Fraudulent Claim Detection

Submitted By-

Yoshita Dhanwaria

Introduction

Fraudulent insurance claims pose a significant challenge for insurers, resulting in considerable financial losses and reduced operational efficiency. As claim volumes continue to grow, traditional manual methods of fraud detection have become inadequate. Adopting data-driven approaches enables insurers to detect and prevent fraud more accurately and efficiently

Problem Statement & Business Objective

Problem Statement
Global Insure processes thousands of claims annually, many of which are fraudulent. Manual fraud detection is slow and often too late, resulting in financial loss. The company needs an efficient way to detect fraud earlier in the process.

Business Objective
Develop a predictive model that uses historical claim and customer data to classify claims as fraudulent or legitimate, enabling faster and more accurate fraud detection.

Data Overview

Source: insurance_claims.csv, containing policy details, incident information, customer demographics, claim amounts, and a binary target fraud_reported (Y/N).

Training–Validation Split:
- Training set: 699 × 0.75 ≈ 525 samples
- Validation set: 699 × 0.25 ≈174 samples

Class Balance:
- Fraudulent: ~25%
- Non-fraudulent: ~75%
- Imbalance ratio ≈3:1 (majority : minority

Data preparation

Missing Value imputation

- Identified and dropped columns with excessive missingness
- Imputed or removed rows for remaining nulls as appropriate.

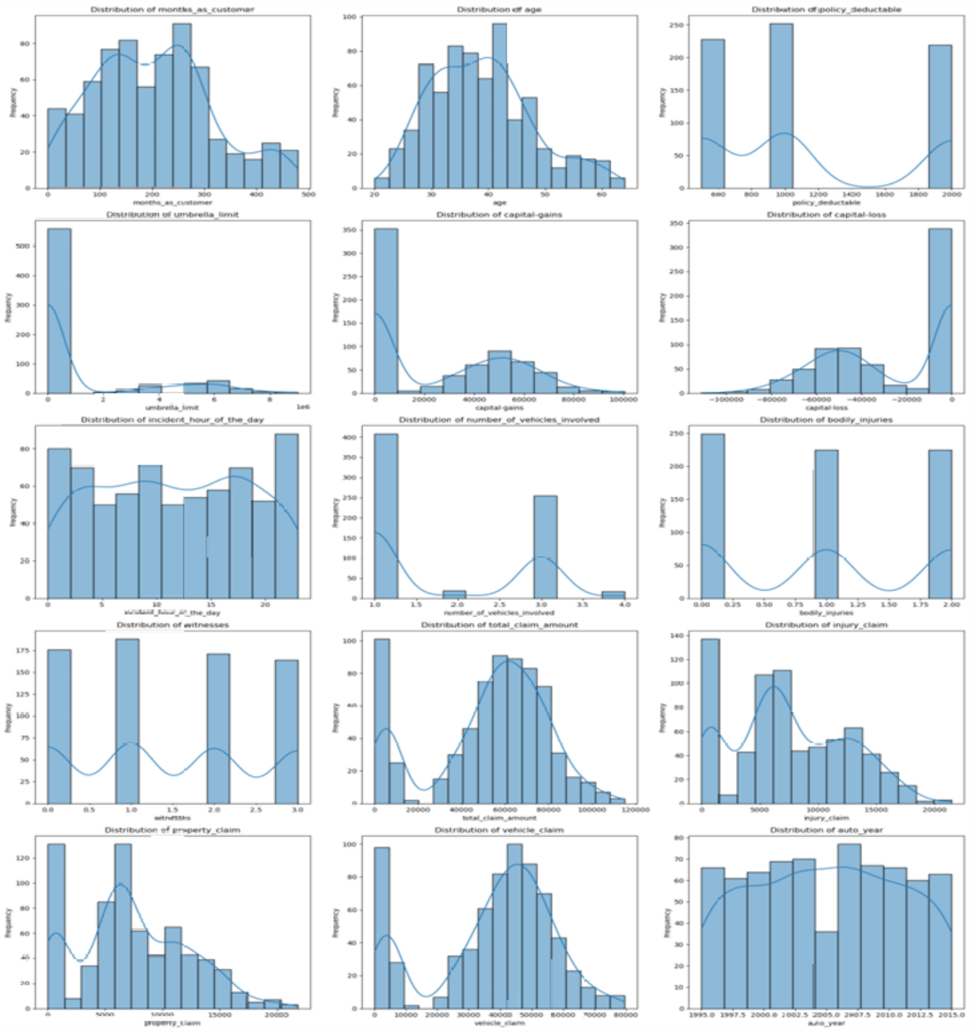
Redundant & Illogical Entries

- Removed duplicate records.
- Dropped features with constant or near-constant values.
- Ensured numeric fields (e.g., policy durations, claim amounts) were non-negative

Data Types

- Converted date fields to datetime objects.
- Cast categorical columns to category dtype.

EDA - Univariate Analysis



Observations from histogram plots:

months_as_customer:
- Mean: 202.57, Median: 199.00
- Skewness: 0.37
- Distribution appears approximately symmetric

age:
- Mean: 38.85, Median: 38.00
- Skewness: 0.51
- Distribution is positively skewed (right-tailed)

policy_deductable:
- Mean: 1150.21, Median: 1000.00
- Skewness: 0.45
- Distribution appears approximately symmetric

umbrella_limit:
- Mean: 1077253.22, Median: 0.00
- Skewness: 1.79
- Distribution is positively skewed (right-tailed)

capital-gains:
- Mean: 25506.01, Median: 0.00
- Skewness: 0.45
- Distribution appears approximately symmetric

capital-loss:
- Mean: -26458.37, Median: -20800.00
- Skewness: -0.41
- Distribution appears approximately symmetric

number_of_vehicles_involved:
- Mean: 1.83, Median: 1.00
- Skewness: 0.49
- Distribution appears approximately symmetric

bodily_injuries:
- Mean: 0.97, Median: 1.00
- Skewness: 0.06
- Distribution appears approximately symmetric

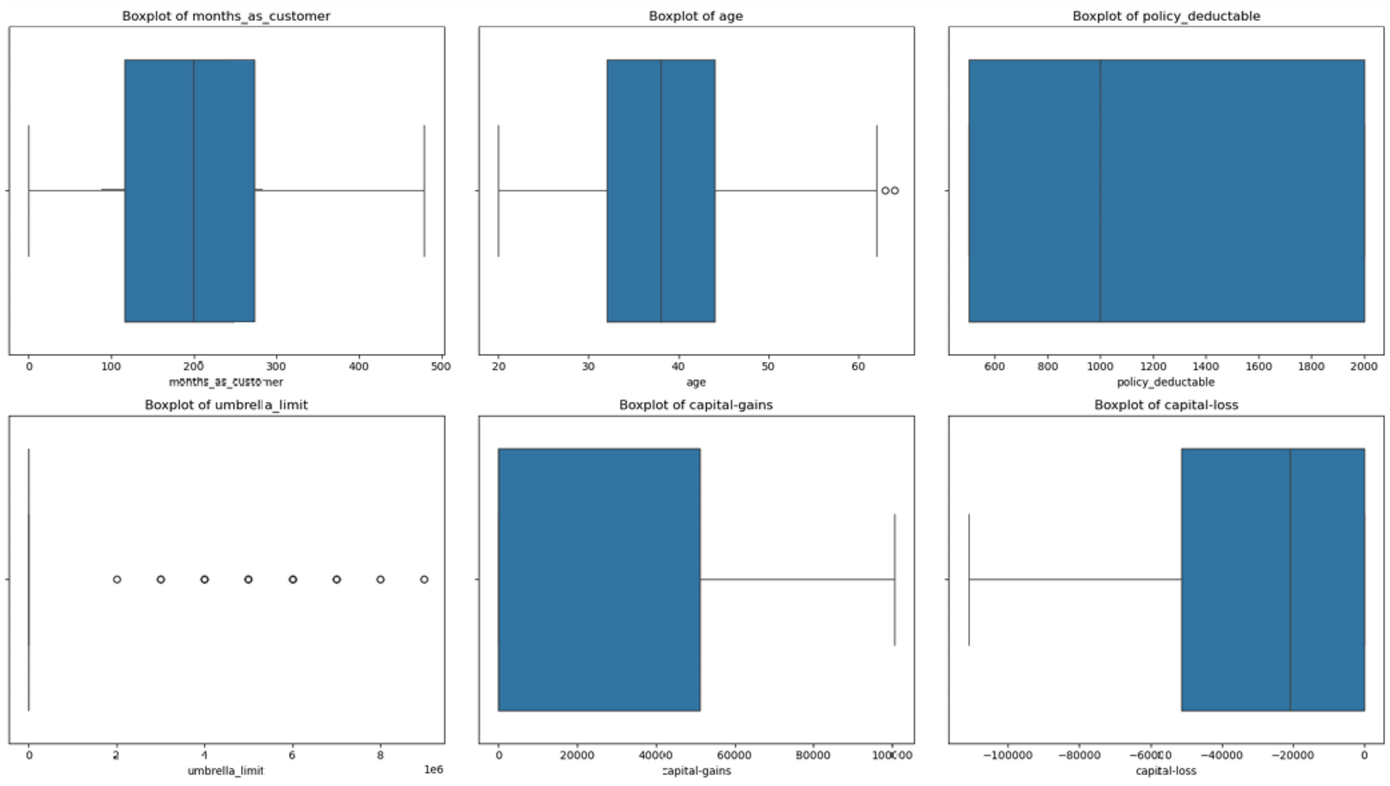
witnesses:
- Mean: 1.46, Median: 1.00
- Skewness: 0.06
- Distribution appears approximately symmetric

total_claim_amount:
- Mean: 52923.61, Median: 58300.00
- Skewness: -0.57
- Distribution is negatively skewed (left-tailed)

injury_claim:
- Mean: 7508.73, Median: 6780.00
- Skewness: 0.27
- Distribution appears approximately symmetric

property_claim:
- Mean: 7399.20, Median: 6780.00
- Skewness: 0.33
- Distribution appears approximately symmetric

