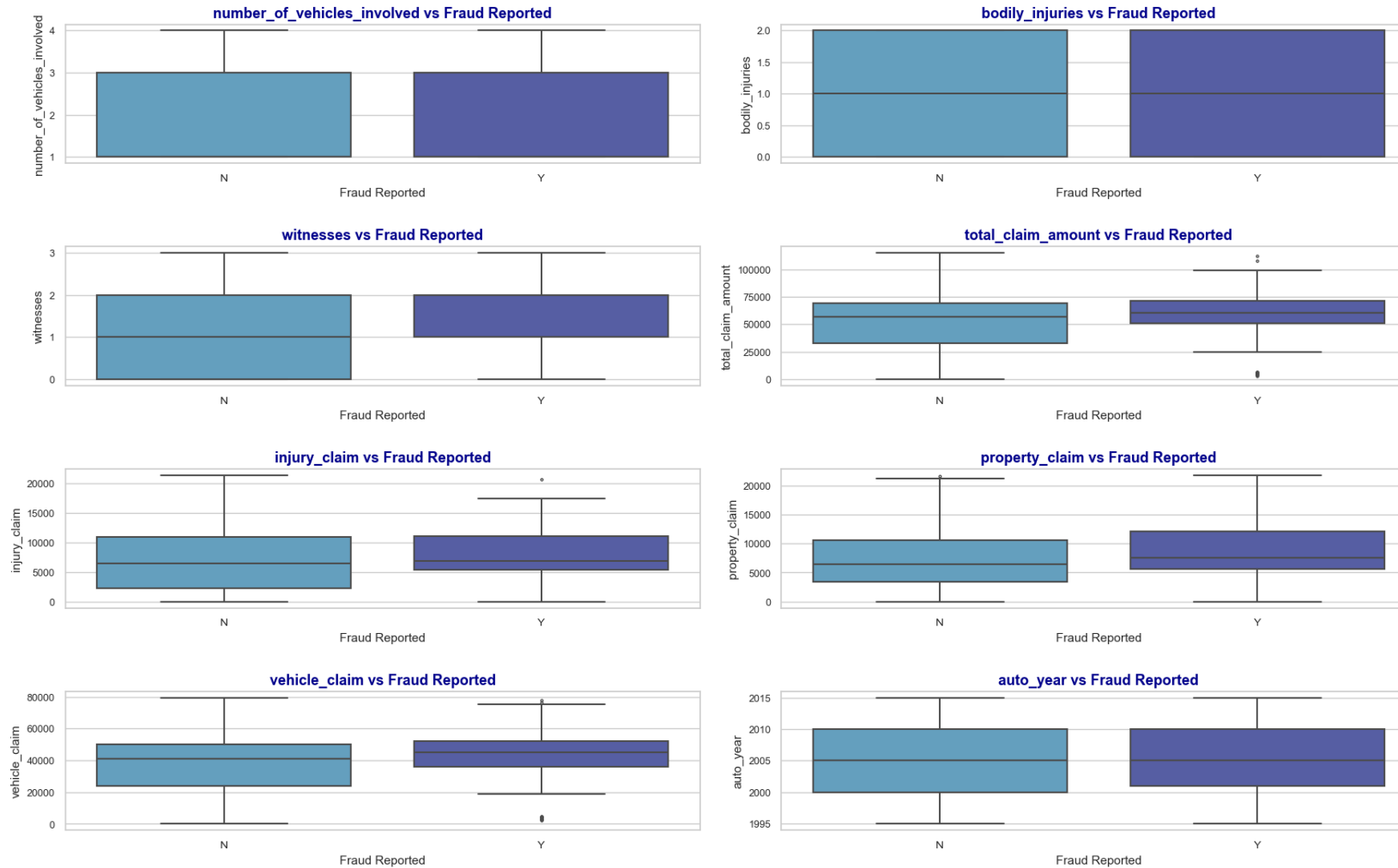
Bivariate Analysis (Categorical columns Vs Target variable)



