FRAUDULENT CLAIM DETECTION

Case Study Report

SAGNIK SAHA
OMKAR TERDAL

Contents

Problem Statement

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The companys current process for identifying fraudulent claims involves manual inspections, which are time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process. This would minimize financial losses and optimize the overall claims handling process.

Business Objective

Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.

1. Data Loading

The insurance claims data has 40 Columns and 1000 Rows. The following data dictionary provides the description for each column present in dataset:

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 40 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   age                          1000 non-null   int64
 2   policy_number                1000 non-null   int64
 3   policy_bind_date             1000 non-null   object
 4   policy_state                 1000 non-null   object
 5   policy_csl                   1000 non-null   object
 6   policy_deductable            1000 non-null   int64
 7   policy_annual_premium        1000 non-null   float64
 8   umbrella_limit               1000 non-null   int64
 9   insured_zip                  1000 non-null   int64
 10  insured_sex                  1000 non-null   object
 11  insured_education_level      1000 non-null   object
 12  insured_occupation           1000 non-null   object
 13  insured_hobbies              1000 non-null   object
 14  insured_relationship         1000 non-null   object
 15  capital-gains                1000 non-null   int64
 16  capital-loss                 1000 non-null   int64
 17  incident_date                1000 non-null   object
 18  incident_type                1000 non-null   object
 19  collision_type               1000 non-null   object
 20  incident_severity            1000 non-null   object
 21  authorities_contacted        909 non-null    object
 22  incident_state               1000 non-null   object
 23  incident_city                1000 non-null   object
 24  incident_location            1000 non-null   object
 25  incident_hour_of_the_day     1000 non-null   int64
 26  number_of_vehicles_involved  1000 non-null   int64
 27  property_damage              1000 non-null   object
 28  bodily_injuries              1000 non-null   int64
 29  witnesses                    1000 non-null   int64
 30  police_report_available      1000 non-null   object
 31  total_claim_amount           1000 non-null   int64
 32  injury_claim                 1000 non-null   int64
 33  property_claim               1000 non-null   int64
 34  vehicle_claim                1000 non-null   int64
 35  auto_make                    1000 non-null   object
 36  auto_model                   1000 non-null   object
 37  auto_year                    1000 non-null   int64
 38  fraud_reported               1000 non-null   object
 39  _c39                         0 non-null      float64
```

2.Data Preparation & Cleaning

- **Null values are present for 2 feature authorities_contacted & _c39.**

```
collision_type                 0.0
incident_severity              0.0
authorities_contacted          9.1
incident_state                 0.0
incident_city                  0.0
incident_location              0.0
incident_hour_of_the_day       0.0
number_of_vehicles_involved    0.0
property_damage                0.0
bodily_injuries                0.0
