Fraudulent Claim Detection

April 7th , 2025

By

Arpan Das - arsu.dwnall@gmail.com

Gagan Deep Madaan – gagandeep170899@gmail.com

Grishma.G – grishmareddy06@gmail.com

Problem Statement

- Global Insure a leading insurance company processes thousands of claims annually. A significant percentage of claims are fraudulent. But fraud is often detected after the company pays out, which causes heavy financial losses.

- The company follows manual inspection which is slow, inefficient and also consumes more resources.

- So the company wants to improve fraud detection using data-driven models by classifying claims as fraudulent or legitimate, early in the process by reducing financial losses and streamline the claims processing. 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 Preparation

- Given the csv file, it contains 1000 rows and 40 columns

- We did basic data feature checks, by using python commands

- From the given columns, _c39 is a null column, rest all columns except authorities_contacted doesn’t have any null values

- Majority of features are of object type or int64 type

- Columns like policy_bind_date and incident_date are in string format (object) and should be converted to datetime

- With the given numerical columns we also did statistical analysis of data

2. Data Cleaning

- In this step we will check the number of missing values in each column and columns having null values will be handled. _c39 has all null values which we dropped

- Also we have checked the columns with unique values and counts and checked for redundant values . Columns like collision_type, property_damage have placeholder values like '?' , which requires data cleaning

- Converting binary variables yes/no to 0 and 1

- Columns like policy_bind_date and incident_date are in string format (object) and we converted it to datetime format

- Changed auto_year column to object type

3. **Train-Validation Split**

- After the data cleaning step, we are splitting the data to train dataset and test dataset. We have Imported required libraries

- And as per the guidelines mentioned we are using stratification on the target variable

- We have split the dataset into 70% train dataset and 30% test dataset.

4. EDA on training data

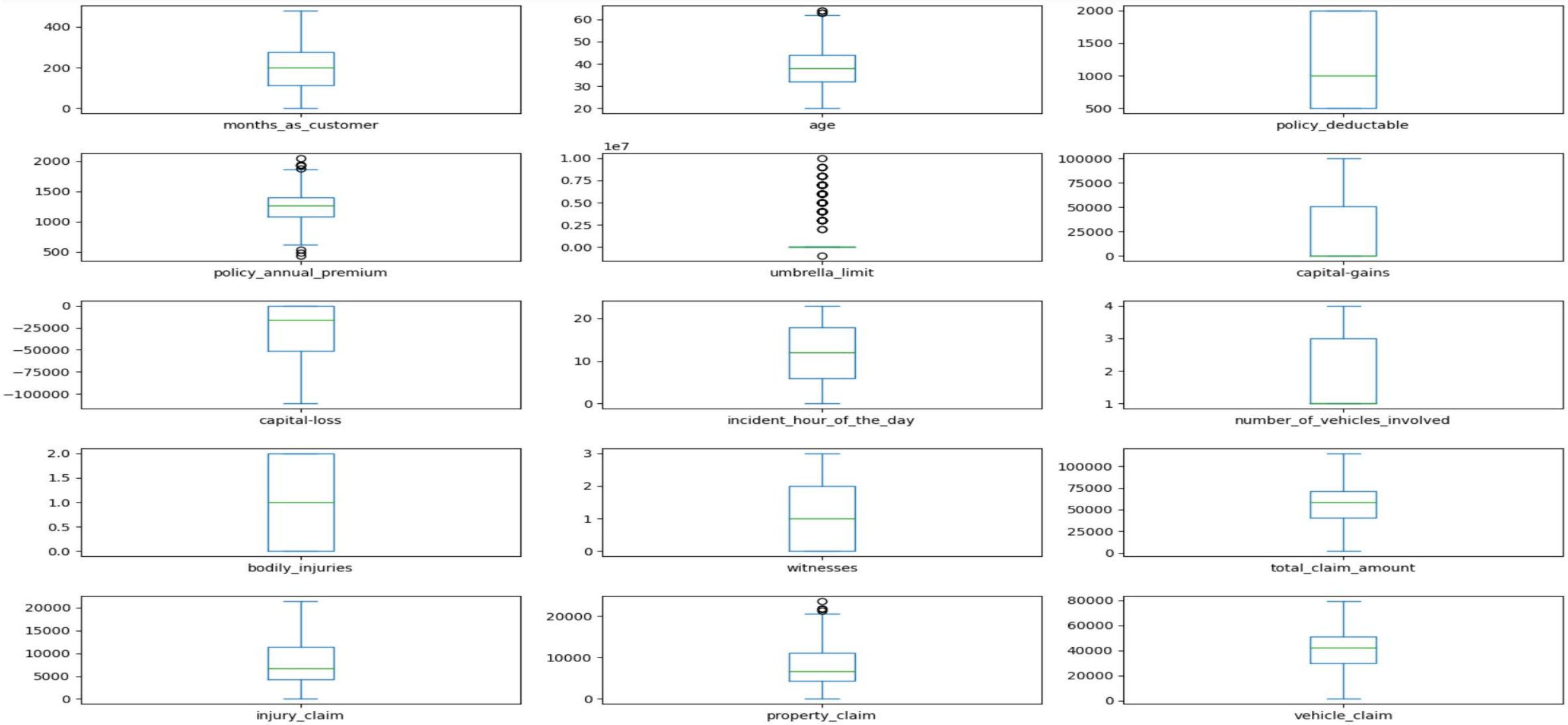
- We performed univariate analysis on the dataset