months_as_customer: Fraud cases appear slightly skewed towards lower customer tenure. Indicates newer customers might have higher fraud likelihood. **age:** Age distributions show that younger age groups (below 30) have a slightly higher occurrence of fraud. Targeted verification strategies might help for younger profiles. **policy_deductable:** Fraudulent claims tend to concentrate in lower deductible brackets. Lower deductibles might incentivize fraudulent claims. **umbrella_limit:** No clear differences observed between fraud and non-fraud cases. Feature might not be a strong predictor of fraud. **capital-gains:** Fraud cases show slightly higher capital gains. This could be an indicator for certain fraud profiles. **capital-loss:** Similar distributions in fraud and non-fraud cases. Likely not impactful for fraud prediction.

Graphs & Insights

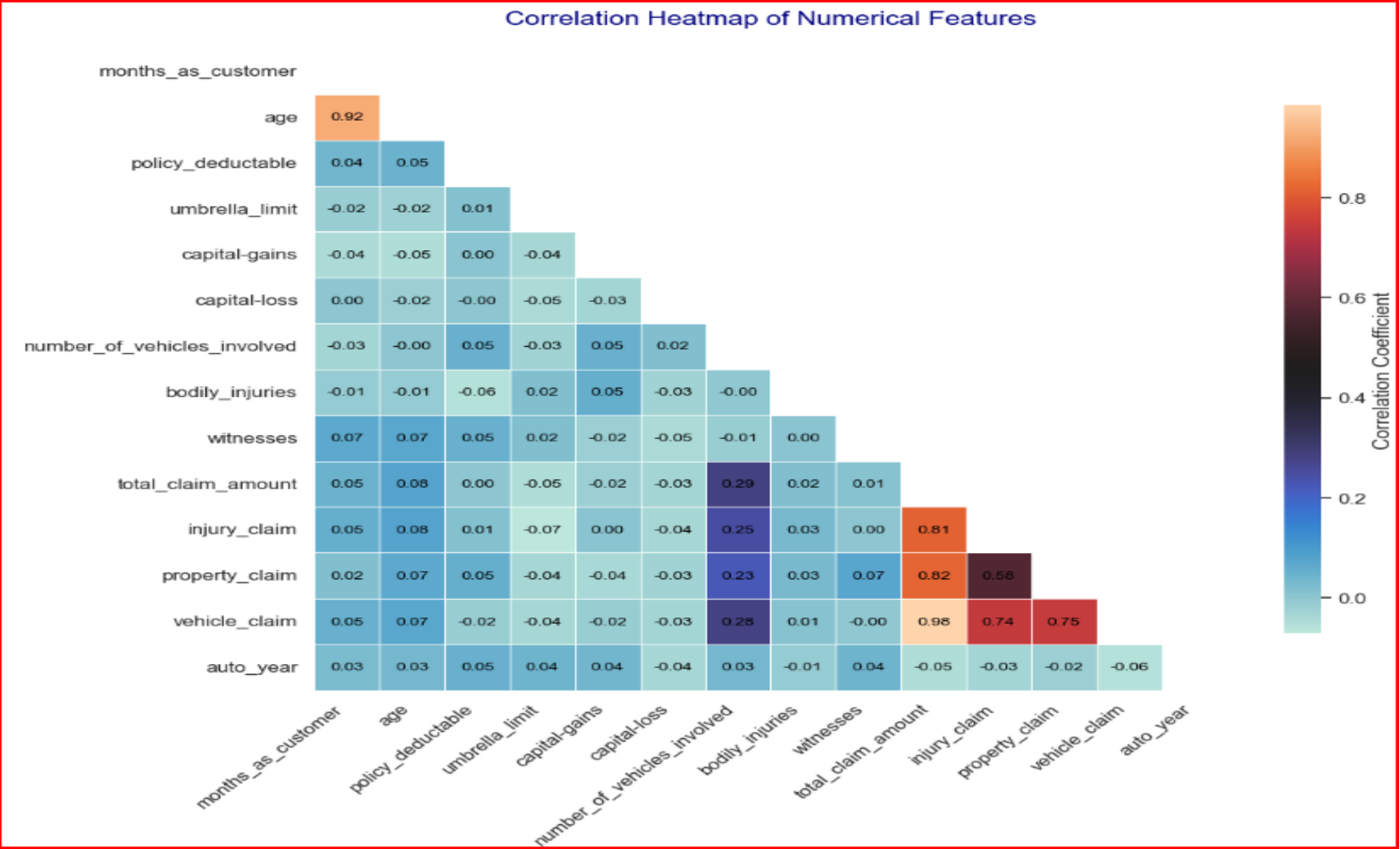
Bivariate Analysis (Categorical columns Vs Target variable)