EDA - Univariate Analysis



Observations from boxplot plots:

months_as_customer:
- Number of outliers: 0
- Outlier range: (-120.25, 509.75)

age:
- Number of outliers: 4
- Outlier range: (14.00, 62.00)

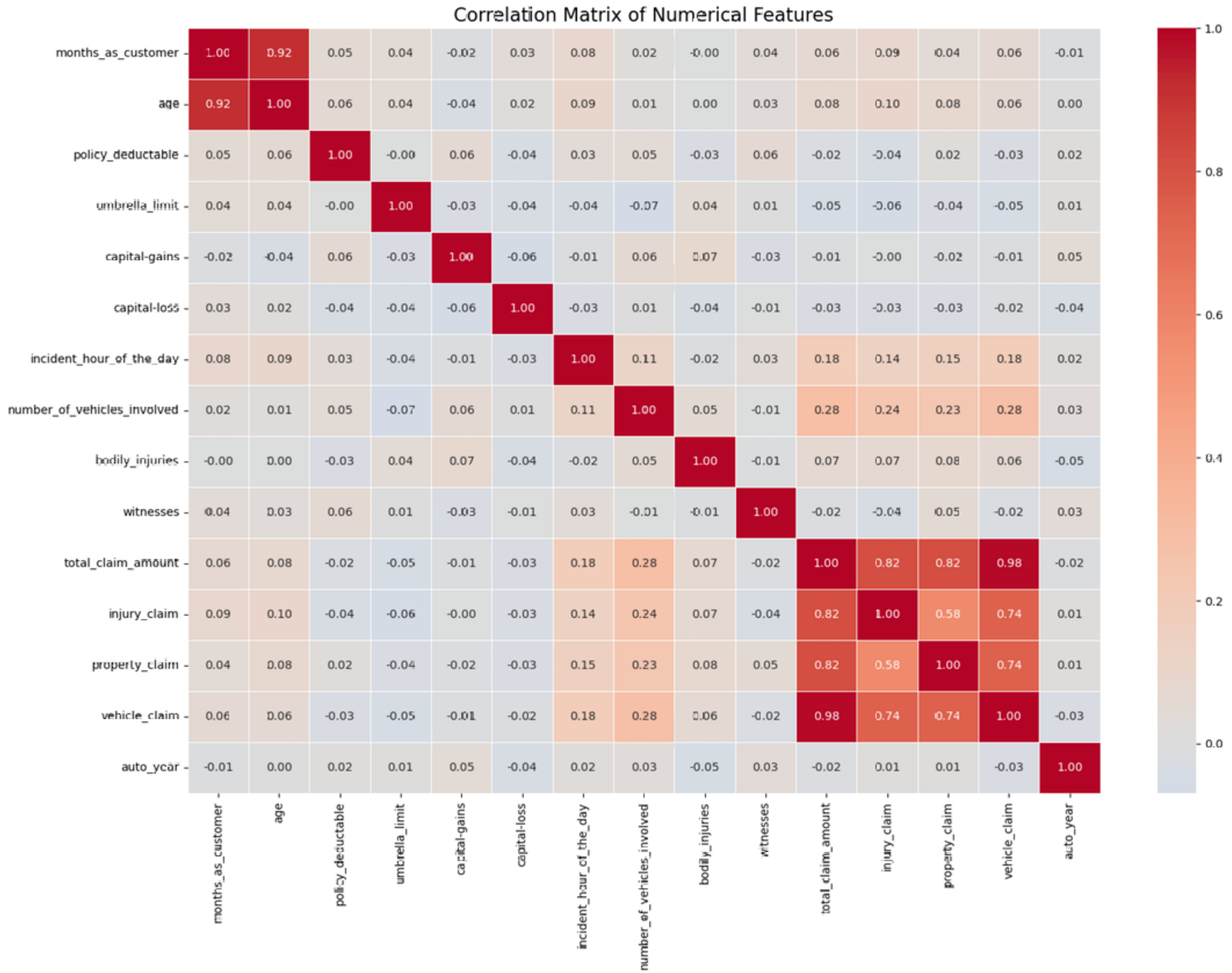
policy_deductable:
- Number of outliers: 0
- Outlier range: (-1750.00, 4250.00)

umbrella_limit:
- Number of outliers: 140
- Outlier range: (0.00, 0.00)

capital-gains:
- Number of outliers: 0
- Outlier range: (-76650.00, 127750.00)

capital-loss:
- Number of outliers: 0
- Outlier range: (-128125.00, 76875.00)

Correlation Matrix



Highly correlated feature pairs (|correlation| > 0.7):

age & months_as_customer: 0.920

injury_claim & total_claim_amount: 0.818

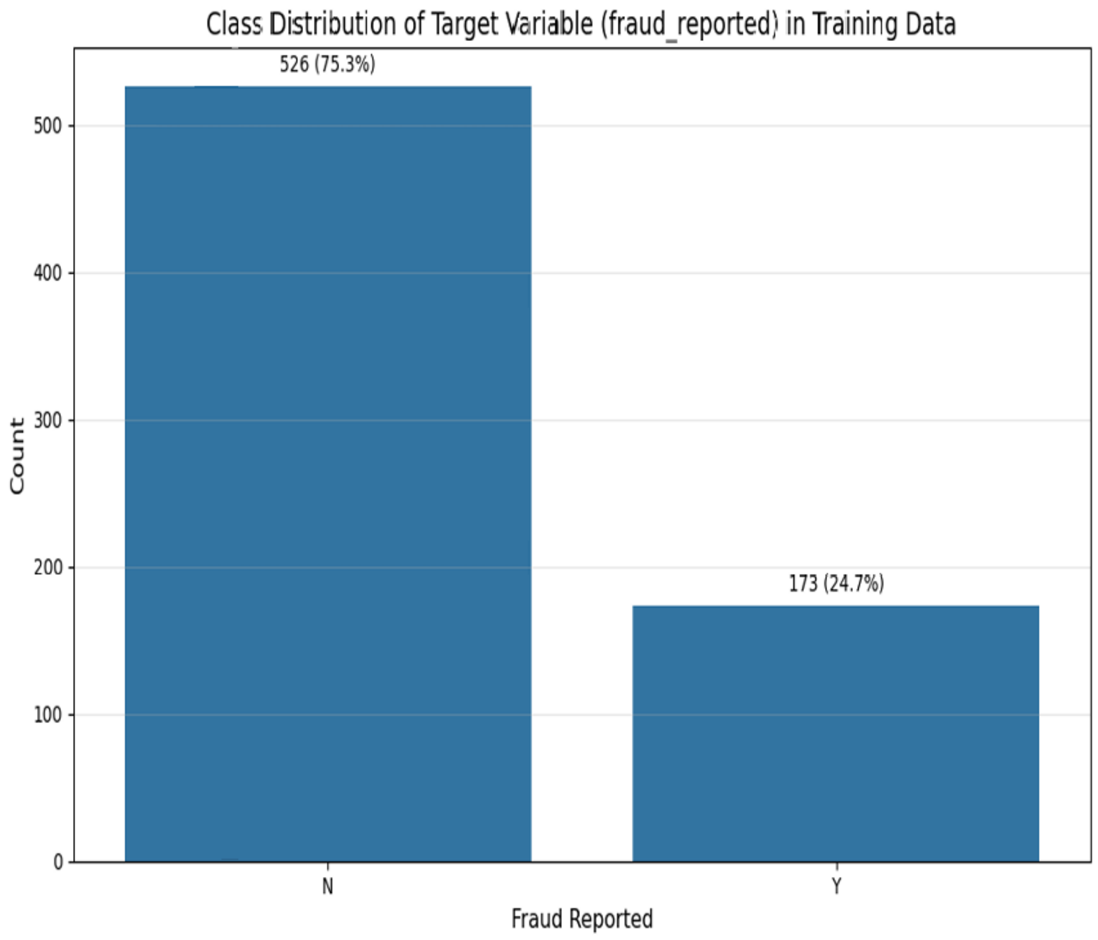
property_claim & total_claim_amount: 0.815

vehicle_claim & total_claim_amount: 0.984

vehicle_claim & injury_claim: 0.743

vehicle_claim & property_claim: 0.742

Class Imbalance analysis



Class imbalance analysis:

Majority class (N): 526 samples (75.25%)

Minority class (Y): 173 samples (24.75%)

