FRAUDULENT CLAIM DETECTION

Vishal Verma
Vinti Singh

INTRODUCTION

Fraudulent insurance claims pose a significant challenge for insurers, resulting in substantial financial losses and inefficiencies. As claim volumes grow, traditional manual detection methods fall short. Data-driven approaches offer a more effective solution for identifying and preventing fraud.

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 & CLEANING

Missing Values

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

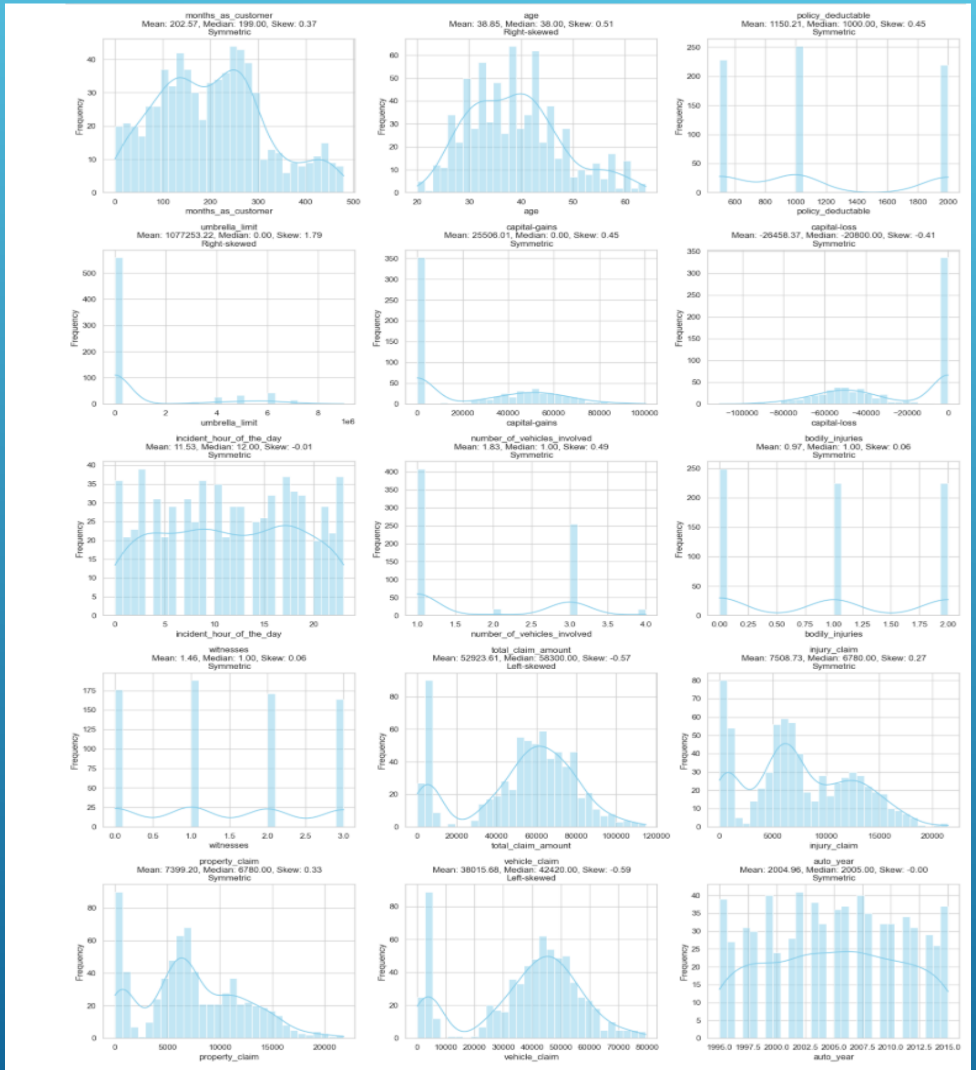
Redundant & Illogical Entries

- 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.

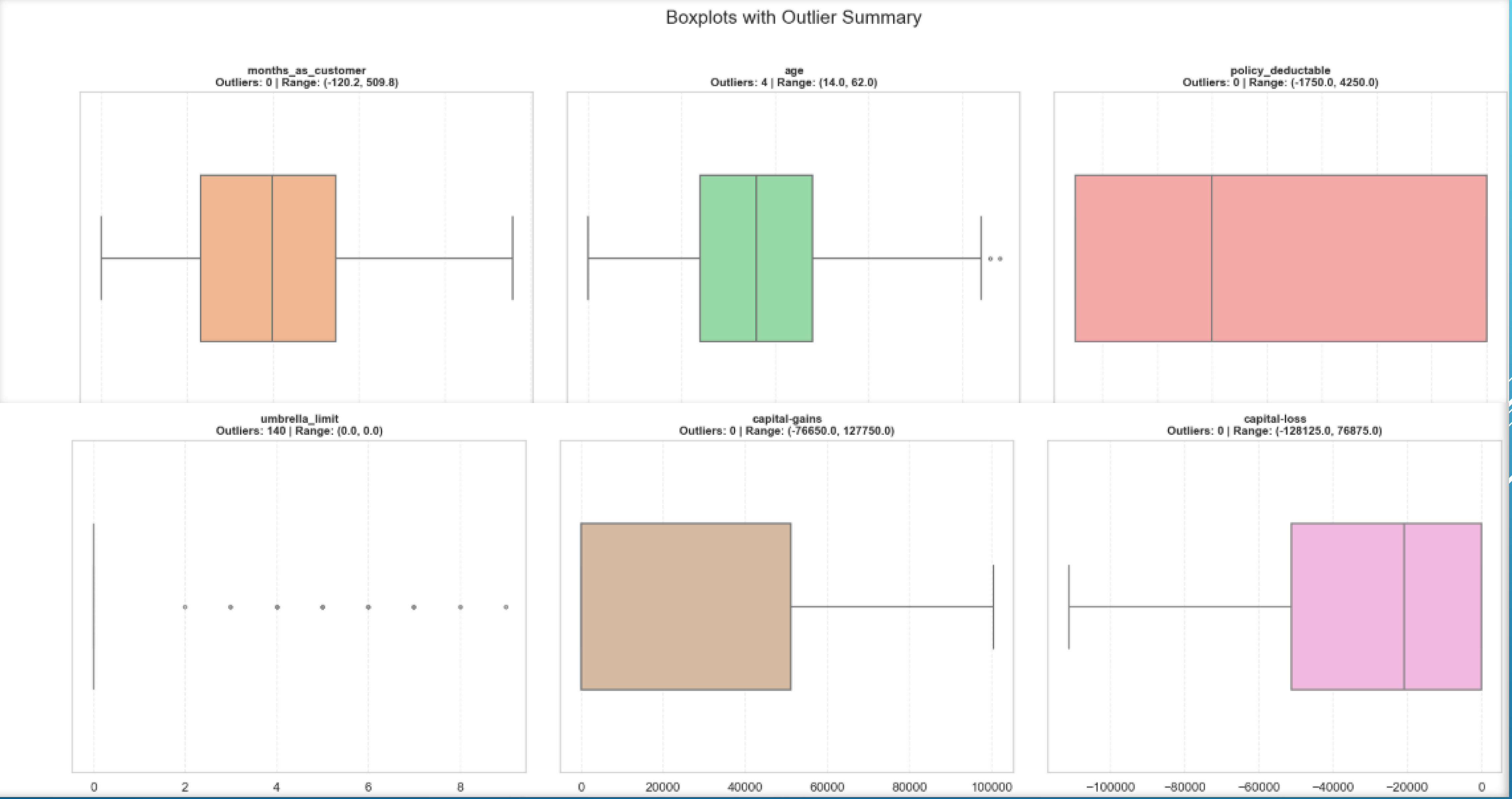
EDA – UNIVARIATE ANALYSIS



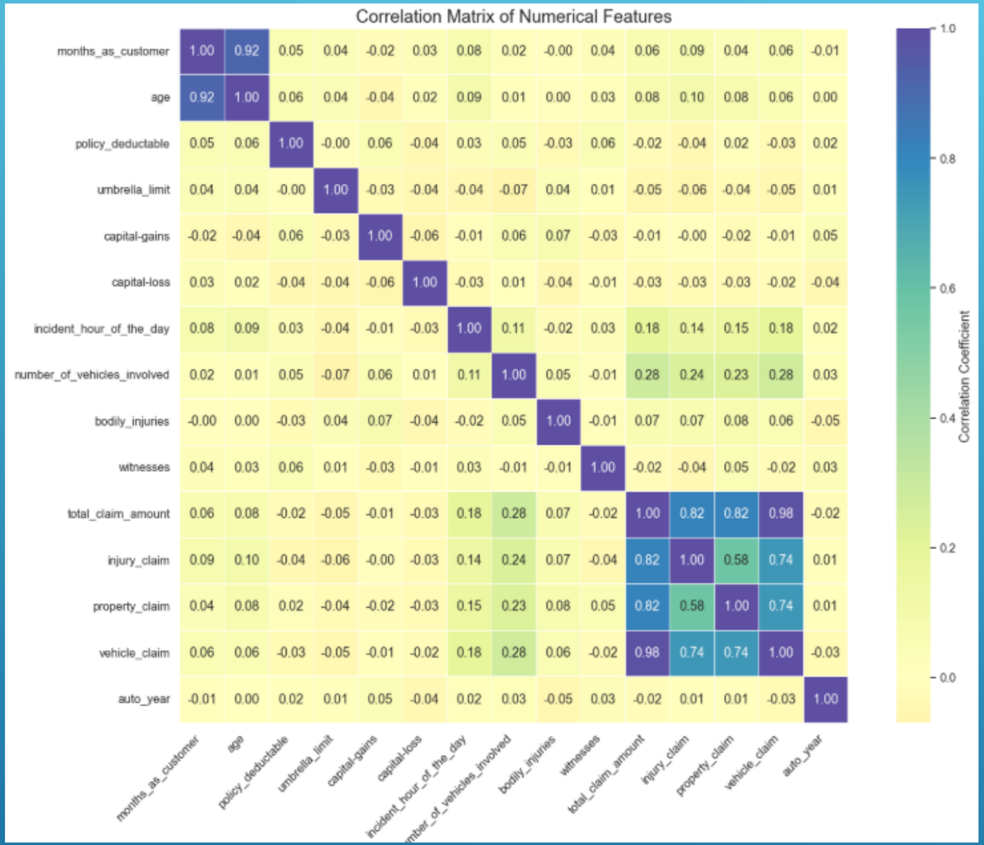
Observations from histogram plots:
months_as_customer:
- Mean: 202.57, Median: 199.00
- Skewness: 0.37
- Distribution appears approximately symmetric
age:
- Mean: 38.85, Median: 38.00
- Skewness: 0.51
- Distribution is positively skewed (right-tailed)
policy_deductable:
- Mean: 1150.21, Median: 1000.00
- 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