number_of_vehicles_involved: Higher fraud likelihood with multiple vehicles involved. This feature could be significant in claim assessments. **bodily_injuries:** Fraud cases show slightly higher bodily injury counts. This may help flag suspicious claims. **witnesses:** More witnesses seem to correlate with fraud cases. This could be leveraged in fraud verification strategies. **total_claim_amount:** Higher total claim amounts are more common in fraud cases. Could be a key predictor in fraud detection. **injury_claim, property_claim, vehicle_claim:** Each component shows increased amounts in fraudulent cases. Segmentation of these claims might provide deeper insights. **auto_year:** No discernible patterns observed between fraud and non-fraud cases. Likely independent of fraud occurrences.

Graphs & Insights

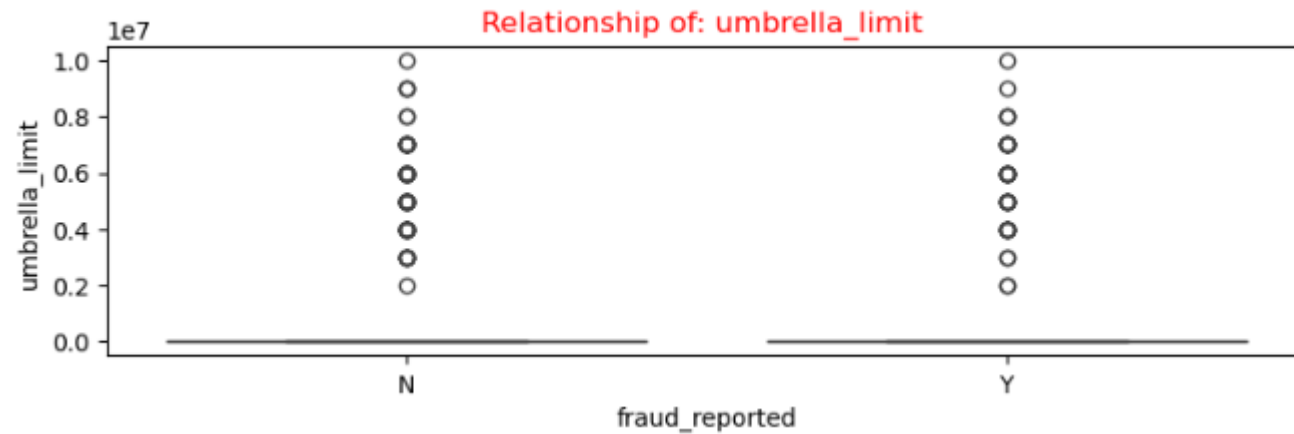
Bivariate Analysis (Identifying potential Multicollinearity)



months_as_customer, age, total_claim_amount, injury_claim, property_claim, vehicle_claim: Heatmap analysis reveals strong correlations among these features. This suggests the potential presence of multicollinearity.