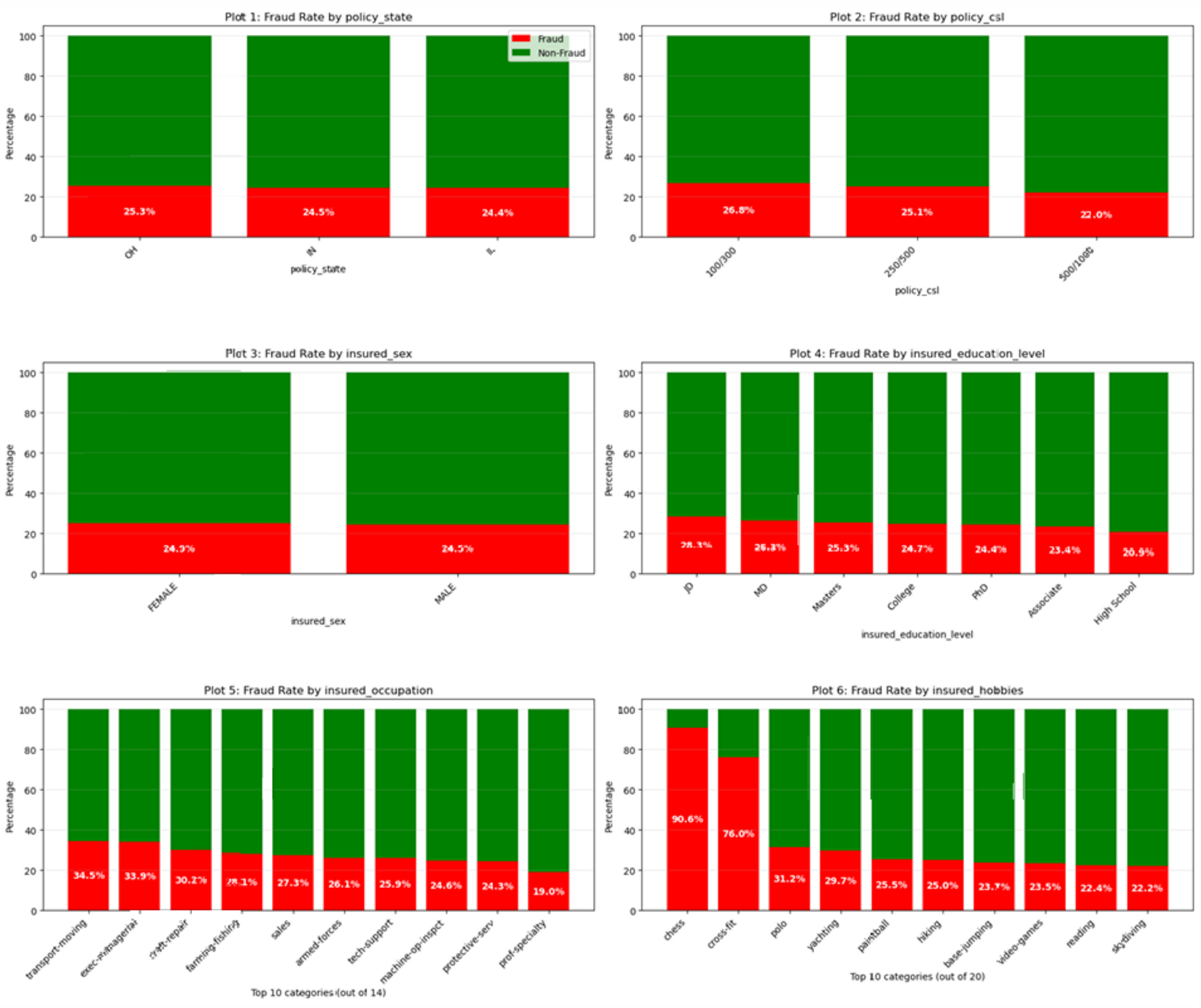
Imbalance ratio (majority:minority): 3.04:1

The dataset shows significant class imbalance. This may affect model performance.

Consider using techniques such as:

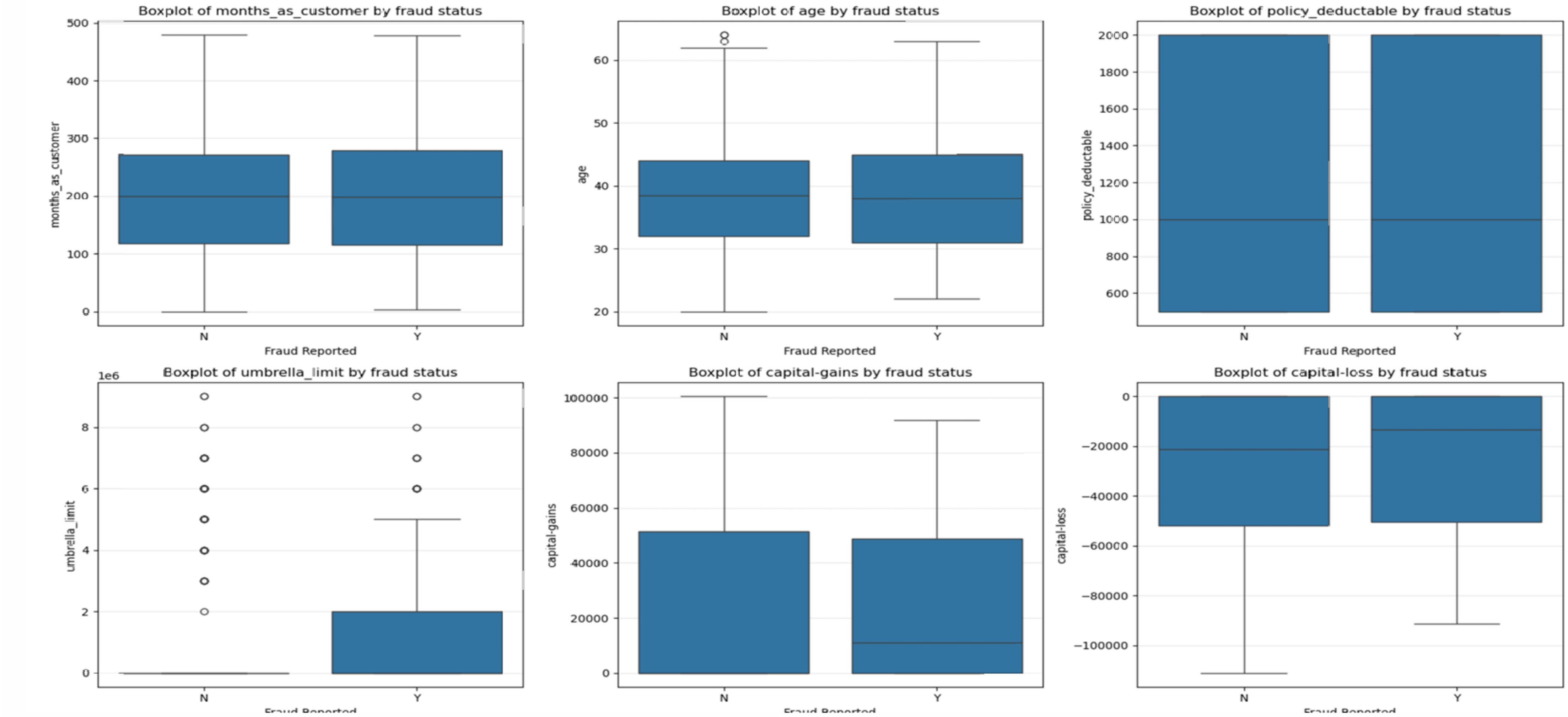
1. Resampling methods (oversampling minority class or undersampling majority class)
2. Using class weights during model training
3. Using algorithms that handle imbalanced data well
4. Using evaluation metrics appropriate for imbalanced datasets (e.g., precision, recall, F1-score, AUC-ROC)

Bivariate Analysis

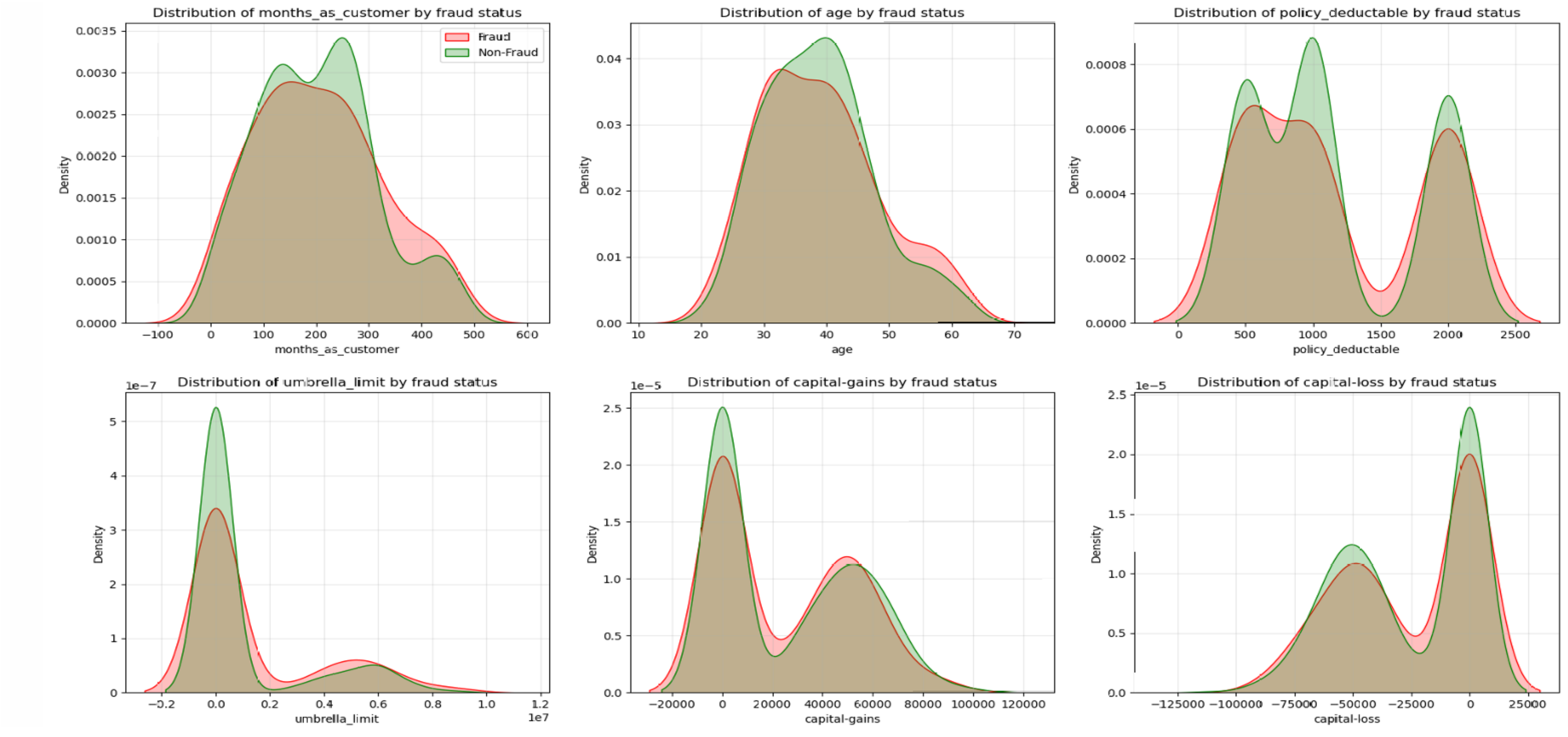


Feature importance based on variance in fraud rates:
incident_severity: 655.5417
insured_hobbies: 437.9118
auto_model: 138.9059 I
ncident_type: 127.9124
collision_type: 97.4883
incident_state: 73.1274
property_damage: 39.8805
insured_occupation: 39.3522
auto_make: 27.8186
insured_relationship: 24.6759
authorities_contacted: 23.6709
incident_city: 14.4581
policy_csl: 6.0253
insured_education_level: 5.3411 police_report_available: 2.1569
policy_state: 0.2506
insured_sex: 0.0773
Categorical features with low variance may not contribute much to explaining fraud.