Observations from histogram plots:
- 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



CORRELATION MATRIX



Highly correlated feature (|correlation| > 0.7):

age vs months_as_customer: 0.920

injury_claim vs total_claim_amount: 0.818

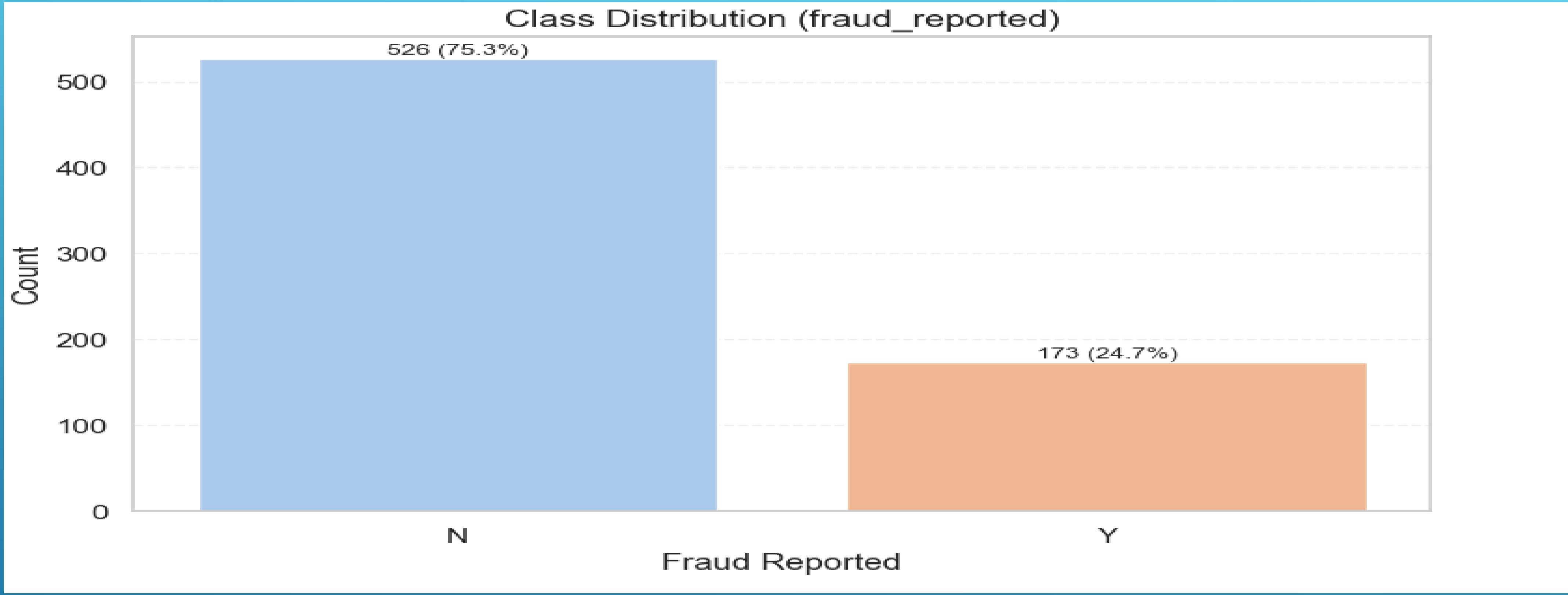
property_claim vs total_claim_amount: 0.815

vehicle_claim vs total_claim_amount: 0.984

vehicle_claim vs injury_claim: 0.743

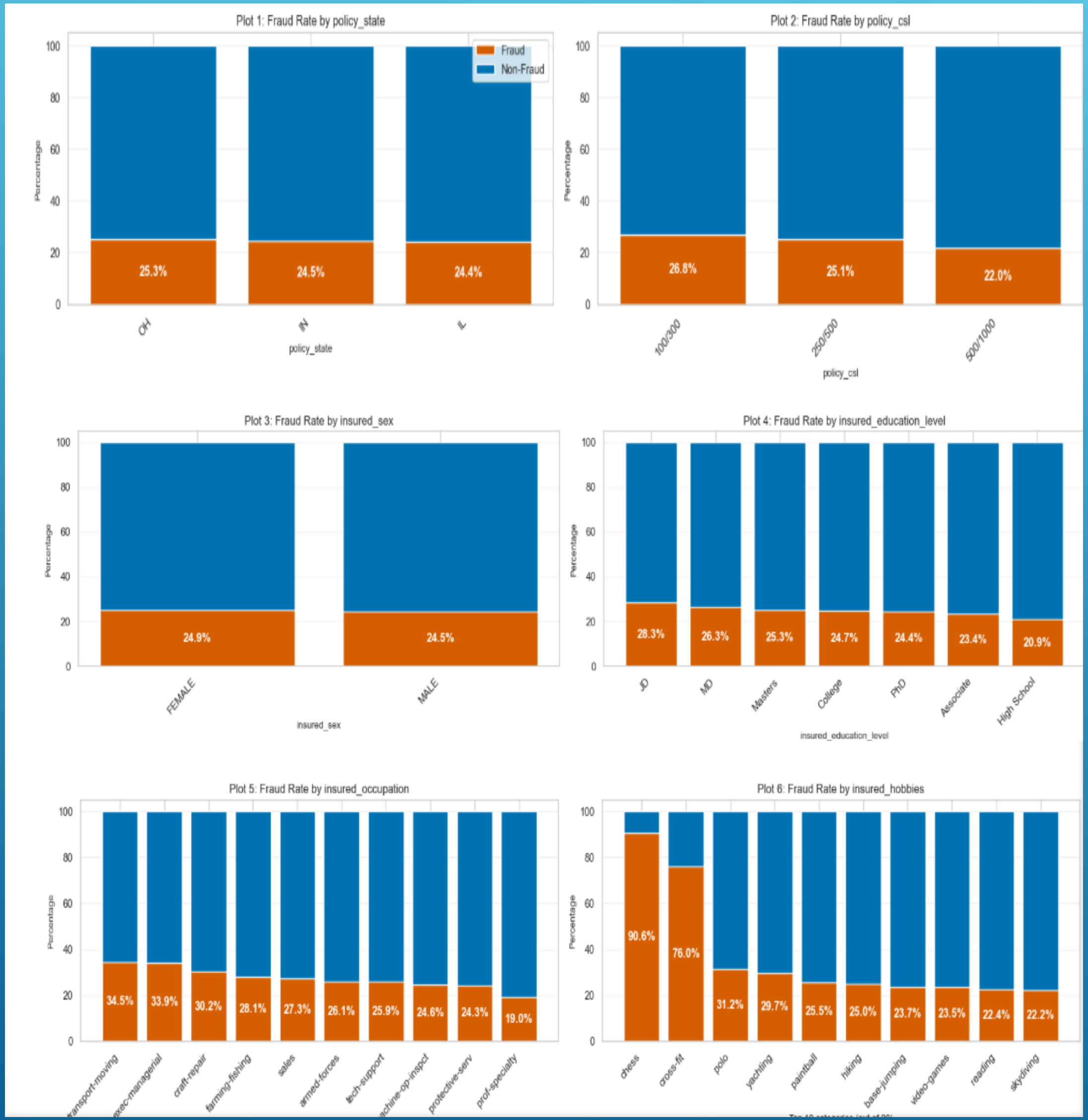
vehicle_claim vs property_claim: 0.742

CLASS IMBALANCE ANALYSIS



Imbalance Ratio: 3.04:1 (N vs Y)
Significant class imbalance detected. This may affect model performance.
Consider: resampling, class weights, or specialized metrics (F1, AUC, etc.)