Graphs & Insights

Bivariate Analysis (Features that Don't Strongly Influence the Prediction)



Umbrella_limit Vs fraud_reported: The distribution of "umbrella_limit" values appears similar for both "Y" and "N" categories. There's no noticeable difference in the central tendency, spread, or pattern that would suggest "umbrella_limit" is informative for predicting fraud. A significant portion of the data points are concentrated near the lower end of the scale, irrespective of the fraud status, with several high-value outliers in both categories. These outliers don't seem to follow any distinct pattern tied to fraud status.

Final Logistic Regression Model

Generalized Linear Model Regression Results

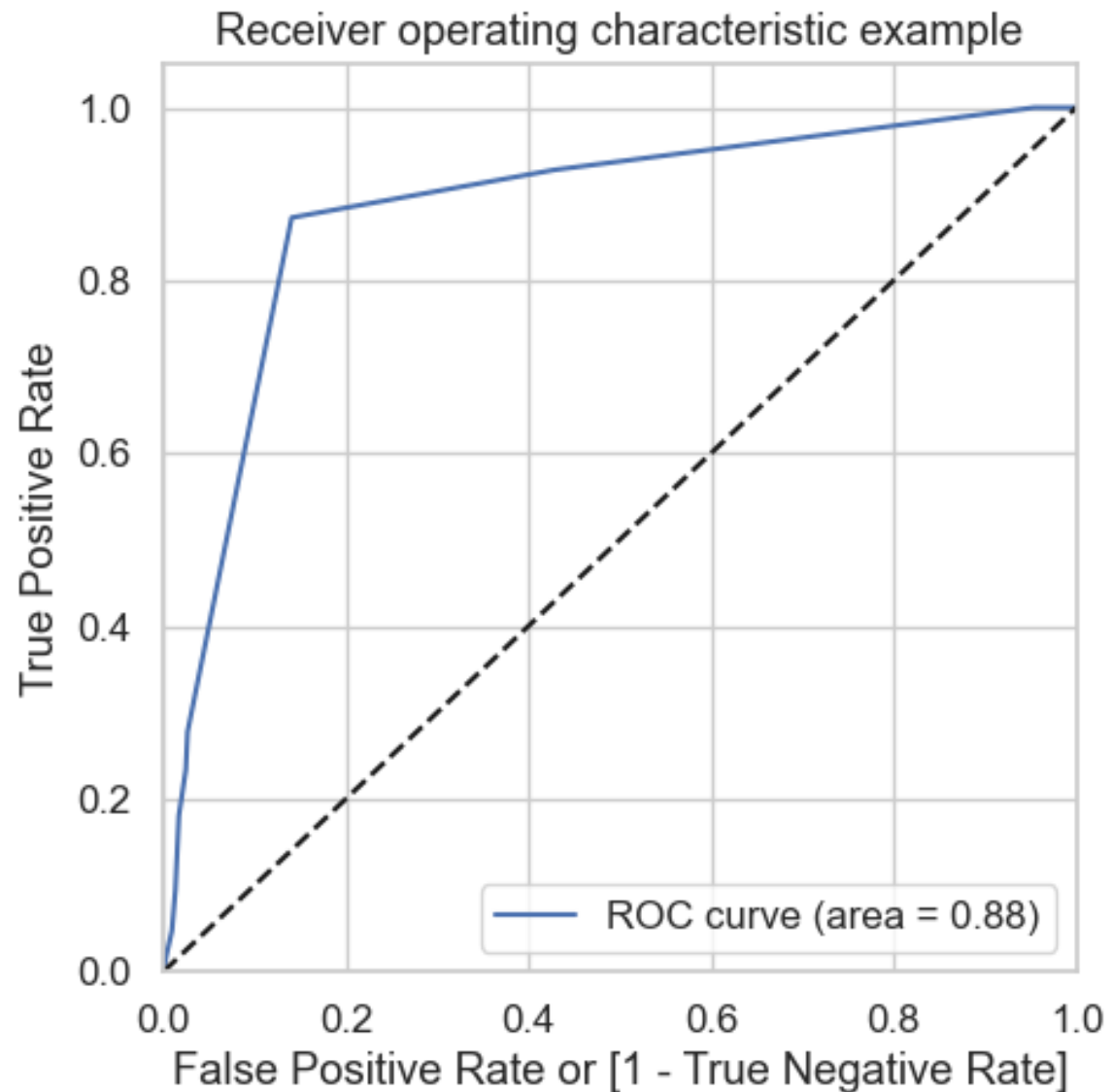
=====						
Dep. Variable:	fraud_reported	No. Observations:	1052			
Model:	GLM	Df Residuals:	1046			
Model Family:	Binomial	Df Model:	5			
Link Function:	Logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-421.41			
Date:	Tue, 08 Apr 2025	Deviance:	842.81			
Time:	14:58:28	Pearson chi2:	2.05e+03			
No. Iterations:	22	Pseudo R-squ. (CS):	0.4430			
Covariance Type:	nonrobust					
=====						
	coef	std err	z	P> z	[0.025	0.975]