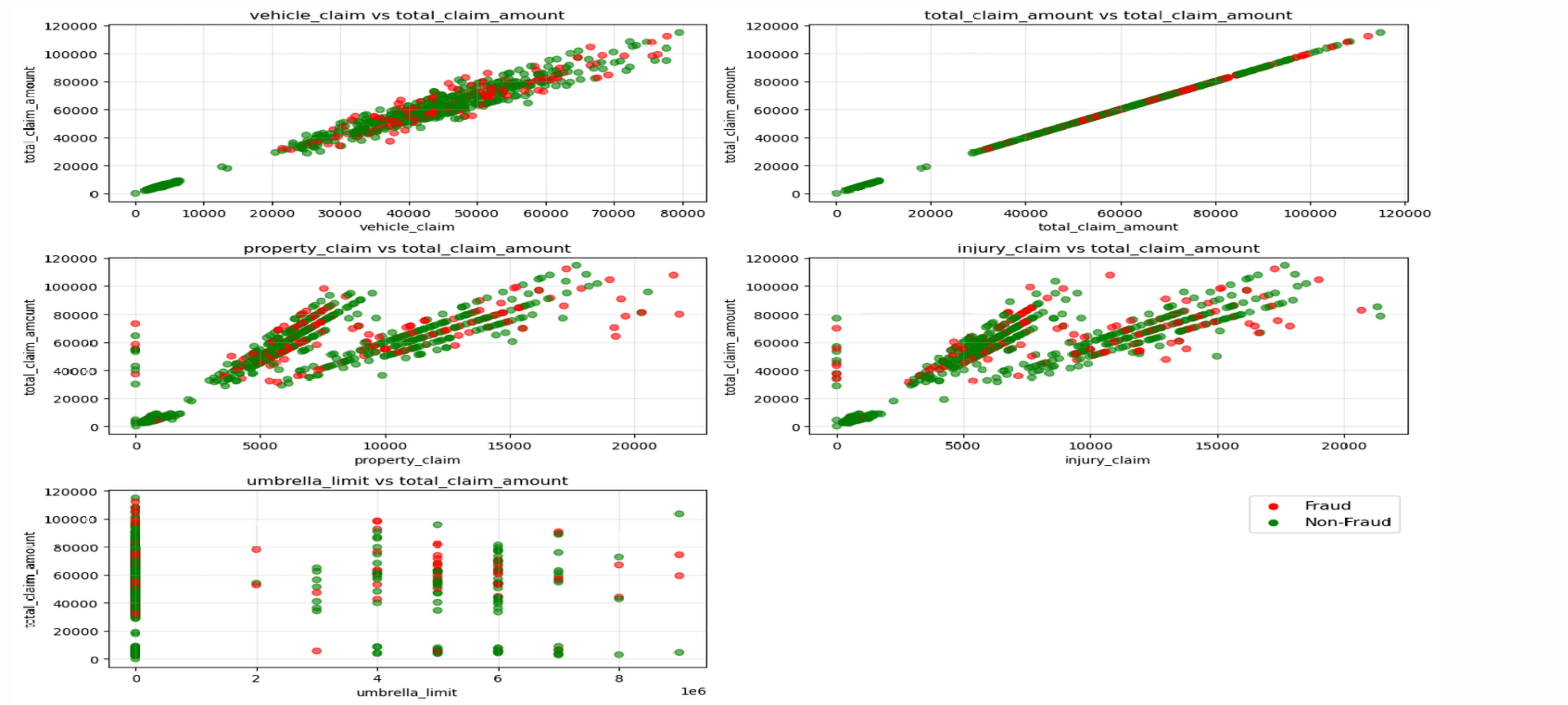
Analyze the association between numerical variables and fraud occurrence



Analyze the association between numerical variables and fraud occurrence



Analyze the association between numerical variables and fraud occurrence



Analyze the association between numerical variables and fraud occurrence

```
Statistical comparison of numerical features between fraud and non-fraud cases:
             months_as_customer                          age          \
                          mean median        std      mean median
fraud_reported
N                   201.277567  199.5  111.618274  38.781369   38.5
Y                   206.479769  199.0  119.634112  39.046243   38.0

                  policy_deductable                    umbrella_limit  \
              std              mean  median        std            mean
fraud_reported
N          8.832824      1144.486692  1000.0  603.196267
9.866920e+05
Y          9.605482      1167.630058  1000.0  634.010529
1.352601e+06

             ... injury_claim property_claim                     \
             ...          std         mean  median          std
fraud_reported  ...
N            ...  4973.403517   6981.083650  6665.0  4765.613193
Y            ...  4663.304029   8670.462428  7420.0  4770.279941

             vehicle_claim                          auto_year          \
                      mean    median          std        mean  median
fraud_reported
N            36109.448669  41305.0  19952.511542  2005.108365  2005.0
Y            43811.502890  45360.0  15040.163982  2004.514451  2004.0


                 std
fraud_reported
N          5.994414
Y          5.856429
```

```
Feature importance based on effect size (Cohen's d):
vehicle_claim                 0.408390
total_claim_amount            0.402311
property_claim                0.354408
injury_claim                  0.252318
umbrella_limit                0.163139
number_of_vehicles_involved   0.132975
witnesses                     0.132495
auto_year                     0.099639
bodily_injuries               0.092743
incident_hour_of_the_day      0.066322
months_as_customer            0.045774
policy_deductable             0.037881
capital-loss                  0.033652
age                           0.029334
capital-gains                 0.028634
dtype: float64
```

Model Selection

Models

- Logistic Regression
- Random Forest Classifier

Logistic Regression + RFECV

Logistic Regression + RFECV:

- Recursive elimination with cross validation selected the top ~52 predictors.

Number of selected features: 52

Selected features:

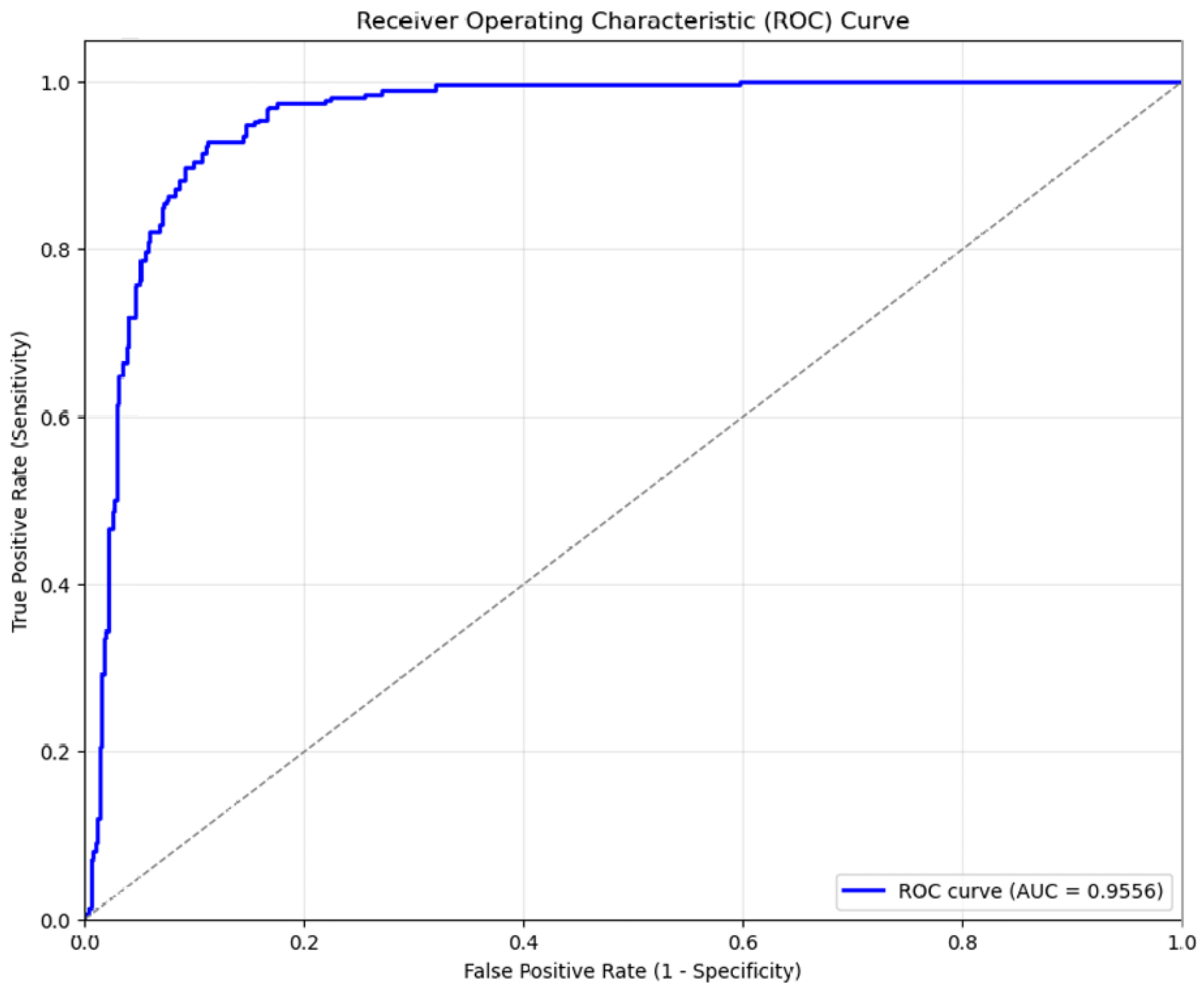
['policy_csl_250/500', 'insured_education_level_JD', 'insured_education_level_MD', 'insured_education_level_PhD', 'insured_occupation_exec-managerial', 'insured_occupation_farming-fishing', 'insured_occupation_handlers-cleaners', 'insured_occupation_other-service', 'insured_occupation_priv-house-serv', 'insured_hobbies_camping', 'insured_hobbies_chess', 'insured_hobbies_cross-fit', 'insured_hobbies_dancing', 'insured_hobbies_golf', 'insured_hobbies_movies', 'insured_hobbies_sleeping', 'insured_hobbies_video-games', 'insured_relationship_not-in-family', 'insured_relationship_own-child', 'insured_relationship_unmarried', 'incident_type_Vehicle Theft', 'collision_type_Side Collision', 'collision_type_Unknown', 'incident_severity_Minor Damage', 'incident_severity_Total Loss', 'incident_severity_Trivial Damage', 'incident_state_NY', 'incident_state_OH', 'incident_state_PA', 'incident_state_WV', 'incident_city_Northbrook', 'property_damage_Unknown', 'property_damage_YES', 'auto_make_Audi', 'auto_make_BMW', 'auto_make_Chevrolet', 'auto_make_Nissan', 'auto_model_A5', 'auto_model_Camry', 'auto_model_Civic', 'auto_model_F150', 'auto_model_Fusion', 'auto_model_Grand Cherokee', 'auto_model_Legacy', 'auto_model_MDX', 'auto_model_Other', 'auto_model_Pathfinder', 'auto_model_Silverado', 'auto_model_Ultima', 'auto_model_Wrangler', 'auto_model_X5', 'age_group_Young']