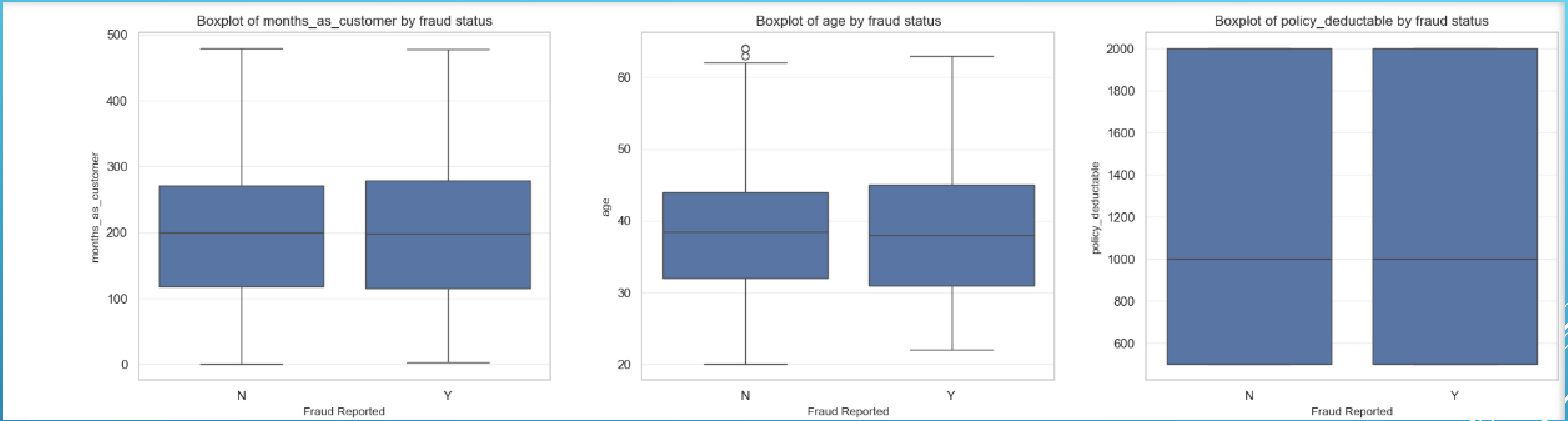
EDA – BIVARIATE ANALYSIS

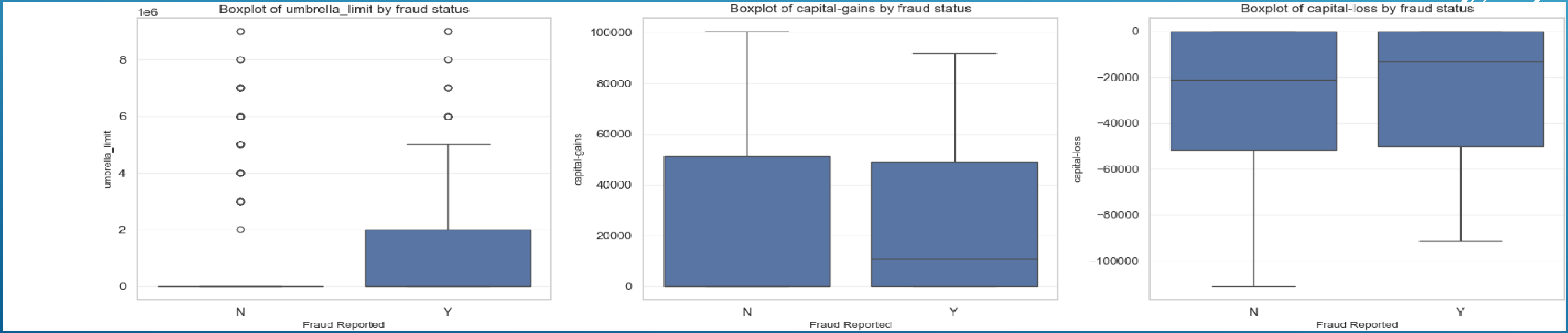


Feature importance based on variance in fraud rates:
incident_severity: 655.5417
insured_hobbies: 437.9118
auto_model: 138.9059
incident_type: 127.9124
collision_type: 97.4883
incident_state: 73.1274
property_damage: 39.8805
insured_occupation: 39.3522
auto_make: 27.8186
insured_relationship: 24.6759
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

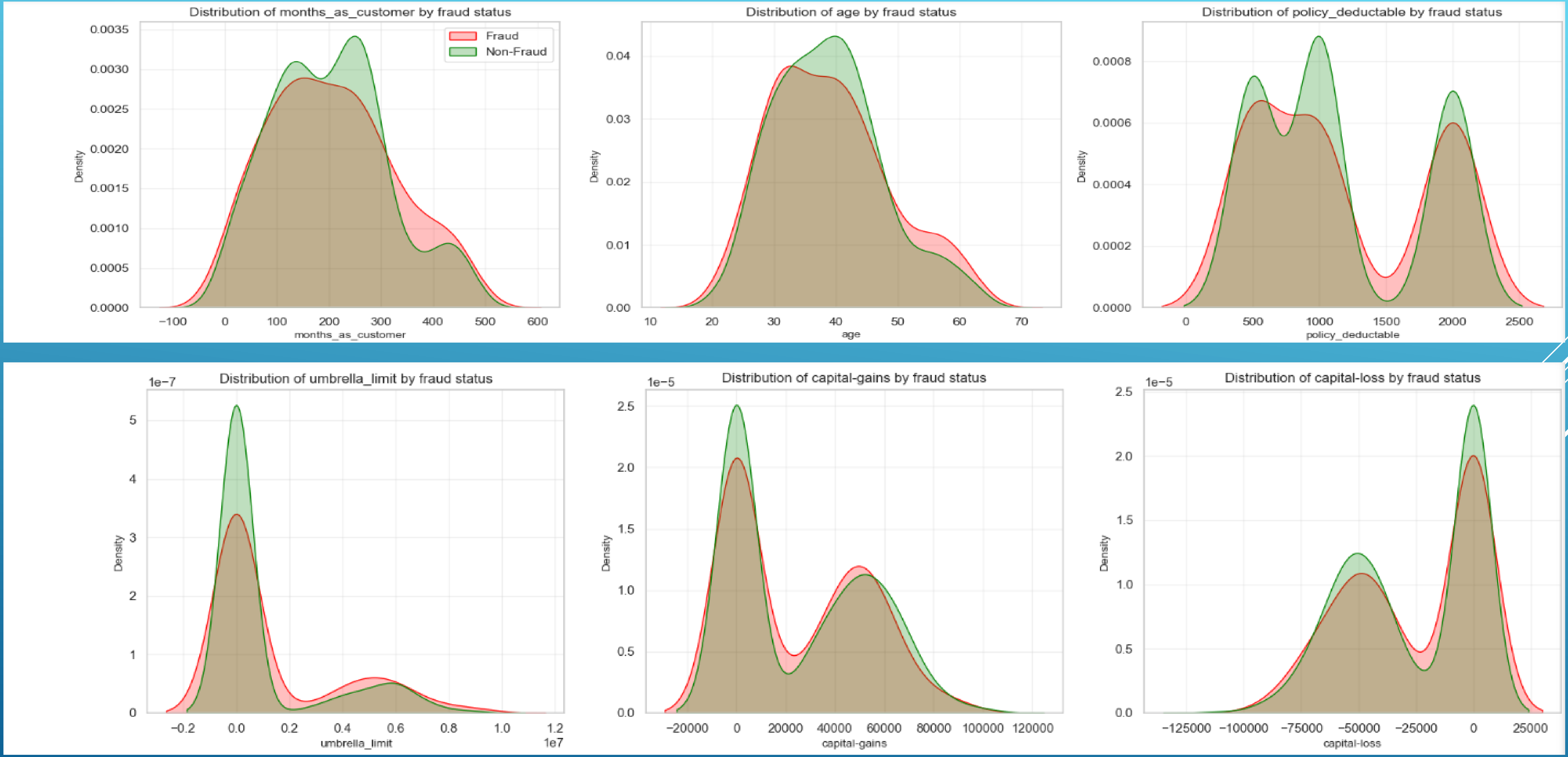
Categorical features with low variance may not contribute much to explaining fraud.