witnesses                      0.0
police_report_available        0.0
total_claim_amount             0.0
injury_claim                   0.0
property_claim                 0.0
vehicle_claim                  0.0
auto_make                      0.0
auto_model                     0.0
auto_year                      0.0
fraud_reported                 0.0
_c39                         100.0
dtype: float64
```

- **Replace the ? by the most common collision type as we are unaware of the type.**
- **Replace NaN values in authorities_contacted column most common values.**
- **Dropped feature "_c39" which is empty.**
- **Identified and removed columns where a large proportion of the values are unique or near unique, as these columns are likely to be identifiers or have very limited predictive power**
    - **policy_number**
    - **incident_location**
    - **insured_zip**
- **Fixed the datatype for datetime features *policy_bind_date & incident_date.***

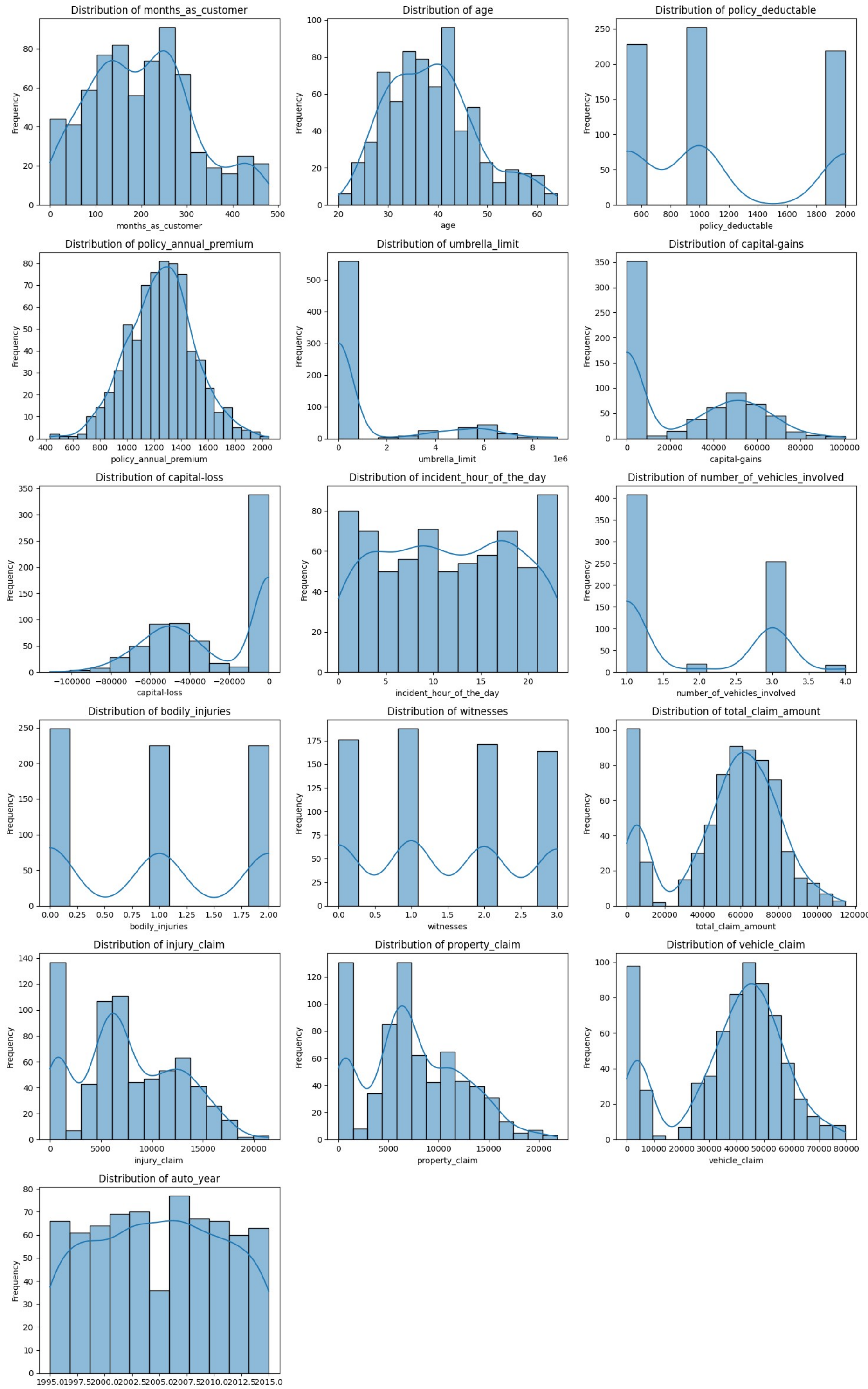
3. Train-Validation Split

- **Performed train-train split with 70% training & 30% validation data.**

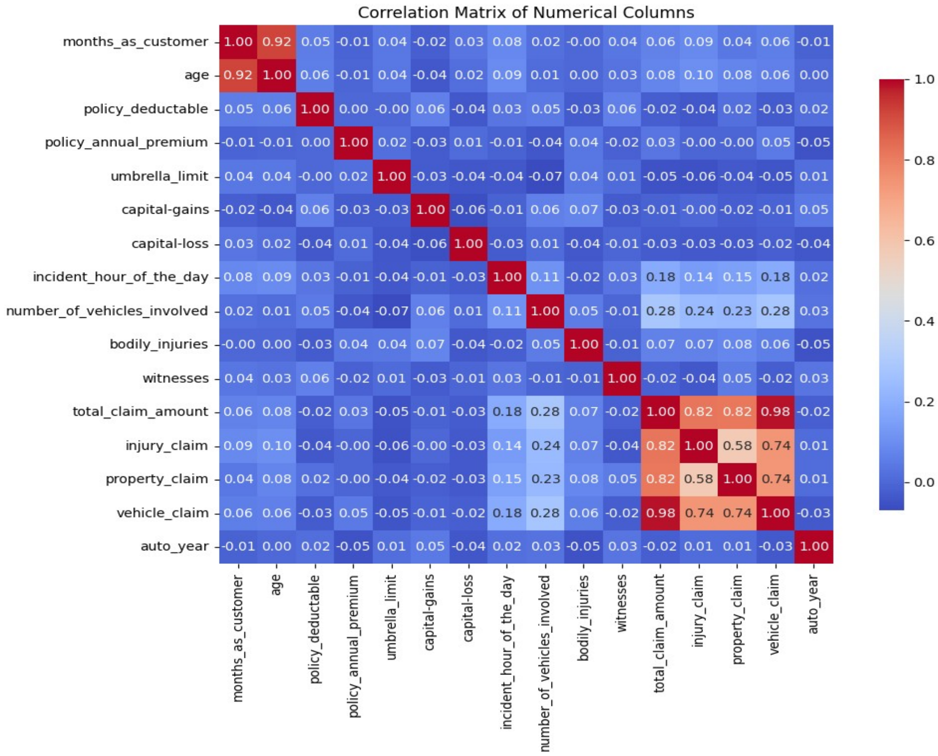
```
print(X_train.shape, X_test.shape)
print(y_train.shape, y_test.shape)
[680]  ✓ 0.0s
... (699, 35) (300, 35)
    (699,) (300,)
```

4. Exploratory Data Analysis on Training Data

4.1 Univariate analysis



4.2 Correlation analysis

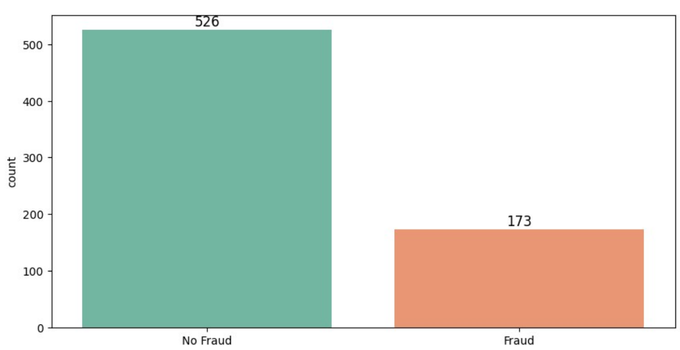


- `The top high correlated features are`

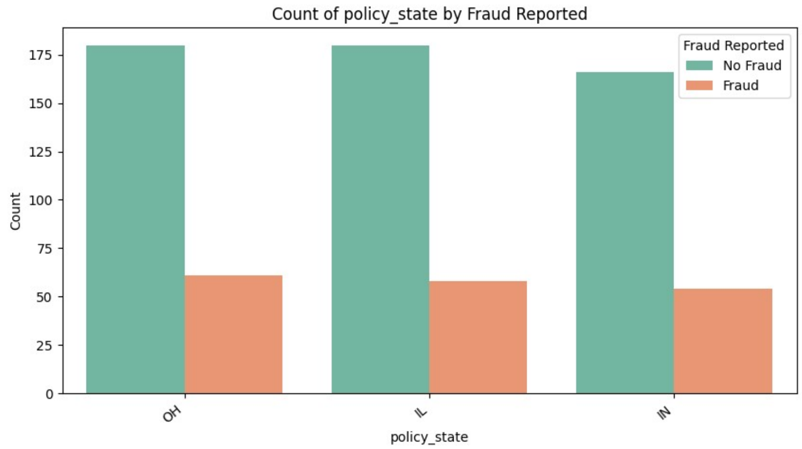
	Feature 1	Feature 2	Correlation
3	vehicle_claim	total_claim_amount	0.984208
0	age	months_as_customer	0.920167
1	injury_claim	total_claim_amount	0.818053
2	property_claim	total_claim_amount	0.815479