const	1.5584	0.136	11.422	0.000	1.291	1.826
insured_hobbies_chess	4.2153	0.445	9.483	0.000	3.344	5.087
insured_hobbies_cross-fit	3.5445	0.403	8.800	0.000	2.755	4.334
insured_hobbies_dancing	-22.7954	1.49e+04	-0.002	0.999	-2.91e+04	2.91e+04
incident_severity_Minor Damage	-3.4093	0.209	-16.291	0.000	-3.819	-2.999
incident_severity_Total Loss	-3.2046	0.235	-13.614	0.000	-3.666	-2.743
=====						

Based on the **Generalized Linear Model (GLM)** Regression Results for the target variable "fraud_reported," here are the five most important insights derived from the coefficients and their statistical significance (p-values), which can help improve the fraud detection process.

- **incident_severity_Total Loss** (coef: -3.2046, p=0.000)
- **incident_severity_Minor Damage** (coef: -3.4093, p=0.000):
- **insured_hobbies_cross-fit** (coef: 3.5445, p=0.000)
- **insured_hobbies_dancing** (coef: -22.7954, p=0.999): Data anomaly; review and clean.
Note: we should have dropped this as the p-value is more than 5%, but kept it as it was optional.
- **insured_hobbies_chess** (coef: 4.2153, p=0.000)