EDA –RELATIONSHIP B/W NUMERICAL FEATURES AND TARGET VARIABLE

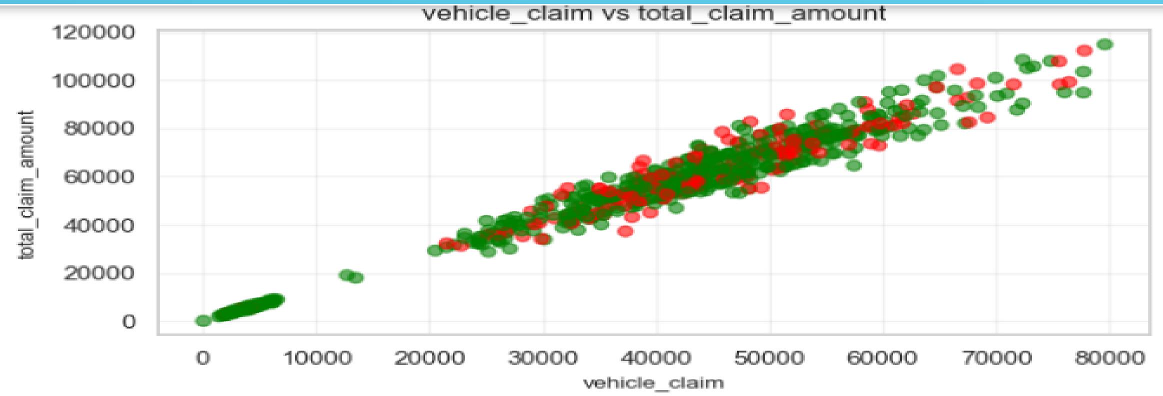


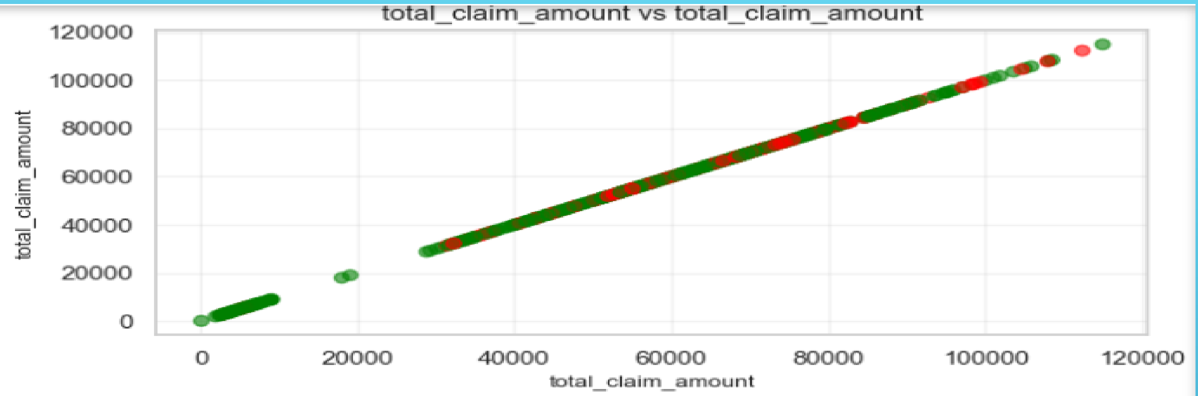


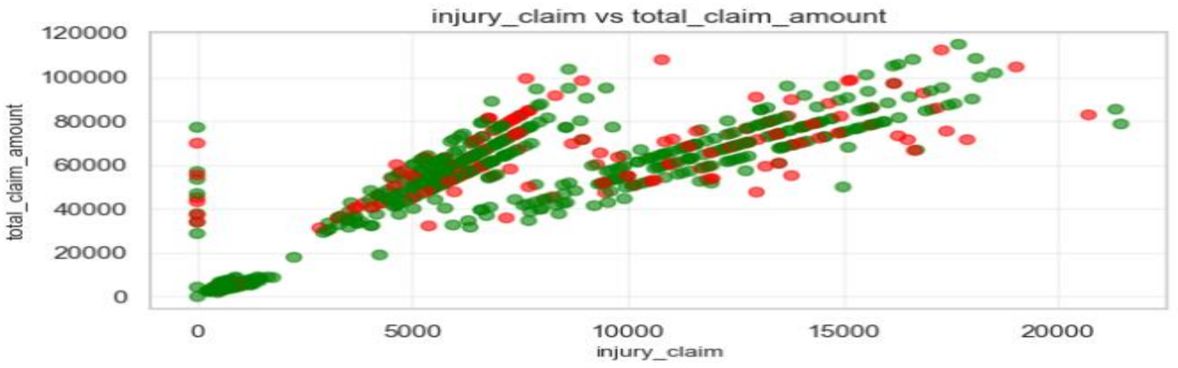
EDA –RELATIONSHIP B/W NUMERICAL FEATURES AND TARGET VARIABLE

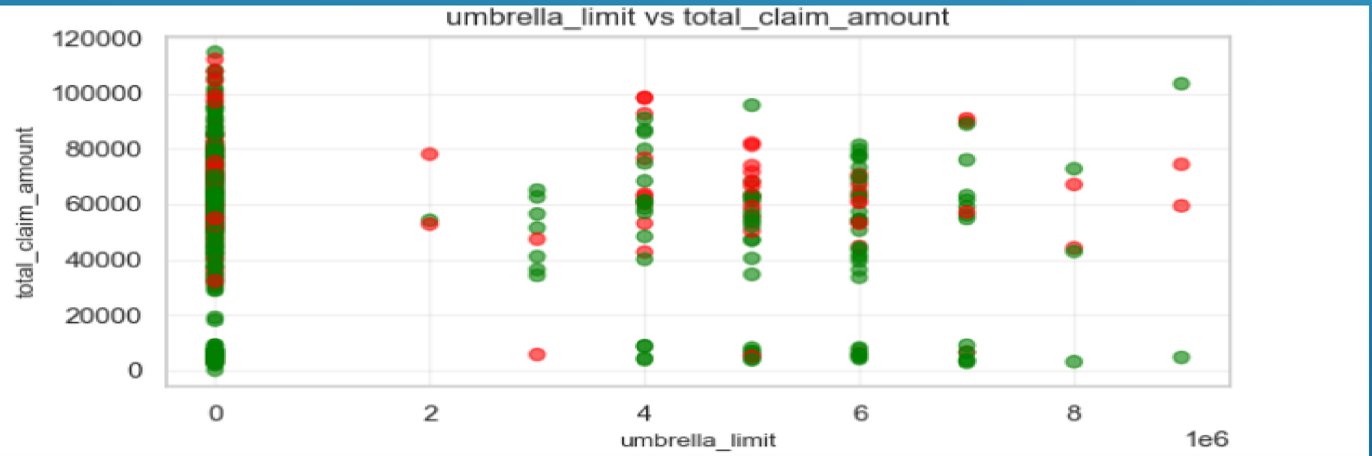


EDA –RELATIONSHIP B/W NUMERICAL FEATURES AND TARGET VARIABLE









MODEL SELECTION

Models

- Logistic Regression
- Random Forest Classifier

LOGISTIC REGRESSION+RFECV

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']