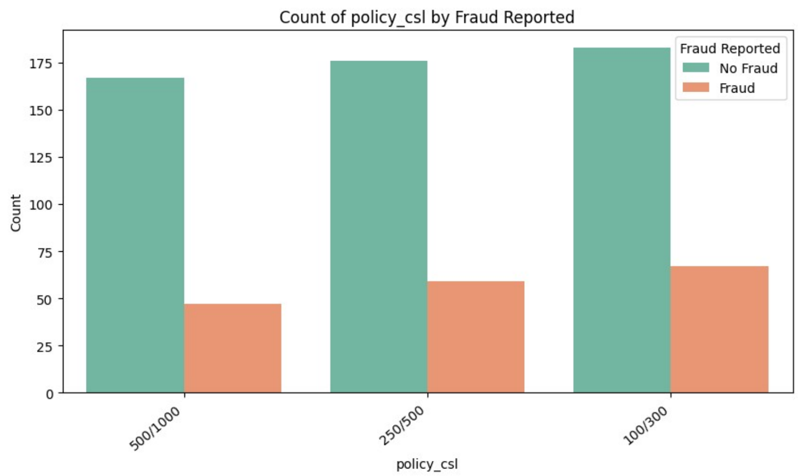
4.3 Check class balance

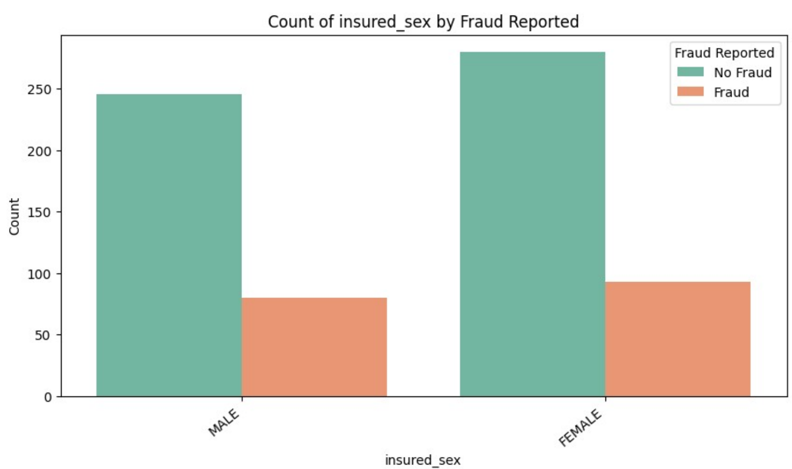
- **Distribution of the target variable to identify potential class imbalances using visualization for better understanding.**

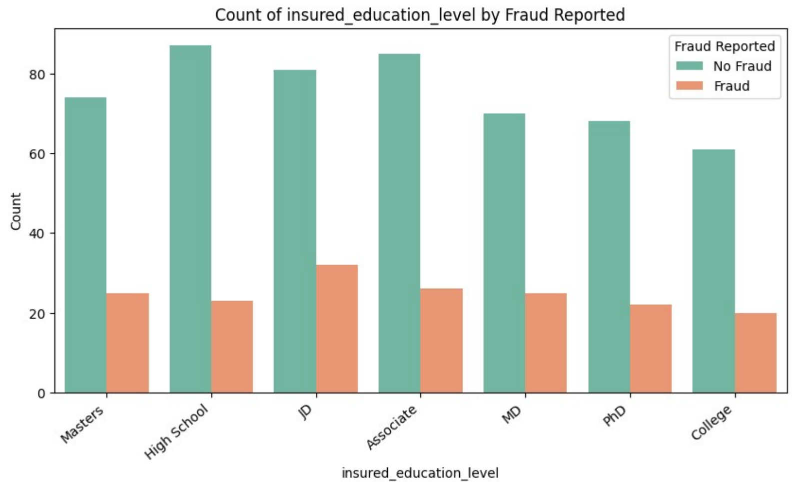


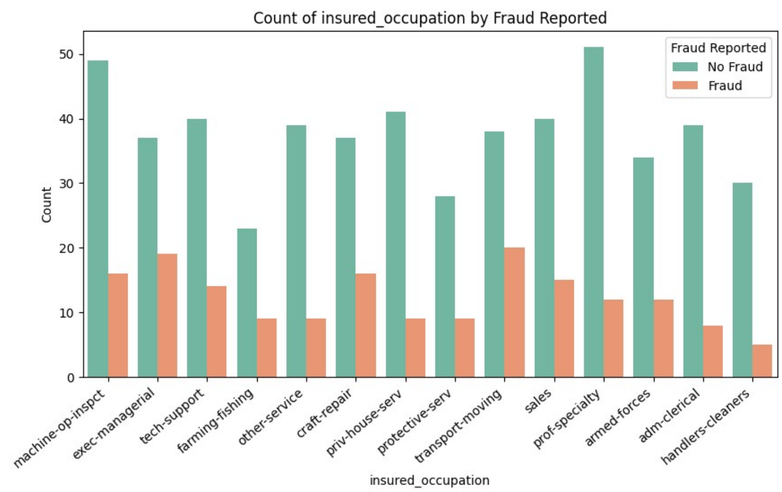
4.4 Bivariate analysis

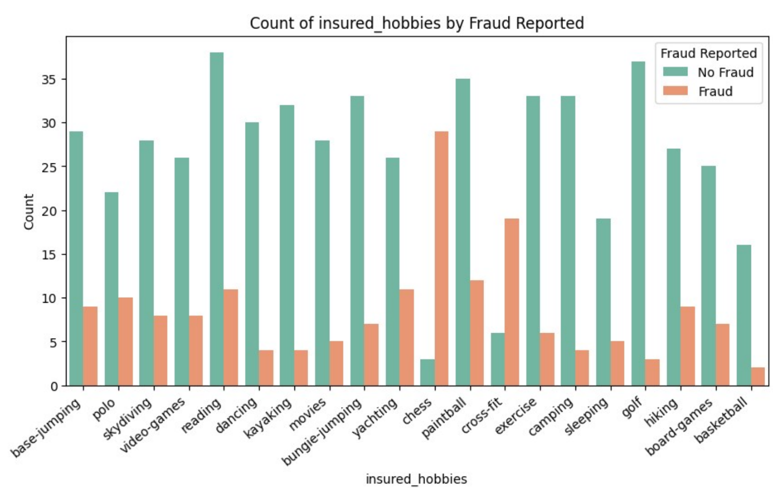


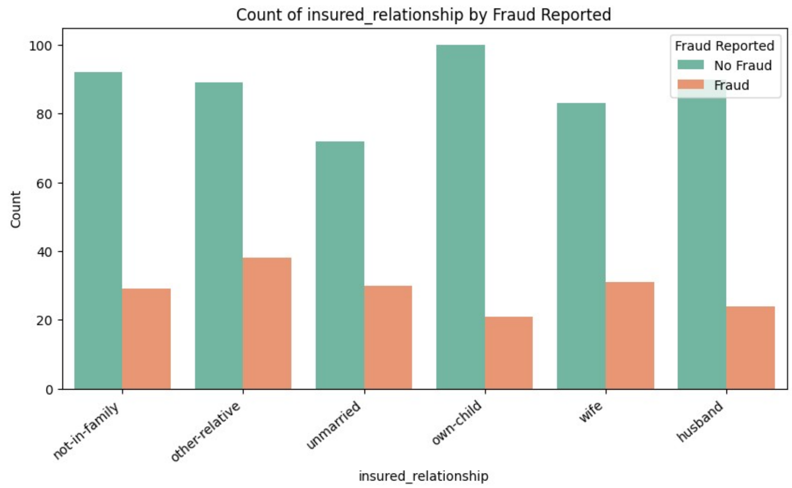


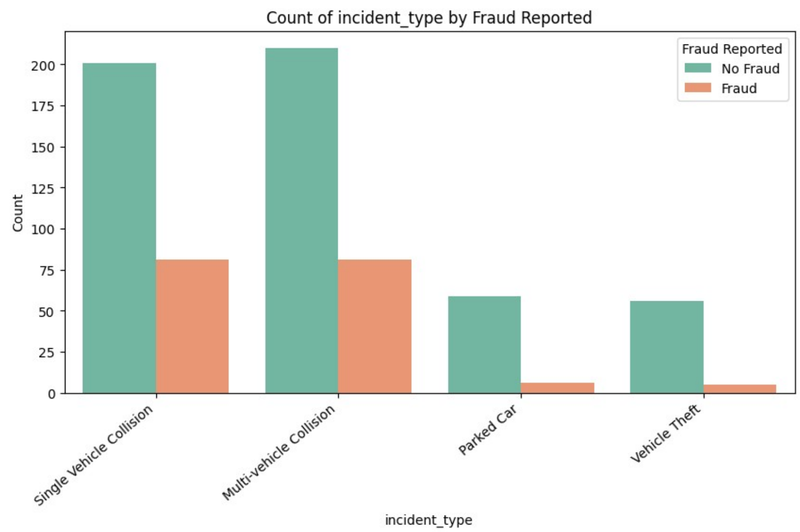


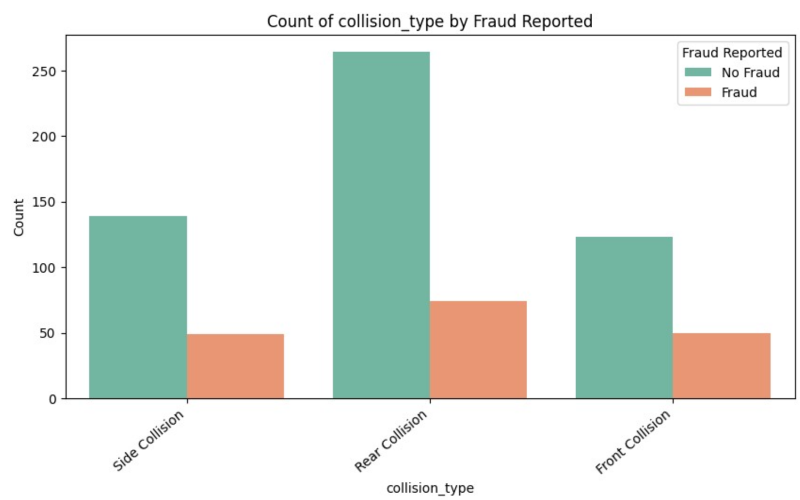

Count of insured_education_level by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
20
40
60
80
Masters
High School
JD
Associate
MD
PhD
College
insured_education_level

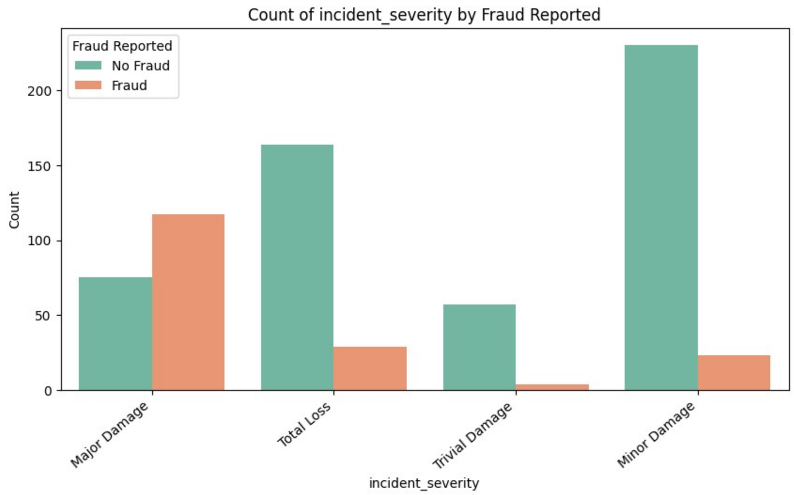

Count of insured_occupation by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
10
20
30
40
50
machine-op-inspct
exec-managerial
tech-support
farming-fishing
other-service
craft-repair
priv-house-serv
protective-serv
transport-moving
sales
prof-specialty
armed-forces
adm-clerical
handlers-cleaners
insured_occupation

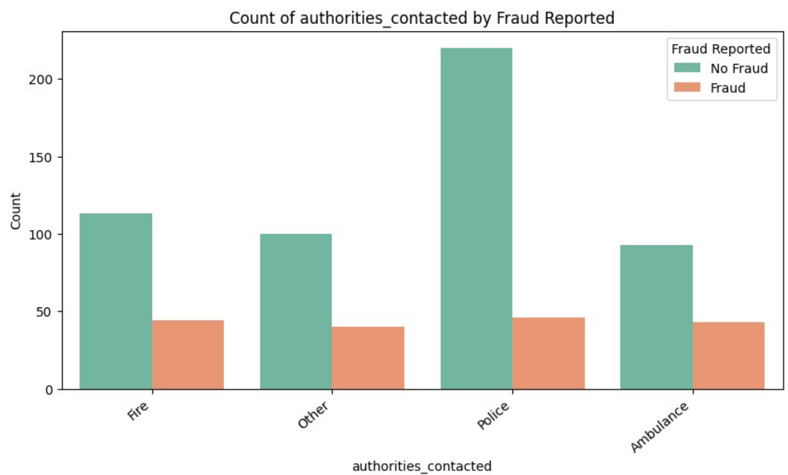

Count of insured_hobbies by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
5
10
15
20
25
30
35
base-jumping
polo
skydiving
video-games
reading
dancing
kayaking
movies
bungie-jumping
yachting
chess
paintball
cross-fit
exercise
camping
sleeping
golf
hiking
board-games
basketball
insured_hobbies

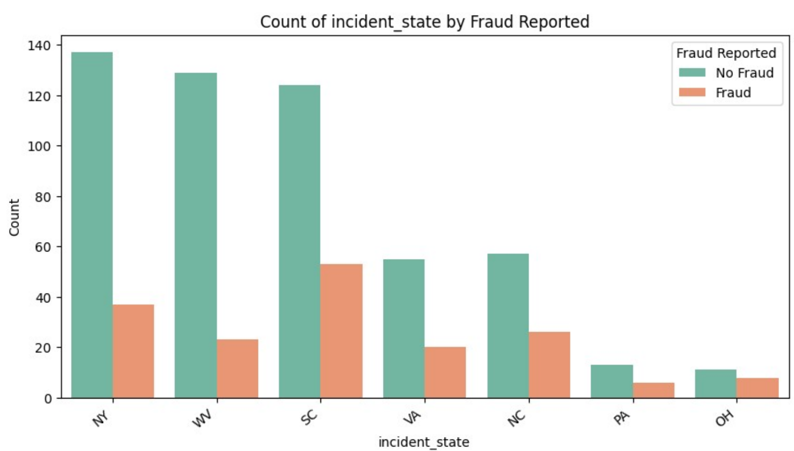

Count of insured_relationship by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
100
80
60
40
20
0
not-in-family
other-relative
unmarried
own-child
wife
husband
insured_relationship

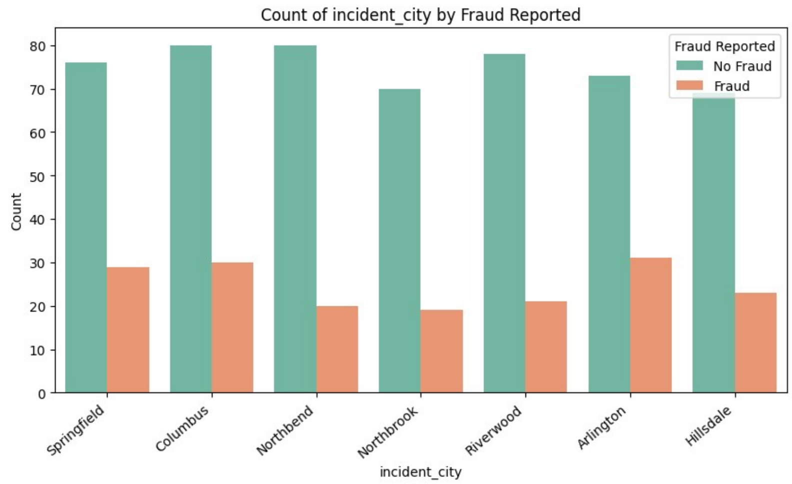