Logistic Regression

```
Optimization terminated successfully.
         Current function value: 0.271020
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:         fraud_reported   No. Observations:                 1052
Model:                          Logit   Df Residuals:                     1001
Method:                           MLE   Df Model:                           50
Date:                Sun, 20 Apr 2025   Pseudo R-squ.:                  0.6090
Time:                        23:21:11   Log-Likelihood:                -285.11
converged:                       True   LL-Null:                       -729.19
Covariance Type:            nonrobust   LLR p-value:                 8.129e-154
=========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                    1.7584      0.399      4.404      0.000       0.976       2.541
policy_csl_250/500                       0.7091      0.247      2.877      0.004       0.226       1.192
insured_education_level_JD               0.8224      0.333      2.469      0.014       0.169       1.475
insured_education_level_MD               1.2107      0.344      3.516      0.000       0.536       1.886
insured_education_level_PhD              0.9629      0.358      2.693      0.007       0.262       1.664
insured_occupation_exec-managerial       0.5531      0.427      1.295      0.195      -0.284       1.390
insured_occupation_farming-fishing      -1.2992      0.614     -2.118      0.034      -2.502      -0.097
insured_occupation_handlers-cleaners    -2.2441      0.606     -3.704      0.000      -3.431      -1.057
insured_occupation_other-service        -1.3984      0.509     -2.747      0.006      -2.396      -0.401
insured_occupation_priv-house-serv      -1.2727      0.498     -2.558      0.011      -2.248      -0.298
insured_hobbies_camping                 -0.9862      0.578     -1.707      0.088      -2.119       0.146
insured_hobbies_chess                    7.1086      0.720      9.875      0.000       5.698       8.519
insured_hobbies_cross-fit                4.5713      0.639      7.154      0.000       3.319       5.824
insured_hobbies_dancing                 -1.9412      0.782     -2.481      0.013      -3.474      -0.408
insured_hobbies_movies                  -0.9414      0.637     -1.477      0.140      -2.190       0.307
insured_hobbies_sleeping                -1.8309      0.542     -3.379      0.001      -2.893      -0.769
insured_hobbies_video-games              1.9561      0.450      4.347      0.000       1.074       2.838
insured_relationship_not-in-family       1.1784      0.330      3.568      0.000       0.531       1.826
insured_relationship_own-child          -0.4121      0.342     -1.205      0.228      -1.082       0.258
insured_relationship_unmarried           0.6007      0.343      1.752      0.080      -0.071       1.273
incident_type_Vehicle Theft             -0.4407      0.734     -0.601      0.548      -1.879       0.997
collision_type_Side Collision           -1.0742      0.279     -3.852      0.000      -1.621      -0.528
collision_type_Unknown                   0.4950      0.628      0.788      0.430      -0.735       1.725
incident_severity_Minor Damage          -5.4665      0.448    -12.202      0.000      -6.345      -4.588
incident_severity_Total Loss            -4.3908      0.356    -12.321      0.000      -5.089      -3.692
incident_severity_Trivial Damage        -5.8259      0.859     -6.781      0.000      -7.510      -4.142
incident_state_NY                       -0.6572      0.298     -2.204      0.028      -1.241      -0.073
```

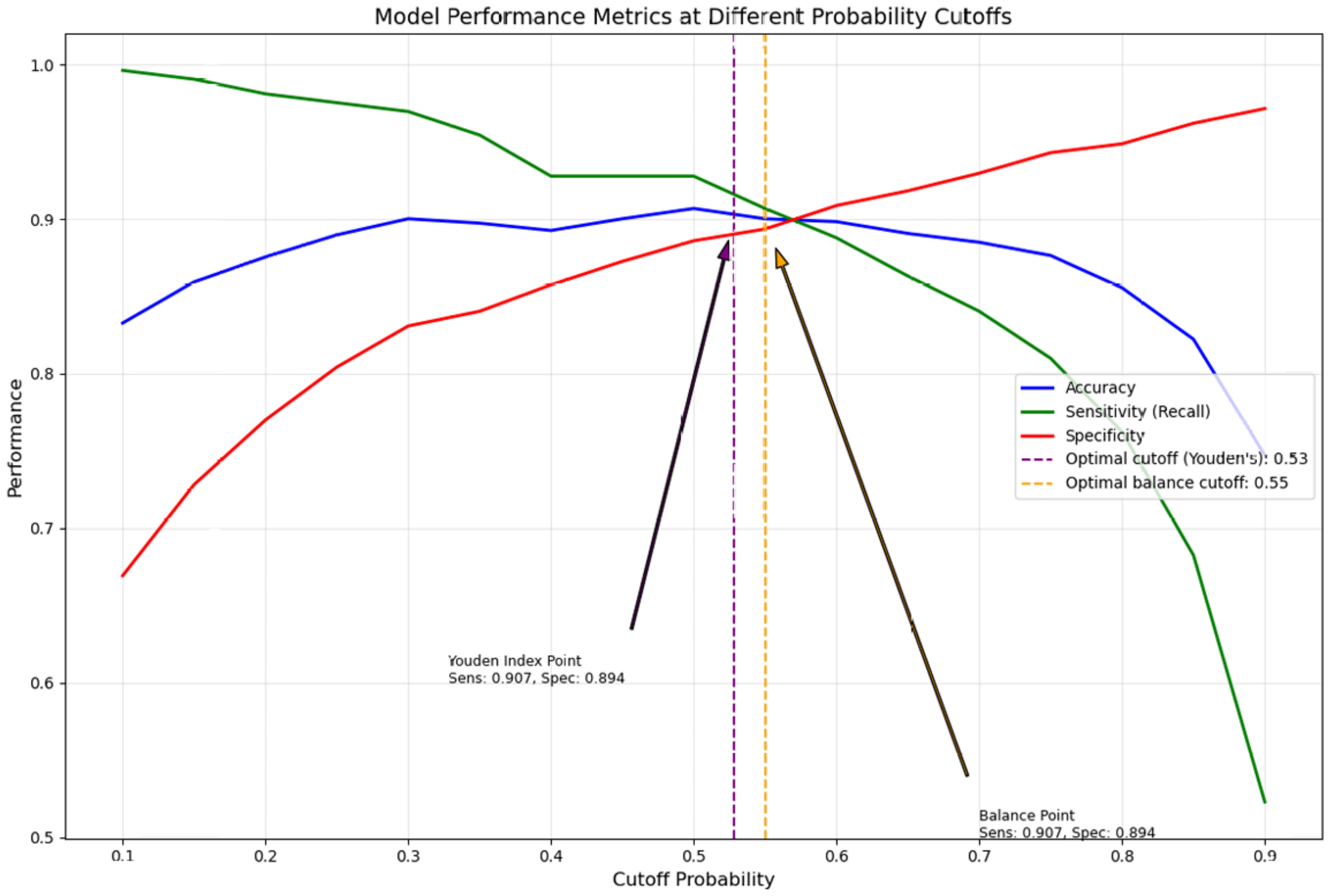
```
VIF values for the refitted model:
                               Feature        VIF
0                                const  13.007234
22              collision_type_Unknown   3.122945
36                   auto_make_Nissan   2.877511
25    incident_severity_Trivial Damage   2.284443
34                      auto_make_BMW   2.277934
49                      auto_model_X5   2.046408
47                   auto_model_Ultima   1.905181
35                auto_make_Chevrolet   1.881893
45               auto_model_Pathfinder   1.868680
20         incident_type_Vehicle Theft   1.859962
46                auto_model_Silverado   1.772034
23      incident_severity_Minor Damage   1.680774
44                    auto_model_Other   1.640283
31             property_damage_Unknown   1.555763
32                 property_damage_YES   1.531561
24        incident_severity_Total Loss   1.377009
29                  incident_state_WV   1.296598
11               insured_hobbies_chess   1.288843
39                    auto_model_F150   1.275981
26                   incident_state_NY   1.258974
28                   incident_state_PA   1.252816
33                     auto_make_Audi   1.244682
17  insured_relationship_not-in-family   1.234295
37                   auto_model_Camry   1.227096
19      insured_relationship_unmarried   1.222783
27                   incident_state_OH   1.216750
2           insured_education_level_JD   1.212843
18      insured_relationship_own-child   1.180312
3           insured_education_level_MD   1.179291
4          insured_education_level_PhD   1.162443
21       collision_type_Side Collision   1.153005
5   insured_occupation_exec-managerial   1.139606
12           insured_hobbies_cross-fit   1.136269
42                  auto_model_Legacy   1.131412
40                  auto_model_Fusion   1.129742
43                     auto_model_MDX   1.129205
16         insured_hobbies_video-games   1.127188
38                   auto_model_Civic   1.117150
```