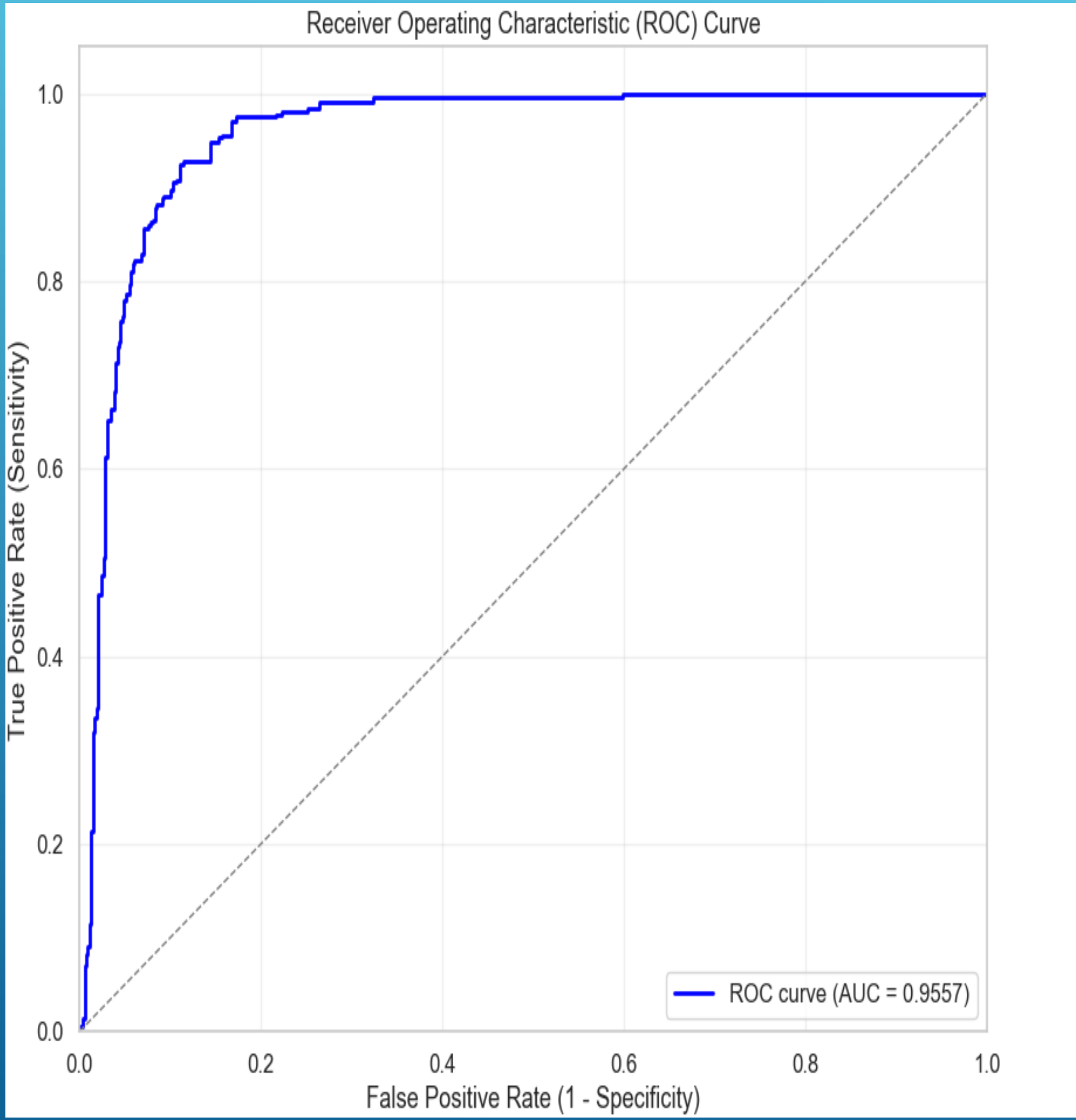
LOGISTIC REGRESSION

```
Optimization terminated successfully.
         Current function value: 0.270871
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:         fraud_reported   No. Observations:                 1052
Model:                          Logit   Df Residuals:                      999
Method:                           MLE   Df Model:                           52
Date:                Sun, 11 May 2025   Pseudo R-squ.:                  0.6092
Time:                        18:51:35   Log-Likelihood:                -284.96
converged:                       True   LL-Null:                       -729.19
Covariance Type:            nonrobust   LLR p-value:                 1.247e-152
==================================================================================================
                                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                                  1.7477      0.400      4.373      0.000       0.964       2.531
policy_csl_250/500                     0.7082      0.247      2.872      0.004       0.225       1.192
insured_education_level_JD             0.8216      0.336      2.449      0.014       0.164       1.479
insured_education_level_MD             1.2077      0.344      3.511      0.000       0.533       1.882
insured_education_level_PhD            0.9790      0.361      2.709      0.007       0.271       1.687
insured_occupation_exec-managerial     0.5662      0.429      1.320      0.187      -0.274       1.407
insured_occupation_farming-fishing    -1.3002      0.613     -2.120      0.034      -2.502      -0.098
insured_occupation_handlers-cleaners  -2.1783      0.632     -3.447      0.001      -3.417      -0.940
insured_occupation_other-service      -1.4148      0.512     -2.763      0.006      -2.418      -0.411
insured_occupation_priv-house-serv    -1.2779      0.499     -2.562      0.010      -2.255      -0.300
insured_hobbies_camping               -0.9977      0.579     -1.724      0.085      -2.132       0.137
insured_hobbies_chess                  7.0837      0.721      9.819      0.000       5.670       8.498
insured_hobbies_cross-fit              4.5590      0.639      7.134      0.000       3.307       5.811
insured_hobbies_dancing               -1.9433      0.784     -2.478      0.013      -3.480      -0.406
insured_hobbies_golf                  -0.1594      0.558     -0.286      0.775      -1.253       0.935
insured_hobbies_movies                -0.9588      0.639     -1.500      0.134      -2.211       0.294
insured_hobbies_sleeping              -1.7921      0.547     -3.275      0.001      -2.865      -0.720
insured_hobbies_video-games            1.9492      0.450      4.330      0.000       1.067       2.832
insured_relationship_not-in-family     1.1859      0.332      3.572      0.000       0.535       1.837
insured_relationship_own-child        -0.3911      0.346     -1.131      0.258      -1.069       0.287
insured_relationship_unmarried         0.5995      0.343      1.748      0.080      -0.073       1.272
incident_type_Vehicle Theft           -0.4552      0.738     -0.617      0.537      -1.902       0.991
collision_type_Side Collision         -1.0720      0.280     -3.835      0.000      -1.620      -0.524
collision_type_Unknown                 0.4954      0.629      0.787      0.431      -0.738       1.728
incident_severity_Minor Damage        -5.4462      0.448    -12.143      0.000      -6.325      -4.567
incident_severity_Total Loss          -4.3645      0.358    -12.175      0.000      -5.067      -3.662
incident_severity_Trivial Damage      -5.8056      0.863     -6.731      0.000      -7.496      -4.115
incident_state_NY                     -0.6497      0.301     -2.157      0.031      -1.240      -0.059
incident_state_OH                      0.8415      0.756      1.112      0.266      -0.641       2.324
incident_state_PA                     -1.2445      0.855     -1.455      0.146      -2.920       0.431
incident_state_WV                     -1.1661      0.326     -3.579      0.000      -1.805      -0.528
incident_city_Northbrook              -1.0728      0.459     -2.339      0.019      -1.972      -0.174
property_damage_Unknown                0.9624      0.289      3.330      0.001       0.396       1.529
property_damage_YES                    0.8091      0.311      2.600      0.009       0.199       1.419
auto_make_Audi                         0.9949      0.557      1.786      0.074      -0.097       2.087
auto_make_BMW                          2.1325      0.712      2.995      0.003       0.737       3.528
auto_make_Chevrolet                   -1.6916      0.662     -2.554      0.011      -2.990      -0.394
auto_make_Nissan                      -0.9644      0.690     -1.398      0.162      -2.316       0.388
auto_model_A5                          0.3517      0.726      0.485      0.628      -1.071       1.774
auto_model_Camry                      -1.3952      0.909     -1.536      0.125      -3.176       0.386
auto_model_Civic                       2.5883      0.722      3.585      0.000       1.173       4.003
auto_model_F150                        0.8345      0.693      1.205      0.228      -0.523       2.192
auto_model_Fusion                     -1.6368      0.833     -1.965      0.049      -3.269      -0.005
auto_model_Grand Cherokee              1.3517      0.681      1.984      0.047       0.017       2.687
auto_model_Legacy                     -2.1517      0.943     -2.282      0.022      -3.999      -0.304
auto_model_MDX                        -1.6200      0.612     -2.648      0.008      -2.819      -0.421
auto_model_Other                      -1.4888      0.431     -3.458      0.001      -2.333      -0.645
auto_model_Pathfinder                 -3.7078      1.187     -3.124      0.002      -6.034      -1.382
auto_model_Silverado                   2.1133      0.927      2.280      0.023       0.297       3.930
auto_model_Ultima                      1.2015      0.974      1.233      0.218      -0.708       3.111
auto_model_Wrangler                   -1.2062      0.705     -1.711      0.087      -2.588       0.175
```

```
VIF values for detecting multicollinearity:
                                 Feature        VIF
0                                  const  13.036403
23                collision_type_Unknown   3.132409
37                      auto_make_Nissan   2.879380
26      incident_severity_Trivial Damage   2.299748
34                        auto_make_Audi   2.282747
35                         auto_make_BMW   2.282130
38                         auto_model_A5   2.257321
51                         auto_model_X5   2.059363
49                     auto_model_Ultima   1.924383
36                   auto_make_Chevrolet   1.882268
47                 auto_model_Pathfinder   1.871605
21           incident_type_Vehicle Theft   1.870257
48                  auto_model_Silverado   1.772085
24        incident_severity_Minor Damage   1.692725
46                      auto_model_Other   1.644159
32               property_damage_Unknown   1.562649
33                   property_damage_YES   1.549599
25          incident_severity_Total Loss   1.388979
11                 insured_hobbies_chess   1.305639
30                     incident_state_WV   1.300934
41                      auto_model_F150   1.277417
27                     incident_state_NY   1.270492
29                     incident_state_PA   1.255231
18    insured_relationship_not-in-family   1.251663
39                      auto_model_Camry   1.227242
20        insured_relationship_unmarried   1.223582
2             insured_education_level_JD   1.218897
28                     incident_state_OH   1.217091
19        insured_relationship_own-child   1.185366
3             insured_education_level_MD   1.180762
4            insured_education_level_PhD   1.164540
22         collision_type_Side Collision   1.153783
5     insured_occupation_exec-managerial   1.143442
12             insured_hobbies_cross-fit   1.141175
15                insured_hobbies_movies   1.133120
44                     auto_model_Legacy   1.132341
42                     auto_model_Fusion   1.130494
17           insured_hobbies_video-games   1.130119
45                        auto_model_MDX   1.129547
14                  insured_hobbies_golf   1.125835
40                      auto_model_Civic   1.117335
6     insured_occupation_farming-fishing   1.114497
50                   auto_model_Wrangler   1.109286
7   insured_occupation_handlers-cleaners   1.109146
1                     policy_csl_250/500   1.107126
31              incident_city_Northbrook   1.101857
52                       age_group_Young   1.099676
43             auto_model_Grand Cherokee   1.095855
13               insured_hobbies_dancing   1.095049
16              insured_hobbies_sleeping   1.094537
9     insured_occupation_priv-house-serv   1.092637
10               insured_hobbies_camping   1.080957
8       insured_occupation_other-service   1.079876

Features with high multicollinearity (VIF > 5):
  Feature        VIF
0   const  13.036403
```

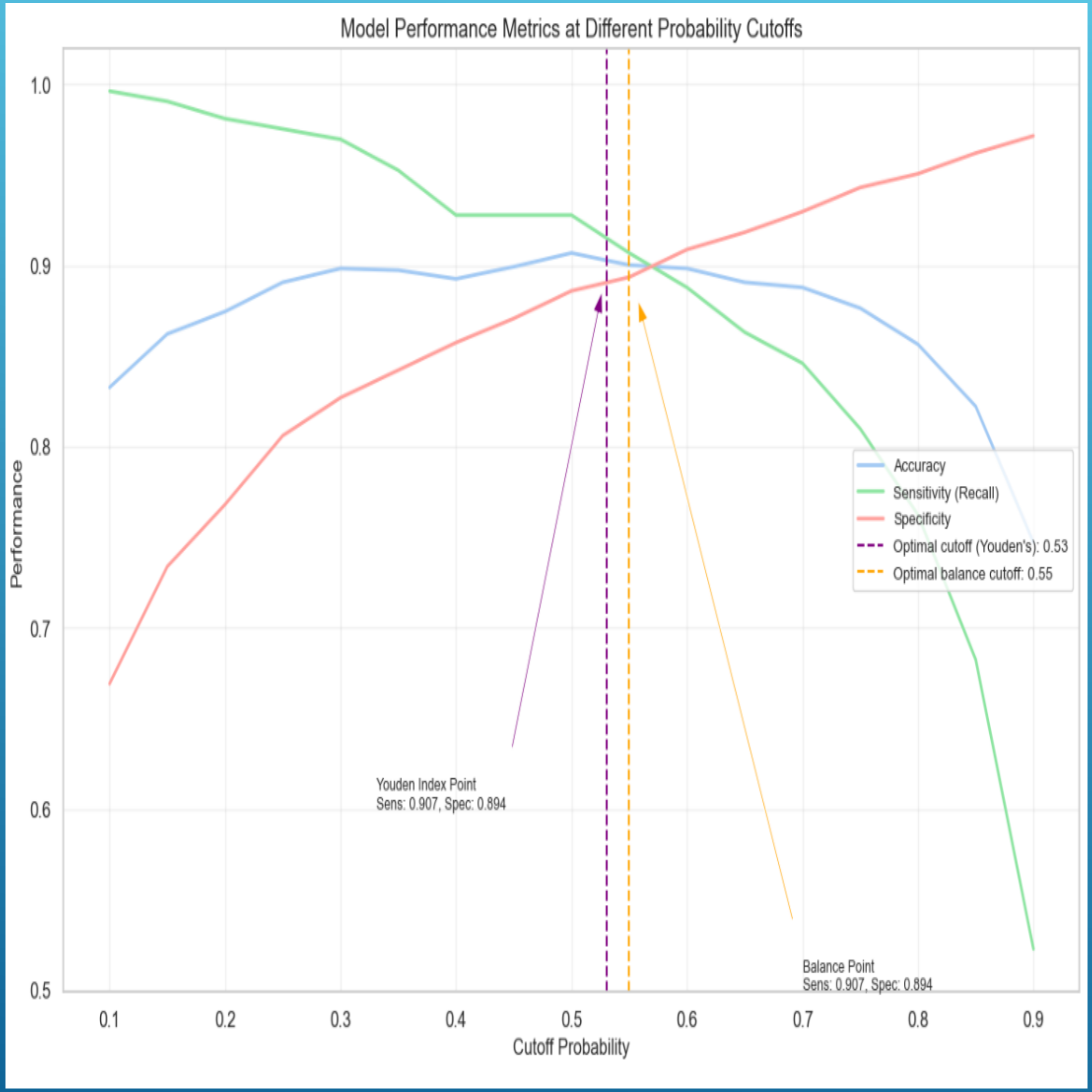
LOGISTIC REGRESSION – ROC CURVE



Optimal threshold based on Youden's index: 0.5309
At this threshold - Sensitivity: 0.9240, Specificity: 0.8897