Count of incident_type by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
200
175
150
125
100
75
50
25
0
Single Vehicle Collision
Multi-vehicle Collision
Parked Car
Vehicle Theft
incident_type

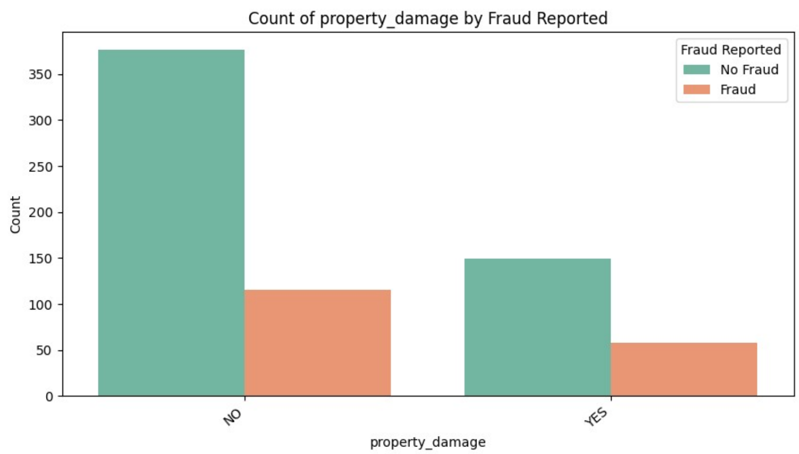

Count of collision_type by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
250
200
150
100
50
0
Side Collision
Rear Collision
Front Collision
collision_type


Count of incident_severity by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
50
100
150
200
Major Damage
Total Loss
Trivial Damage
Minor Damage
incident_severity

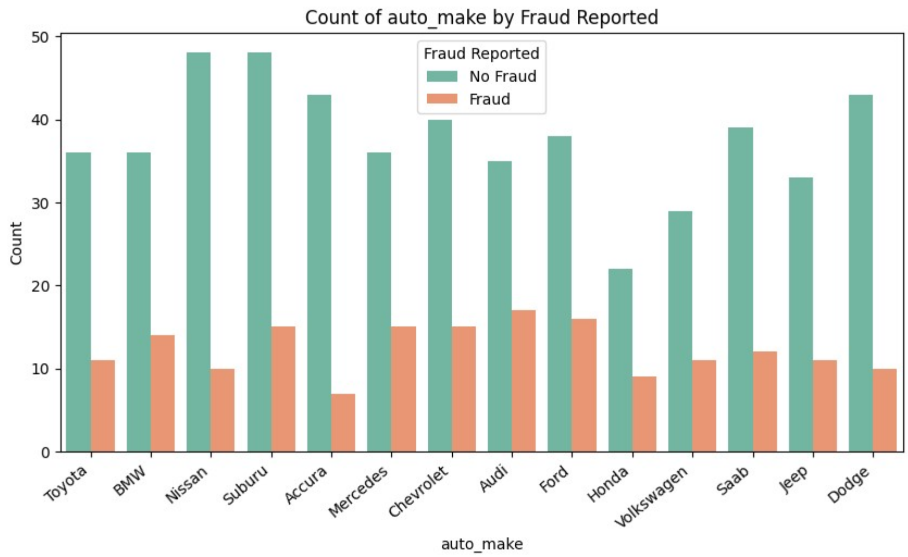

Count of authorities_contacted by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
50
100
150
200
Fire
Other
Police
Ambulance
authorities_contacted

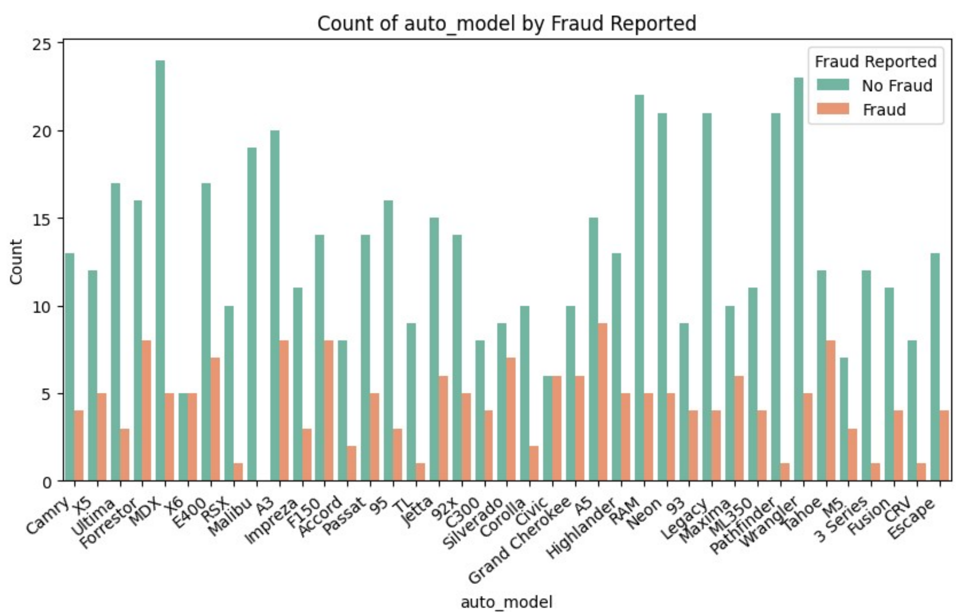

Count of incident_state by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
20
40
60
80
100
120
140
NY
WV
SC
VA
NC
PA
OH
incident_state


Count of incident_city by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
10
20
30
40
50
60
70
80
Springfield
Columbus
Northbend
Northbrook
Riverwood
Arlington
Hillsdale
incident_city

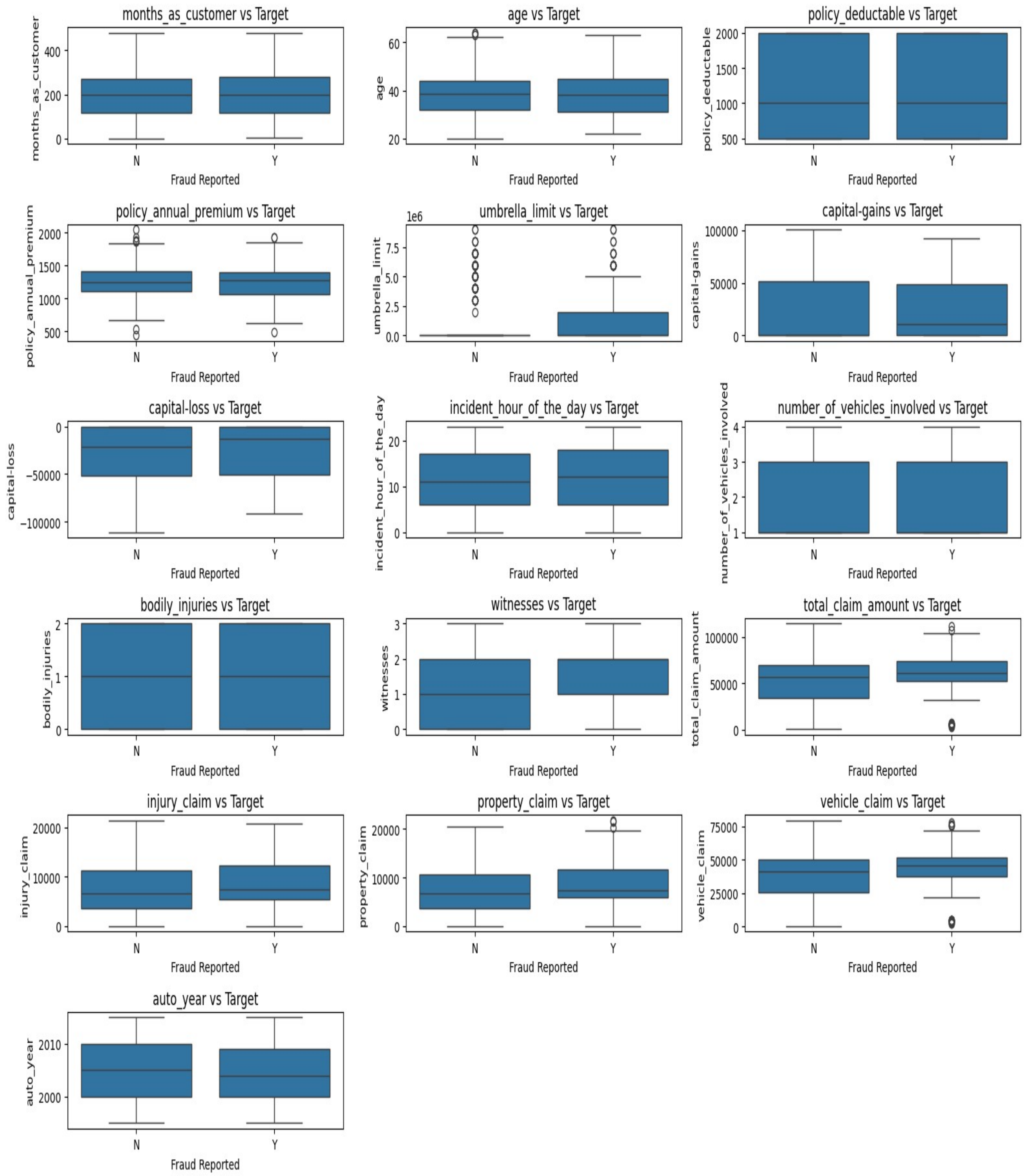

Count of property_damage by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
50
100
150
200
250
300
350
NO
YES
property_damage


Count of police_report_available by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
50
100
150
200
250
300
350
NO
YES
police_report_available


Count of auto_make by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
10
20
30
40
50
Toyota
BMW
Nissan
Suburu
Accura
Mercedes
Chevrolet
Audi
Ford
Honda
Volkswagen
Saab
Jeep
Dodge
auto_make


Count of auto_model by Fraud Reported
Fraud Reported
No Fraud
Fraud
Count
0
5
10
15