Logistic Regression – ROC Curve



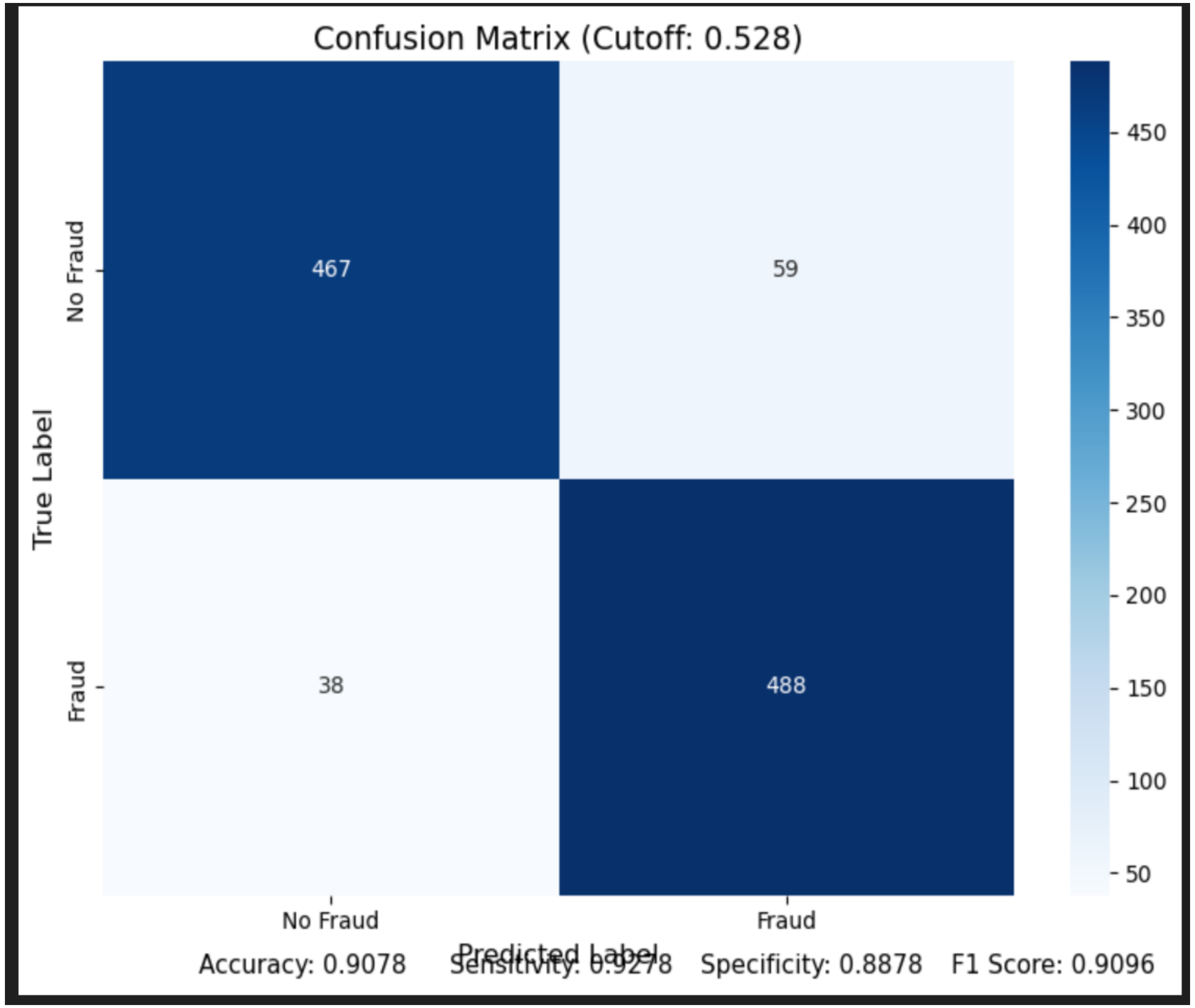
Optimal threshold based on Youden's index: 0.5282 At this threshold -
Sensitivity: 0.9278,
Specificity: 0.8878
Optimal cutoff value: 0.5282

Logistic Regression – Optimal Cutoff



Optimal cutoff where sensitivity and specificity are closest: 0.5500
At this cutoff - Sensitivity: 0.9068, Specificity: 0.8935
Accuracy at this cutoff: 0.9002

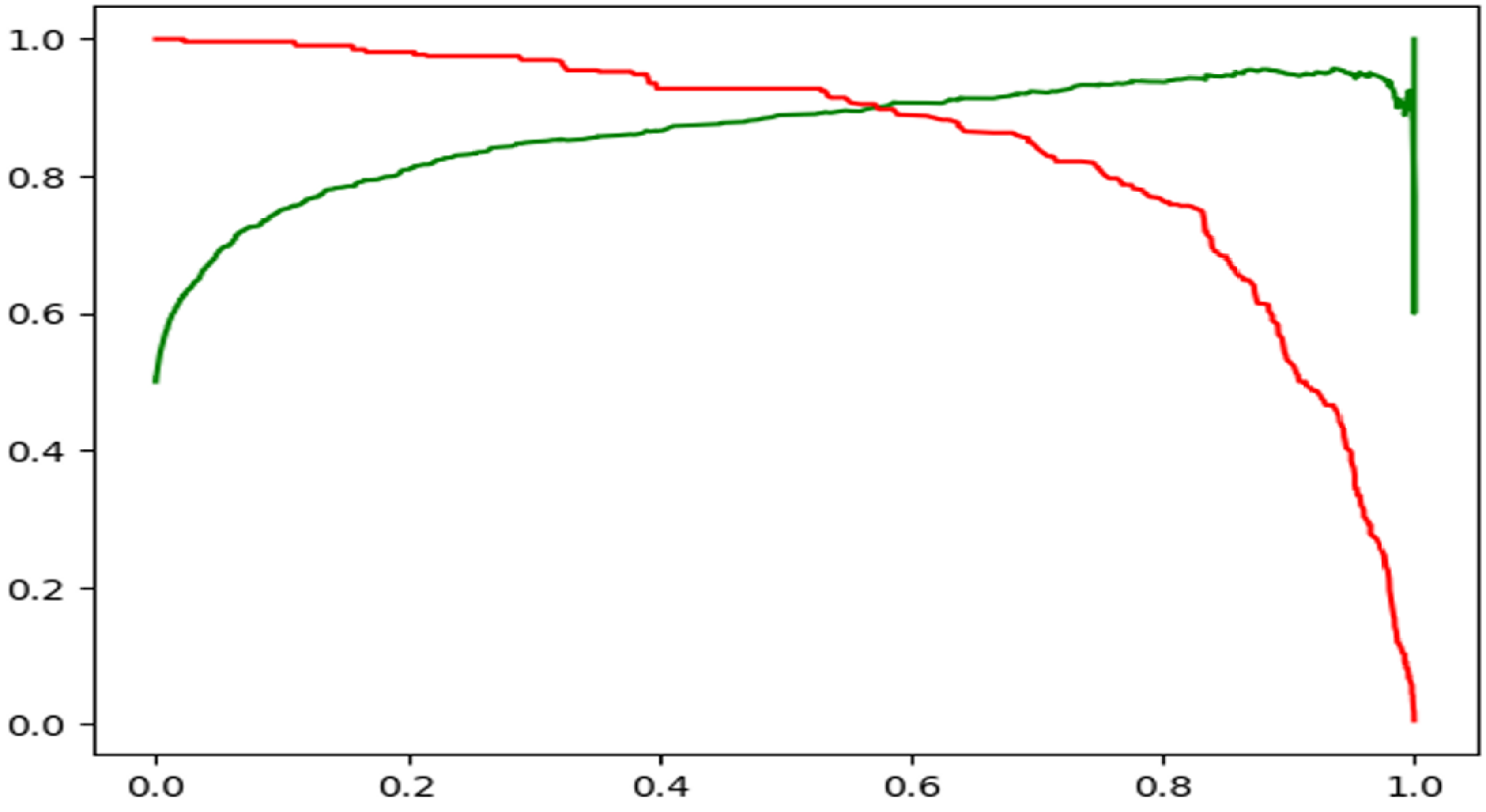
Confusion Matrix



Confusion Matrix using optimal cutoff:
[[467 59]
 [38 488]]

Model performance metrics using optimal cutoff (0.5282):
Accuracy: 0.9078
Sensitivity (True Positive Rate): 0.9278
Specificity (True Negative Rate): 0.8878
Precision: 0.8921
Recall: 0.9278
F1 Score: 0.9096