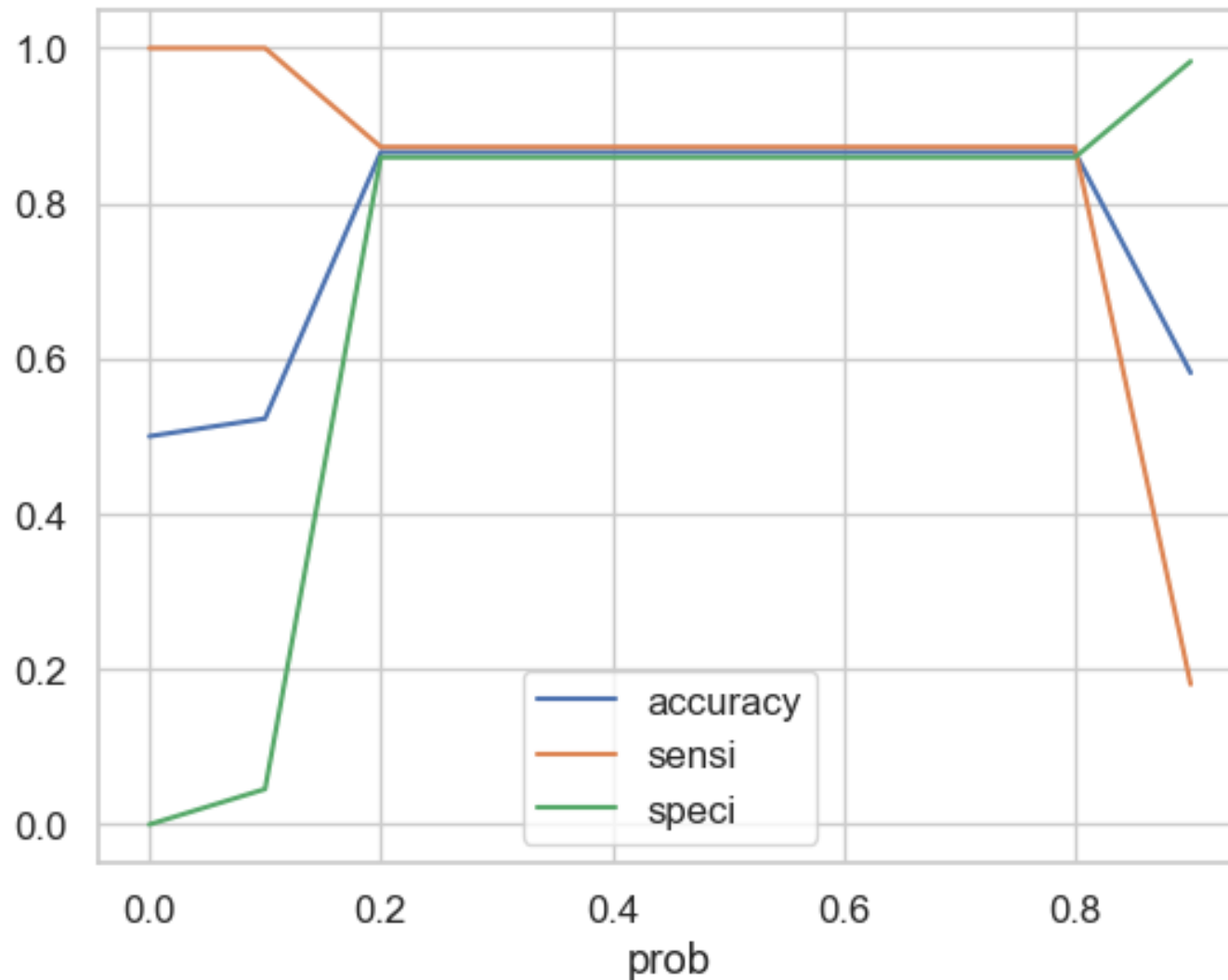
Receiver Operating Characteristic (ROC) Curve



The Receiver Operating Characteristic (ROC) curve illustrates the model's ability to distinguish between fraudulent and non-fraudulent claims, with an Area Under the Curve (AUC) of 0.88, indicating strong predictive performance

Finding Optimal Cut Off value- accuracy, sensitivity, and specificity

Finding Optimal Cut Off value- accuracy, sensitivity, and specificity



The plot displays the relationship between probability thresholds and key performance metrics—accuracy (blue), sensitivity (orange), and specificity (green)—for the fraud detection model. This helps identify the optimal threshold to balance fraud detection logistic regression model.

We choose **0.4** as our cut-off value for our logistic regression model as any value between 0.2 to 0.8 will provide the same probabilities

Logistic Regression

Evaluation Metrics – Train and Test datasets

Metric	Train dataset	Validation dataset
Accuracy	86.6%	84%
Sensitivity	87.3%	90.5%
Specificity	85.9%	81.9%
Precision	86.1%	62%
Recall	87.3%	90.5%
F1-Score	86.7%	73.6%

Random Forests

Evaluation Metrics – Train and Test datasets

Metric	Train dataset	Validation dataset
Accuracy	90.4%	77.7%
Sensitivity	93.3%	75.7%
Specificity	87.5%	78.3%
Precision	88.2%	53.3%
Recall	93.3%	75.7%
F1-Score	90.7%	62.6%