20
25
Camry
X5
Ultima
Forrestor
MDX
X6
E400
RSX
Malibu
A3
Impreza
F150
Accord
Passat
95
TL
Jetta
92x
C300
Silverado
Corolla
Civic
Grand Cherokee
A5
Highlander
RAM
Neon
93
Legacy
Maxima
ML350
Pathfinder
Wrangler
Tahoe
M5
3 Series
Fusion
CRV
Escape
auto_model

4.4.2 Relationships between numerical features and the target variable to understand their impact on the target outcome.



6. Feature Engineering

6.1 Perform resampling

- **Using RandomOverSampler technique to balance the data and handle class imbalance.**
- **This method increases the number of samples in the minority class by randomly duplicating them, creating synthetic data points with similar characteristics.**
- **This helps prevent the model from being biased toward the majority class and improves its ability to predict the minority class more accurately.**
- **Results after resampling**

```
Original training set shape: (699, 35), (699,)
Resampled training set shape: (1052, 35), (1052,)
```

6.2 Feature Creation

- **Below new features are created using the existing features**
    - **age_group – Using the age column, created age group bins.**
    - **vehicle_age - Derived the age of the vehicle based on the year value.**
    - **policy_age_years – Derived the policy age at the time of the incident using incident_date & policy_bind_date.**
    - **injury_claim_ratio, property_claim_ratio, vehicle_claim_ratio – Derived the claim ratio from the total types of claims and there claim amount.**
    - **claim_per_vehicle – Created claim per vehicle.**
    - **auto_full_name – Derived by using the auto_make & auto_model.**

6.3 Handle redundant columns

- Removed the below features which have low correlation and redundant.