Precision Recall Curve

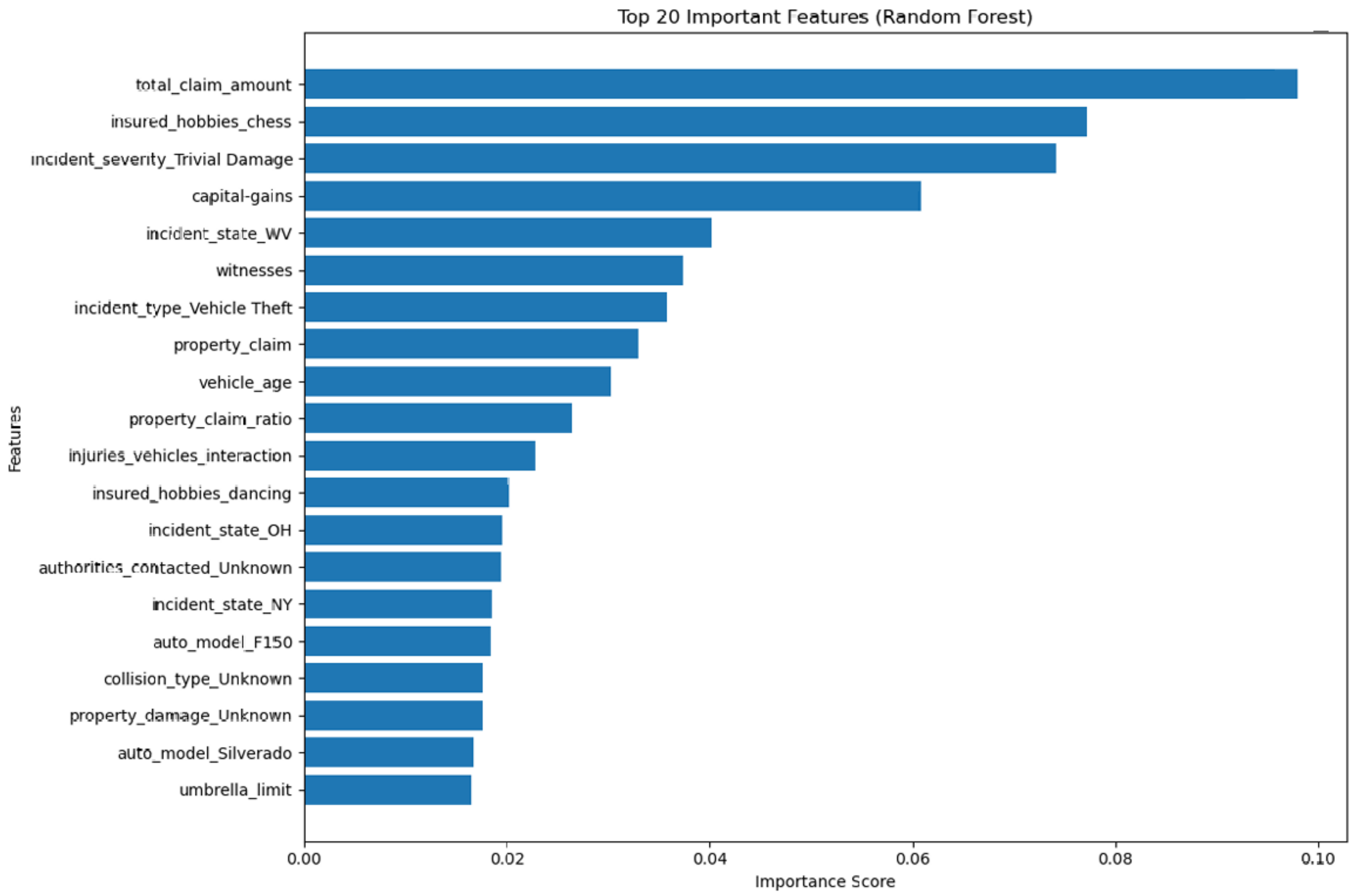


Random Forest

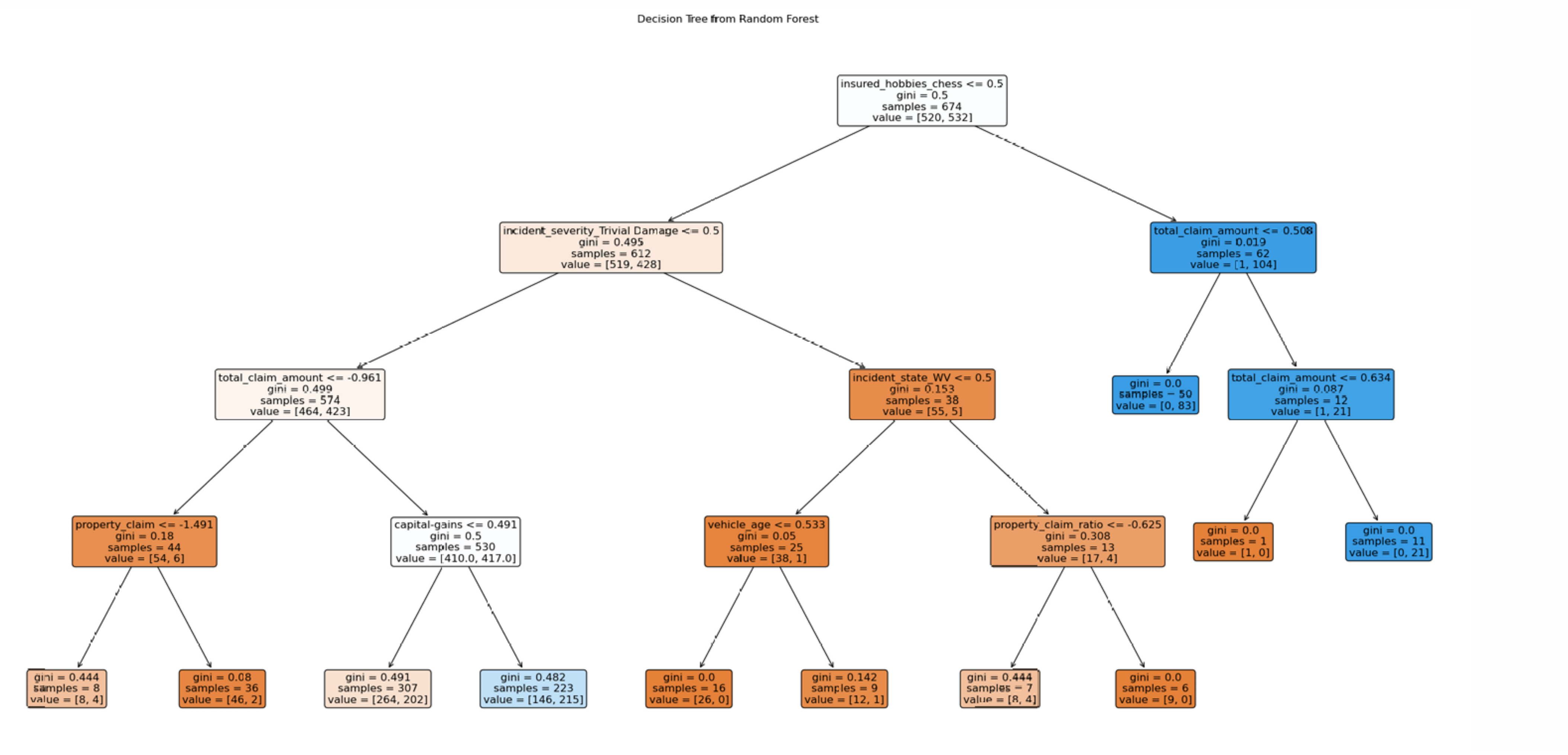
- Feature importance thresholding (0.01) retained 28 variables; top features included.
- Number of selected features based on importance threshold (0.01): 28
- Selected features based on importance threshold:

  ['total_claim_amount', 'insured_hobbies_chess', 'incident_severity_Trivial Damage', 'capital-gains', 'incident_state_WV', 'witnesses', 'incident_type_Vehicle Theft', 'property_claim', 'vehicle_age', 'property_claim_ratio', 'injuries_vehicles_interaction', 'insured_hobbies_dancing', 'incident_state_OH', 'authorities_contacted_Unknown', 'incident_state_NY', 'auto_model_F150', 'collision_type_Unknown', 'property_damage_Unknown', 'auto_model_Silverado', 'umbrella_limit', 'capital-loss', 'insured_hobbies_board-games', 'auto_model_95', 'incident_city_Riverwood', 'policy_deductable', 'insured_hobbies_movies', 'incident_day_of_week', 'insured_hobbies_bungie-jumping']