Optimal cutoff value: 0.5309

LOGISTIC REGRESSION – ROC CURVE

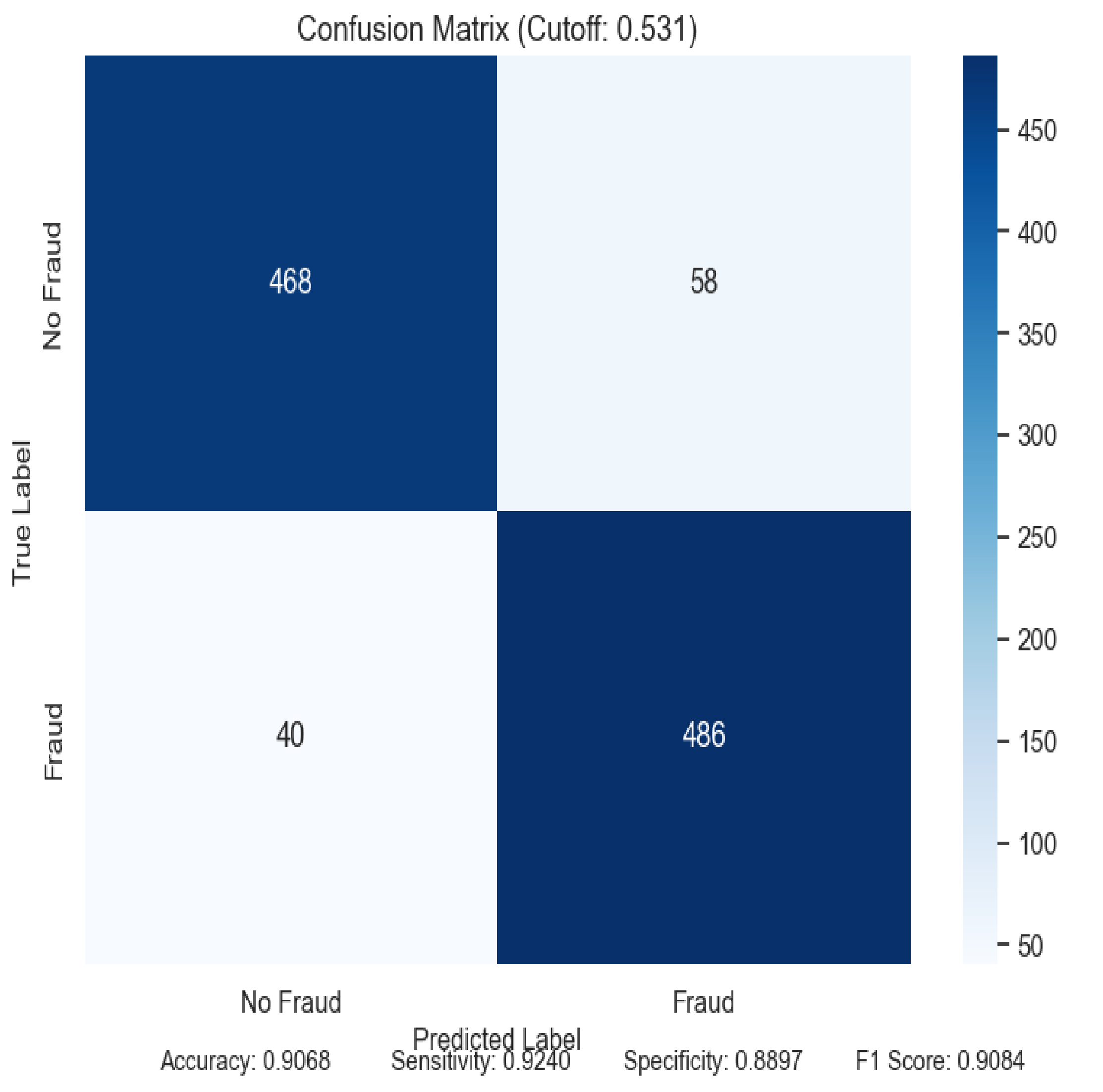


Optimal cutoff where sensitivity and specificity are closest: 0.5500

At this cutoff - Sensitivity: 0.9068, Specificity: 0.8935

Accuracy at this cutoff: 0.9002

LOGISTIC REGRESSION – CONFUSION ATRIX



Confusion Matrix using optimal cutoff:
[[468 58]
 [40 486]]

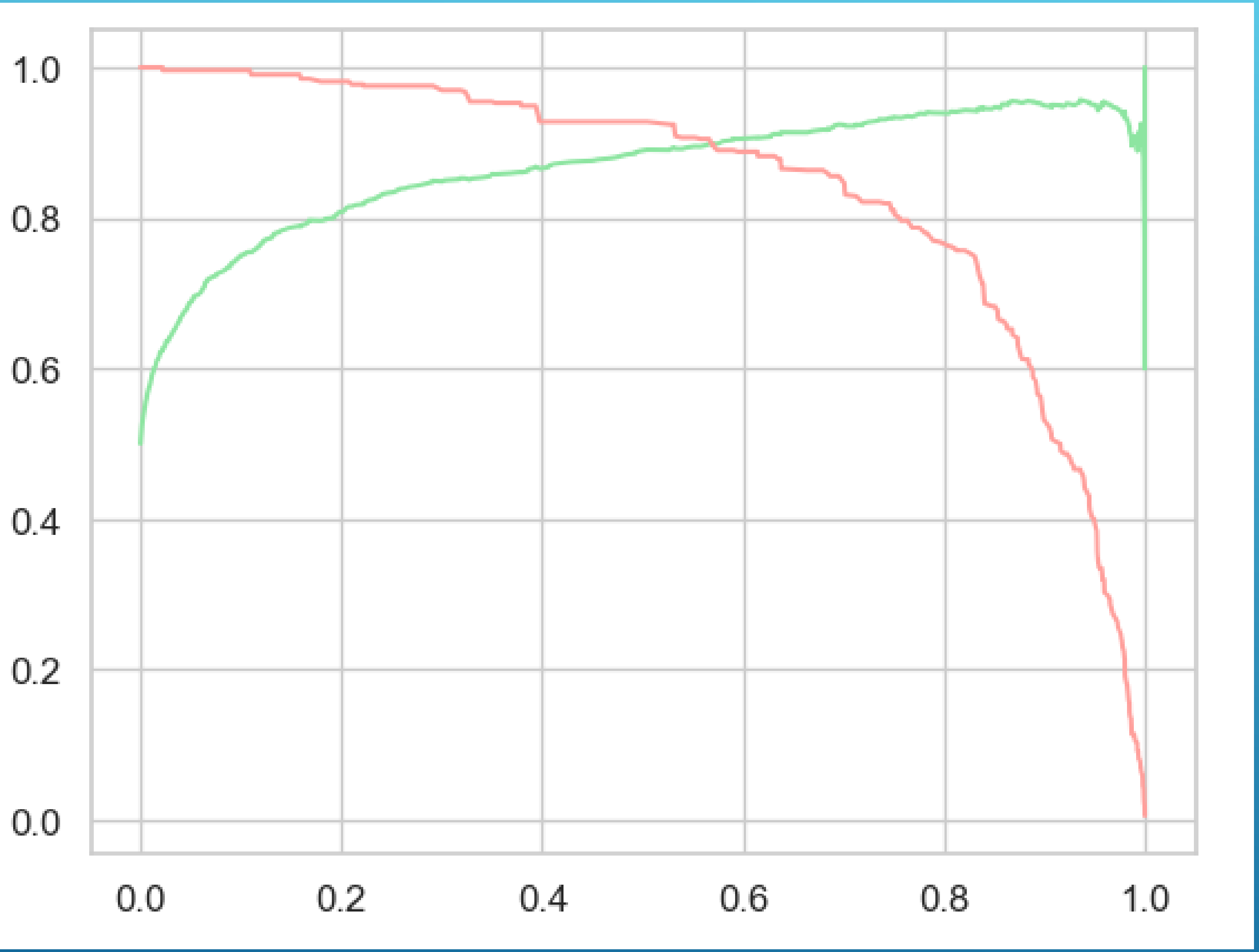
Model performance metrics using optimal cutoff (0.5309):
Accuracy: 0.9068