    - policy_csl
    - policy_bind_date
    - incident_date
    - auto_year
    - auto_model
    - auto_make
    - incident_hour_of_the_day
    - csl_per_person

- `csl_per_accident`

6.4 Combine values in Categorical Columns

- **From EDA found that for insured_hobbies feature vs target variable, chess and cross-fit are among the top. So apart from these 2 values, replaced all values in other categories.**

6.5 Dummy variable creation

- **Transform categorical variables into numerical representations using dummy variables. Ensure consistent encoding between training and validation data.**
- **The shape for both training and validation data after dummy variable creation**

Train	**(1052, 109)**
Test	**(300, 109)**

6.6 Feature scaling

- **Scale numerical features to a common range to prevent features with larger values from dominating the model. Choose a scaling method appropriate for the data and the chosen model. Apply the same scaling to both training and validation data.**

7. Model Building

7.1 Logistic Regression

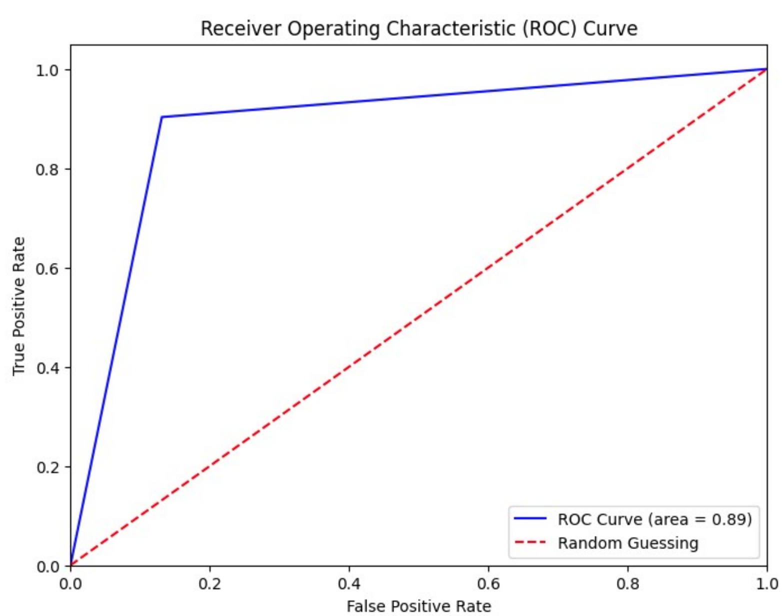
Top features based on RPECV

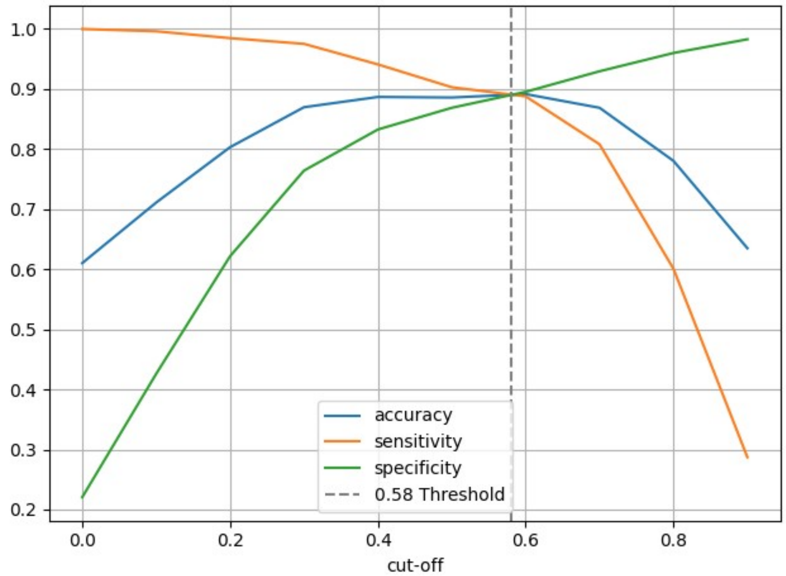
- **insured_occupation_exec-managerial,**
- **insured_occupation_handlers-cleaners,**
- **insured_occupation_other-service,**
- **insured_occupation_priv-house-serv,**
- **insured_education_level_PhD,**
- **insured_education_level_MD,**
- **insured_education_level_JD,**
- **insured_occupation_craft-repair,**
- **insured_relationship_unmarried,**

- **collision_type_Side Collision,**
- **insured_relationship_not-in-family,**
- **insured_hobbies_chess,**
- **incident_state_WV,**
- **incident_city_Columbus,**
- **incident_city_Northbrook,**
- **incident_severity_Minor Damage,**
- **incident_severity_Total Loss,**
- **incident_severity_Trivial Damage,**
- **incident_state_OH,**
- **incident_state_NY,**
- **incident_city_Northbend,**
- **incident_state_VA,**
- **incident_type_Vehicle Theft,**
- **insured_relationship_own-child,**
- **insured_hobbies_cross-fit,**
- **insured_occupation_transport-moving,**
- **insured_occupation_protective-serv,**
- **auto_full_name_Accura-RSX,**
- **age_group_60+,**
- **property_damage_YES,**
- **incident_city_Riverwood,**
- **auto_full_name_Chevrolet-Silverado,**
- **auto_full_name_Ford-Fusion,**
- **auto_full_name_Ford-F150,**
- **auto_full_name_Ford-Escape,**
- **auto_full_name_Jeep-Wrangler,**
- **auto_full_name_Chevrolet-Malibu,**
- **auto_full_name_BMW-X6,**

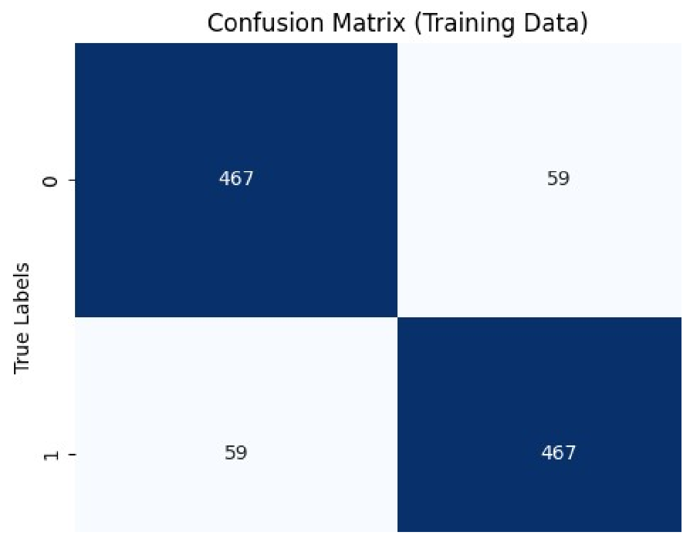
- **auto_full_name_Toyota-Camry,**
- **auto_full_name_Nissan-Pathfinder,**
- **auto_full_name_Saab-92x,**
- **auto_full_name_Suburu-Legacy,**
- **auto_full_name_Jeep-Grand Cherokee,**
- **auto_full_name_Honda-CRV,**
- **auto_full_name_Honda-Civic,**
- **auto_full_name_Mercedes-ML350,**
- **auto_full_name_Mercedes-C300,**
- **auto_full_name_Accura-TL,**
- **auto_full_name_Audi-A3,**
- **auto_full_name_Audi-A5,**
- **auto_full_name_BMW-3 Series,**
- **auto_full_name_BMW-M5ROC Curve**
- **witnesses,**

Trade-off plot between accuracy, sensitivity and specificity

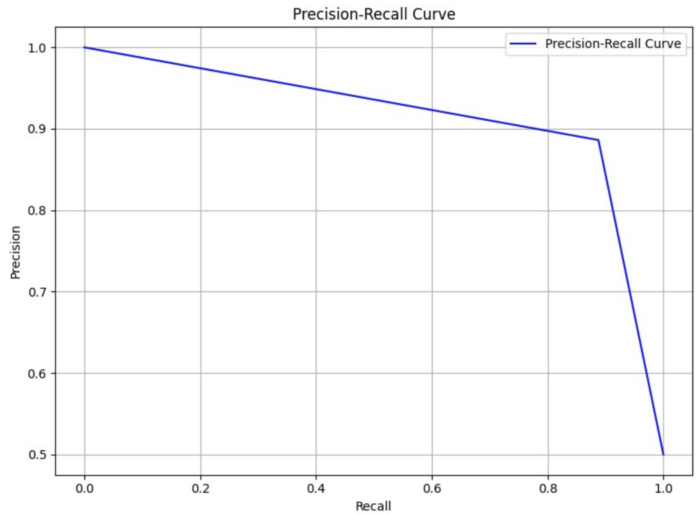




- **Summary of Model (Training)**
  - **Probability Cutoff: 0.58**
  - **Model Accuracy: 0.89**
  - **Sensitivity (Recall): 0.9**