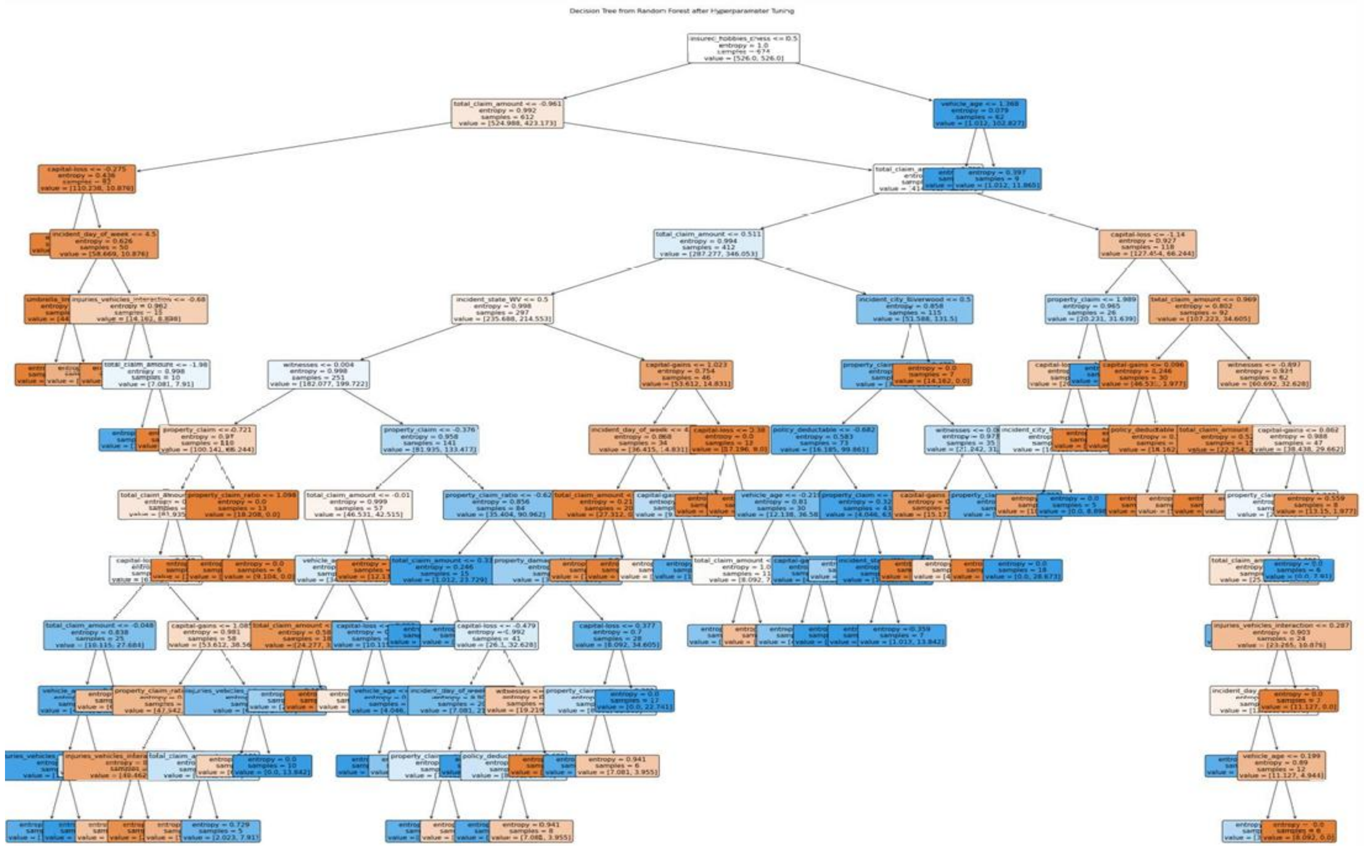
Random Forest – Feature Importance



Random Forest– Decision Tree based on feature selection



Random Forest - Hyperparameter Tuning

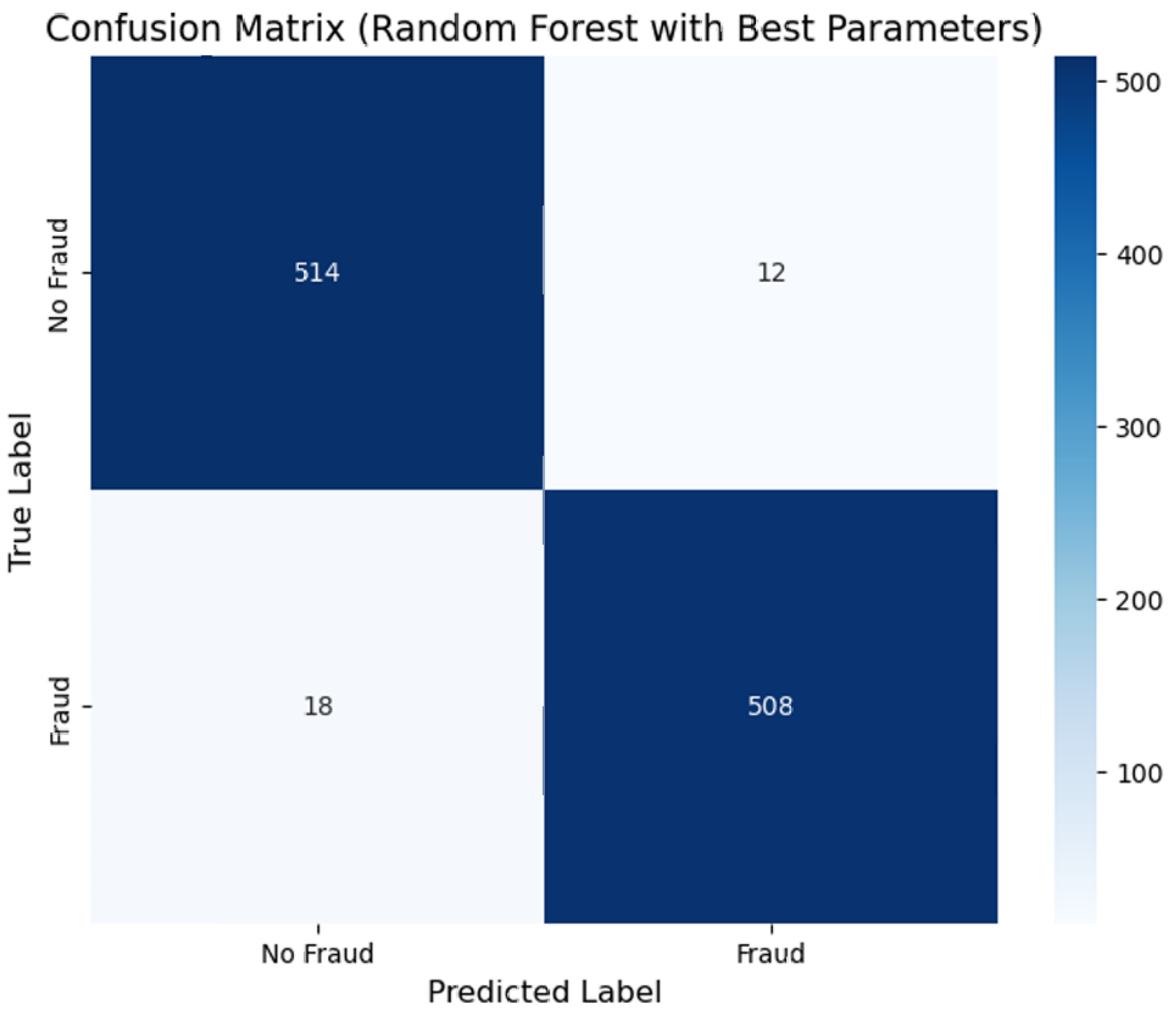


Starting grid search for hyperparameter tuning...
Fitting 5 folds for each of 972 candidates, totalling 4860 fits

Best Parameters:
{'bootstrap': True, 'class_weight': 'balanced_subsample', 'criterion': 'entropy', 'max_depth': 12, 'max_features': 0.5, 'min_samples_leaf': 5, 'min_samples_split': 5, 'n_estimators': 200}

Best ROC-AUC Score: 0.9342

Random Forest – Confusion Matrix



Confusion Matrix:
[[514 12]
 [18 508]]

Random Forest Model with Best Parameters:
Accuracy: 0.9715
Sensitivity (True Positive Rate): 0.9658
Specificity (True Negative Rate): 0.9772
Precision: 0.9769
Recall: 0.9658
F1 Score: 0.9713

Prediction & Model evaluation

Model	Validation Accuracy	Sensitivity (TPR)	Specificity (TNR)	Precision	Recall	F1-Score
Logistic Regression Optimized cutoff (0.5282)	0.8000	0.6757	0.8407	0.5814	0.6757	0.6250
Random Forest (Hyperparameter Tuning)	0.7233	0.3378	0.8496	0.4237	0.3378	0.3759

Logistic Regression (Optimized cutoff at 0.5282)

- Achieves 80.00% validation accuracy
- Shows good sensitivity/recall at 67.57% (effectively captures true positives)
- Maintains high specificity at 84.07% (effectively identifies true negatives)
- Delivers precision of 58.14% (moderate confidence in positive predictions)
- Results in F1-Score of 62.50% (balanced performance between precision and recall)

Random Forest

- Reaches 72.33% validation accuracy
- Demonstrates poor sensitivity at only 33.78% (misses many positive cases)
- Maintains high specificity at 84.96% (slightly better than Logistic Regression)
- Shows lower precision at 42.37% (less confidence in positive predictions)
- Results in a substantially lower F1-Score of 37.59%

Prediction and Model Evaluation: Conclusion

The **Logistic Regression** model with a tuned probability threshold outperforms Random Forest in detecting fraudulent claims. While both models demonstrate comparable **specificity** (accurately identifying non-fraudulent claims), Logistic Regression excels in **sensitivity**, capturing more true fraud cases. It also achieves a better trade-off between **precision and recall**, as reflected in a higher **F1-Score**. This makes it a more effective choice for fraud detection, where spotting fraudulent activity is critical.

Questions

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

Analyzing Historical Claim Data

1.Exploratory Data Analysis (EDA):

Uncover relationships between features and fraudulent behavior

2.Feature Engineering:

Create derived variables (e.g., claim-to-policy ratio) to enhance signal detection

3.Statistical Analysis:

Detect anomalies and outliers that may indicate potential fraud

4.Predictive Modeling:

Use machine learning models like **Logistic Regression** and **Random Forest** to capture hidden fraud patterns

5.ROC Curve Analysis:

Identify optimal probability cutoffs to balance fraud detection and false positives

6.Model Evaluation:

Measure effectiveness using metrics such as **sensitivity** (recall) and **specificity**

Questions

2. Which features are most predictive of fraudulent behaviour?

Key Predictive Features Identified

1. **Total Claim Amount**
   - Higher claim values are more frequently associated with fraud
2. **Hobbies (e.g., Chess, Dancing)**
   - Certain hobbies reflect demographic trends linked to fraudulent behavior
3. **Incident Severity**
   - Minor or trivial damage claims show a stronger association with fraud
4. **Capital Gains/Losses**
   - Serve as indicators of the claimant's financial standing
5. **Geographic Location**
   - States like **WV, NY, and OH** exhibit higher fraud incidence
6. **Vehicle Type**
   - Specific models (e.g., **F150**, **Silverado**) are linked with higher fraud rates
7. **Incident Characteristics**
   - Claims involving **vehicle theft** or **ambiguous collision types** raise fraud likelihood
8. **Property Damage Reporting**
   - Inconsistent or suspicious property damage reports are red flags

Questions

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

Predicting Fraud for New Claims

1.The **Logistic Regression** model achieved **80.00% validation accuracy** with strong **sensitivity** at **67.57%**
2.An **optimal probability threshold (~0.55)** was identified to balance false positives and false negatives
3.The model generates **probability scores** indicating the likelihood of fraud for each claim
4.The **Random Forest** model offers an alternative, though with **lower sensitivity** (**33.78%**)
5.These models can be **deployed in production** to score and flag incoming claims for potential fraud

Questions

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

Key Insights for Enhancing Fraud Detection

1. **Cutoff Optimization is Key**
   - The default 0.5 threshold is suboptimal for imbalanced datasets; fine-tuning improves detection
2. **Claim Characteristics Matter**
   - Flags should be raised for **trivial damage** and certain **vehicle models** with high fraud incidence
3. **Geographic Risk Patterns**
   - Claims originating from specific states warrant closer inspection
4. **Balancing Detection and Experience**
   - Adjusting the threshold helps manage the trade-off between **catching fraud** and **avoiding false alarms**
5. **Tiered Review Strategy**
   - Implement a multi-level review process based on fraud **probability scores**
6. **Model Choice Matters**
   - **Logistic Regression** with a tuned cutoff outperforms more complex models like Random Forest in this use case
7. **Use Demographics Responsibly**
   - Features like **hobbies** and **occupation** reveal patterns but must be used with care to avoid bias