Sensitivity (True Positive Rate): 0.9240

Specificity (True Negative Rate): 0.8897

Precision: 0.8934
Recall: 0.9240
F1 Score: 0.9084

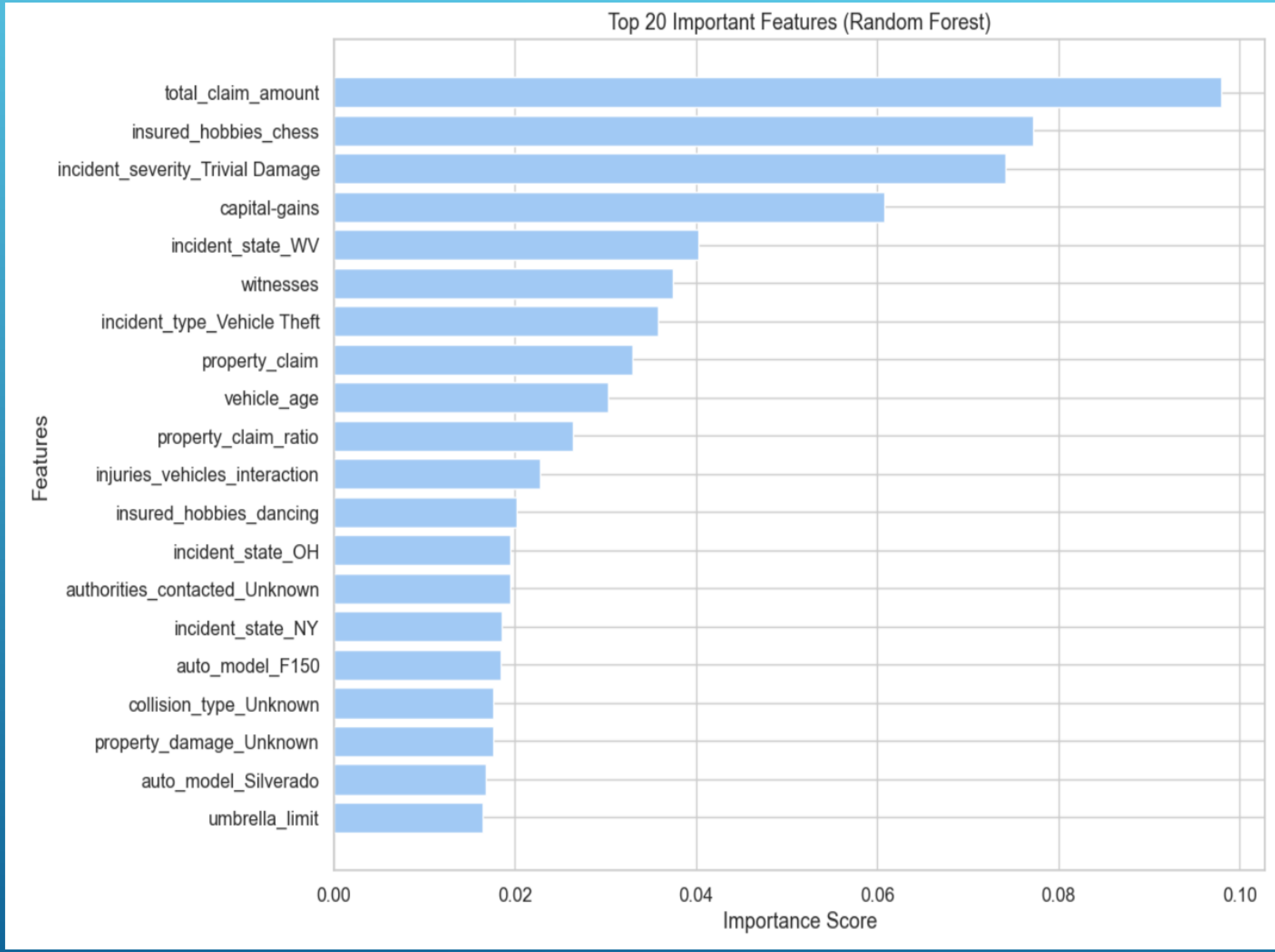
LOGISTIC REGRESSION – PRECISION – RECALL CURVE



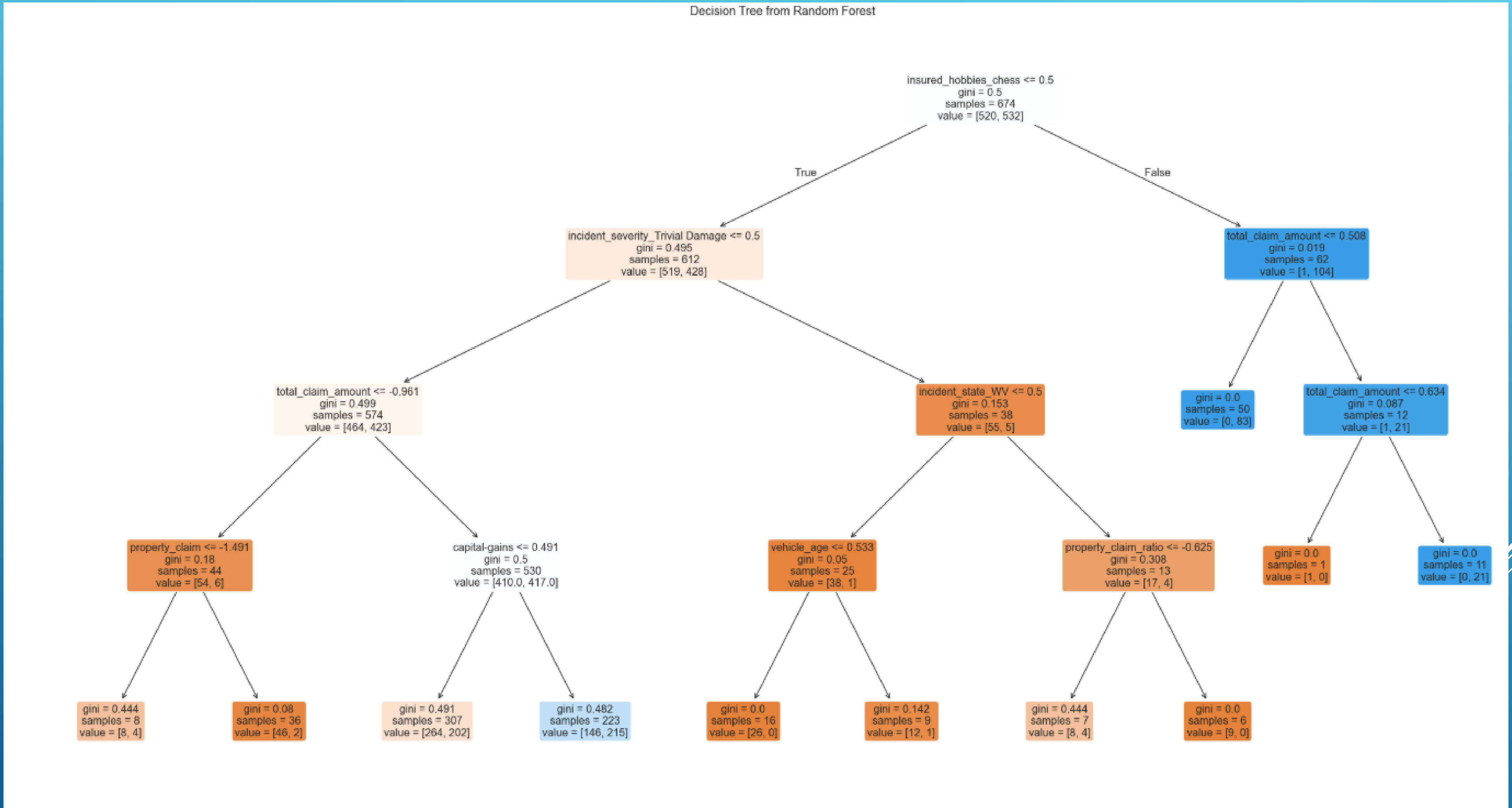
RANDOM FOREST

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']