  - **Specificity: 0.87**
  - **Precision: 0.87**
  - **F1 Score: 0.89**
- **Confurion Matrix**



- **precision-recall curve**



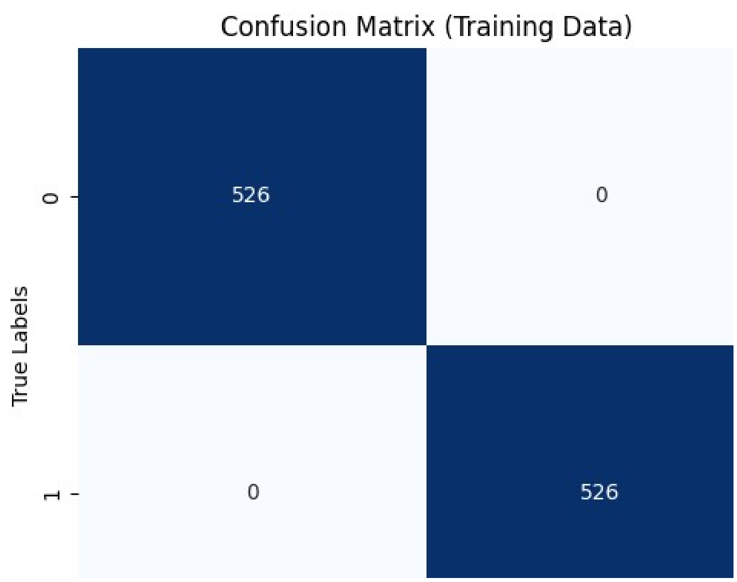
7.2 Random Forest Model

- **Top features with importance score**

```
Selected features for Random Forest:
48  incident_severity_Minor Damage    0.089925
13               claim_per_vehicle    0.060661
36            insured_hobbies_chess    0.058664
49    incident_severity_Total Loss    0.055817
0               months_as_customer    0.042461
2             policy_annual_premium    0.042267
8                      vehicle_age    0.033206
9                 policy_age_years    0.032261
10              injury_claim_ratio    0.029883
12             vehicle_claim_ratio    0.028098
3                    capital-gains    0.027516
4                     capital-loss    0.025487
11            property_claim_ratio    0.022038
```

- **Best estimator found**
  - **max_depth=15,**
  - **max_features=5,**
  - **min_samples_leaf=10,**
  - **min_samples_split=20,**
  - **n_estimators=15**

- **Summary of Model (Training)**
  - **OOB Score: 0.85**
  - **Model Accuracy: 0.89**
  - **Sensitivity (Recall): 0.91**
  - **Specificity: 0.87**
  - **Precision: 0.88**
  - **F1 Score: 0.89**
- **Confurion Matrix**



8. Prediction and Model Evaluation

Model	Hyperparameter Tuning	Training data performance	Test data performance
Logistic Regression	Probability Cutoff: 0.58	Model Accuracy: 0.89 Sensitivity (Recall): 0.9 Specificity: 0.87 Precision: 0.87 F1 Score: 0.89	Model Accuracy: 0.84 Sensitivity (Recall): 0.78 Specificity: 0.85 Precision: 0.64 F1 Score: 0.7
Random Forest	max_depth=15, max_features=5, min_samples_leaf=10, min_samples_split=20, n_estimators=15	OOB Score: 0.85 Model Accuracy: 0.88 Sensitivity (Recall): 0.9 Specificity: 0.86 Precision: 0.87 F1 Score: 0.88	OOB Score: 0.82 Model Accuracy: 0.82 Sensitivity (Recall): 0.74 Specificity: 0.82 Precision: 0.58 F1 Score: 0.65

9. Conclusion

- **Feature Importance: Both models identified features such as *incident_severity, insured_hobbies, vehicle_age, injury_claim and vehicle_claim* as highly important for predicting fraud.**
- **Correlation Analysis: High correlations were found between some features (e.g., *vehicle_claim and total_claim_amount*), indicating potential multicollinearity, which was addressed during feature selection.**
- **Both logistic regression and random forest models provide robust performance for fraud detection, with logistic regression offering a slightly better F1 score.**
- **The optimal cutoff was chosen based on the trade-off between sensitivity and specificity, ensuring a balanced approach to fraud detection.**
- **Feature selection and correlation analysis helped in improving model interpretability and reducing redundancy.**
- **For deployment, either model can be considered, but logistic regression may be preferred for its simplicity and interpretability, while random forest can be chosen for potentially better handling of complex, non-linear relationships.**
- **Further improvements can be made by exploring advanced ensemble methods, additional feature engineering, or addressing class imbalance with techniques like SMOTE or cost-sensitive learning.**