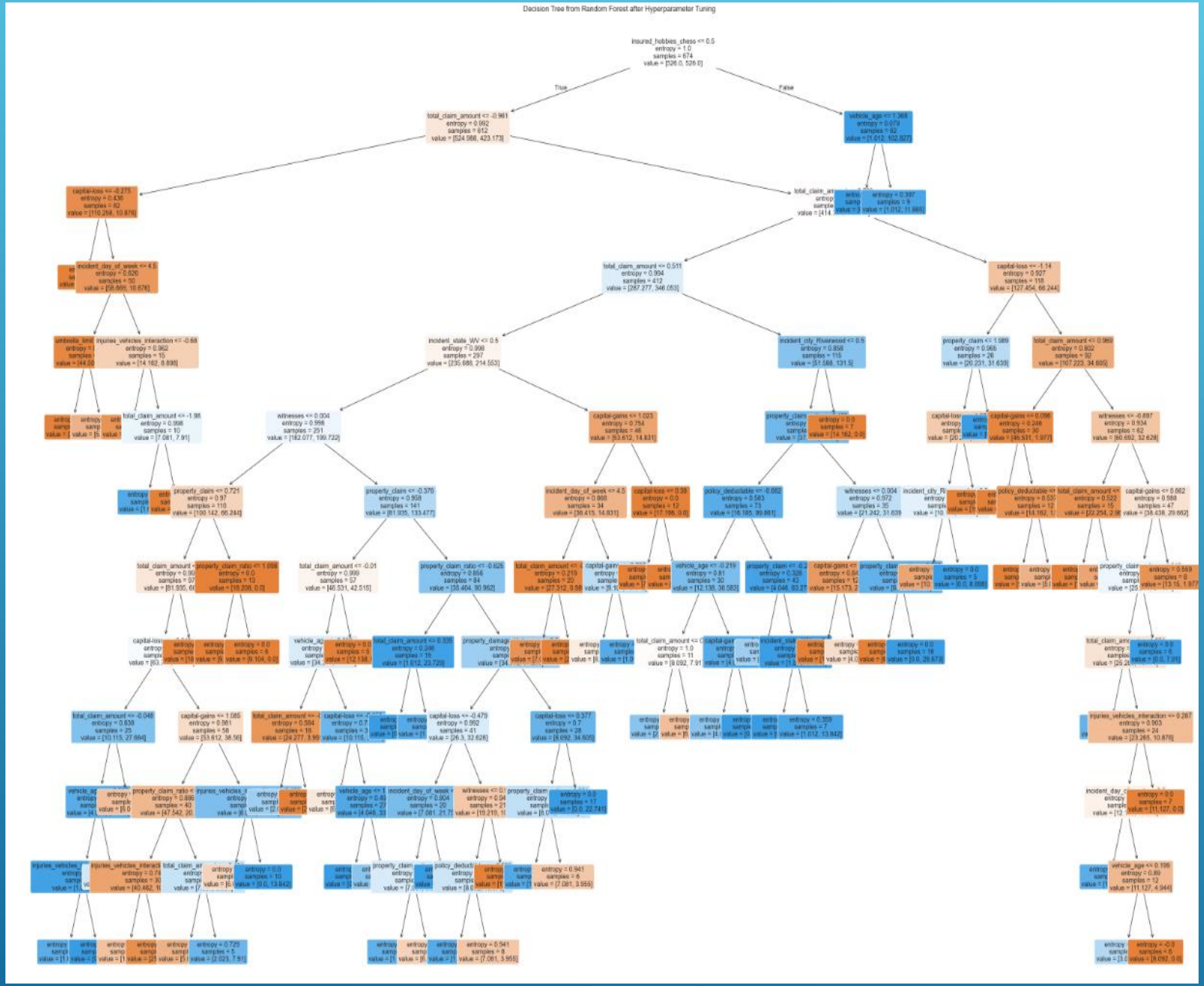
RANDOM FOREST – FEATURE IMPORTANCE



RANDOM FOREST – DECISION TREE BASED ON FEATURE

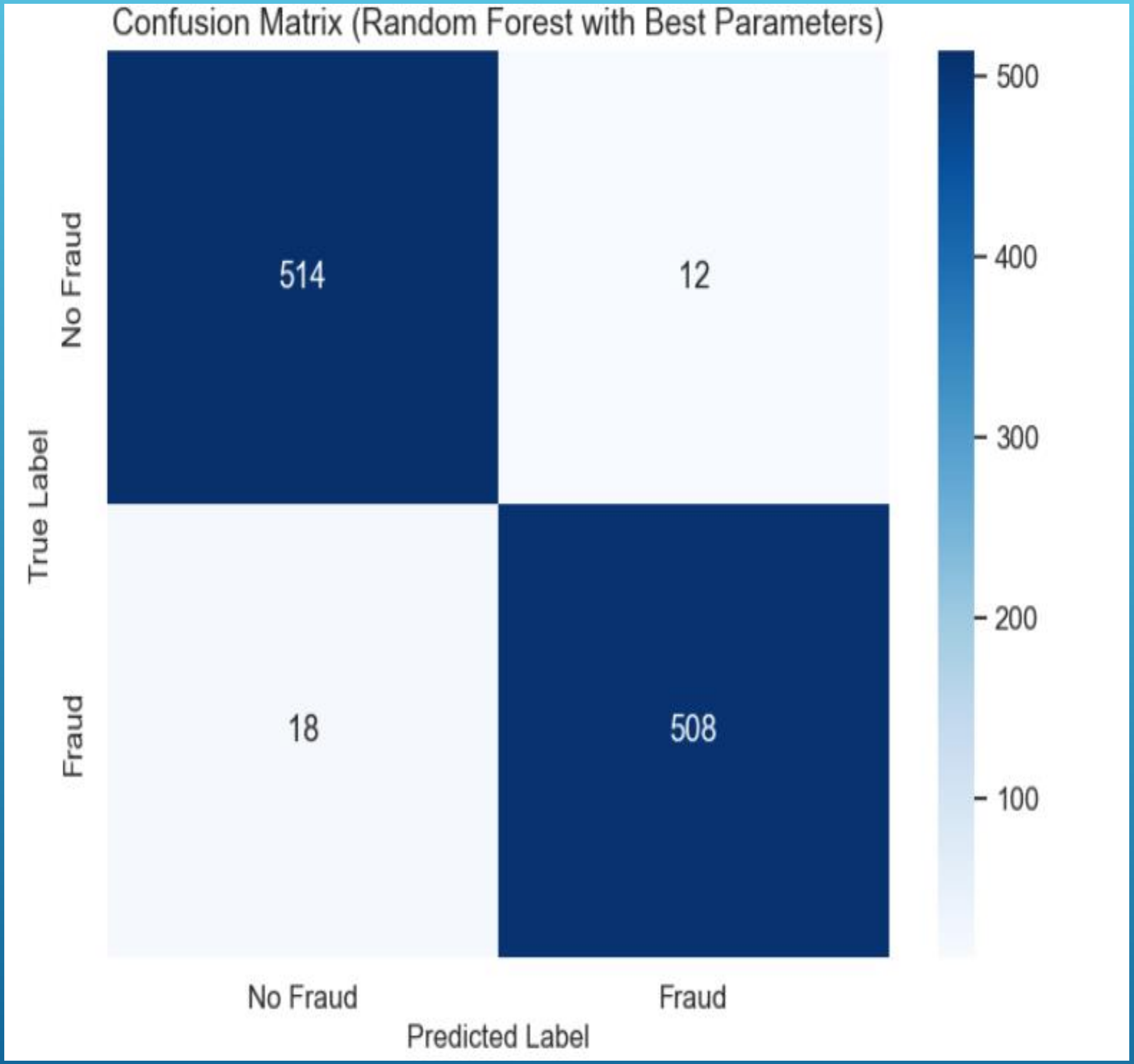


RANDOM FOREST – HYPERPARAMETER TUNING



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 AND MODEL EVALUATION

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

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)

QUESTIONS

1. What methods can be used to analyze historical insurance claims data for identifying potential fraud patterns?

- EDA to uncover variable relationships and trends linked to fraud.
- Feature engineering to create derived metrics like claim-to-premium ratios and claim frequency.
- Outlier detection using statistical or unsupervised methods to flag anomalies.
- Predictive modeling (e.g., logistic regression, random forest) to detect complex fraud patterns.
- ROC curve analysis to find optimal fraud detection thresholds.
- Model evaluation using sensitivity, specificity, and precision-recall metrics to handle class imbalance effectively.

QUESTIONS

2. Which factors most strongly indicate potential fraudulent behavior in insurance claims?

- Total Claim Amount – Higher claim amounts are often associated with increased fraud risk.
- Customer Demographics – Certain hobbies (e.g., chess, dancing) may correlate with specific fraud patterns.
- Incident Severity – Minor or trivial damage claims show a higher likelihood of being fraudulent.
- Capital Gains/Losses – Claimants with notable financial fluctuations may present higher fraud risks.
- Geographic Location – States like West Virginia (WV), New York (NY), and Ohio (OH) exhibit elevated fraud rates.
- Vehicle Type – Models such as the Ford F-150 and Chevy Silverado are more frequently involved in suspicious claims.
- Incident Type – Claims involving vehicle theft or unspecified collision types tend to raise red flags.
- Property Damage Reporting – Delayed or inconsistent property damage reporting can signal fraudulent intent.

QUESTIONS

3. Is it possible to predict the likelihood of fraud in new insurance claims using historical data?

- Logistic Regression Performance – Achieved 80% validation accuracy with a sensitivity of 67.57%, indicating strong performance in identifying actual fraud cases.
- Probability Threshold Optimization – An optimal cutoff of approximately 0.55 was identified to balance false positives and false negatives.
- Fraud Probability Scores – The model generates a probability score for each claim, indicating the likelihood of fraud.
- Random Forest Benchmark – Offers an alternative model with slightly lower sensitivity (33.78%) but useful for comparison and ensemble strategies.
- Deployment Capability – These models can be integrated into claim processing systems to automatically score and flag suspicious claims in real time.

QUESTIONS

3. Is it possible to predict the likelihood of fraud in new insurance claims using historical data?

- Logistic Regression Performance – Achieved 80% validation accuracy with a sensitivity of 67.57%, indicating strong performance in identifying actual fraud cases.
- Probability Threshold Optimization – An optimal cutoff of approximately 0.55 was identified to balance false positives and false negatives.
- Fraud Probability Scores – The model generates a probability score for each claim, indicating the likelihood of fraud.
- Random Forest Benchmark – Offers an alternative model with slightly lower sensitivity (33.78%) but useful for comparison and ensemble strategies.
- Deployment Capability – These models can be integrated into claim processing systems to automatically score and flag suspicious claims in real time.

QUESTIONS

4. What actionable insights from the model can enhance the fraud detection strategy?

- Optimize Probability Thresholds – A fixed 0.5 threshold is suboptimal for imbalanced datasets; tuning cutoffs improves detection without overwhelming false positives.
- Prioritize High-Risk Claim Attributes – Claims involving minor damage or specific vehicle models should receive heightened scrutiny.
- Leverage Geographic Trends – Certain states consistently show higher fraud risk, suggesting the need for regional fraud flags.
- Balance Sensitivity and Customer Experience – Cutoff adjustments can strike a balance between catching fraud and minimizing disruption for genuine claimants.
- Implement Tiered Reviews – Use model-generated probability scores to route claims into different levels of manual or automated review.
- Model Selection Matters – Logistic regression with optimized thresholds outperforms more complex models in terms of practical fraud detection effectiveness.
- Use Demographics with Caution – Patterns in hobbies or occupations should be considered carefully to avoid bias or unfair profiling.