- We have identified and selected numerical columns from training data for univariate analysis

- Plotted all the numerical columns to understand their distribution

4. EDA on training data contd…

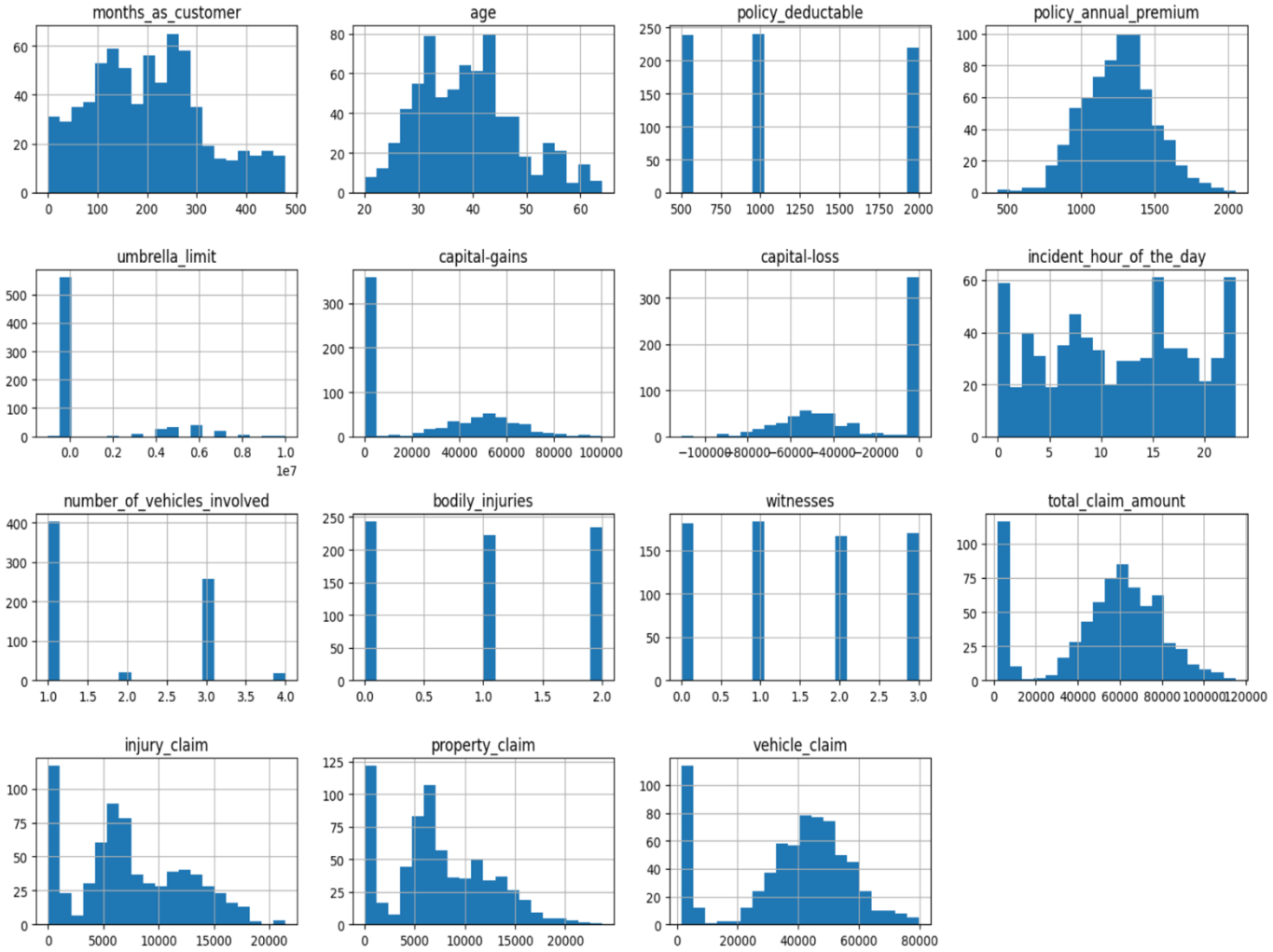
Boxplots - To check data spread and outliers



Insight from the graphs : We did not find any major outliers

4. EDA on training data contd…

Histograms - For univariate distributions



Few insights from the graphs:

- months_as_customer and incident_hour_of_the day doesn't follow normal distribution as expected as they are time stamps

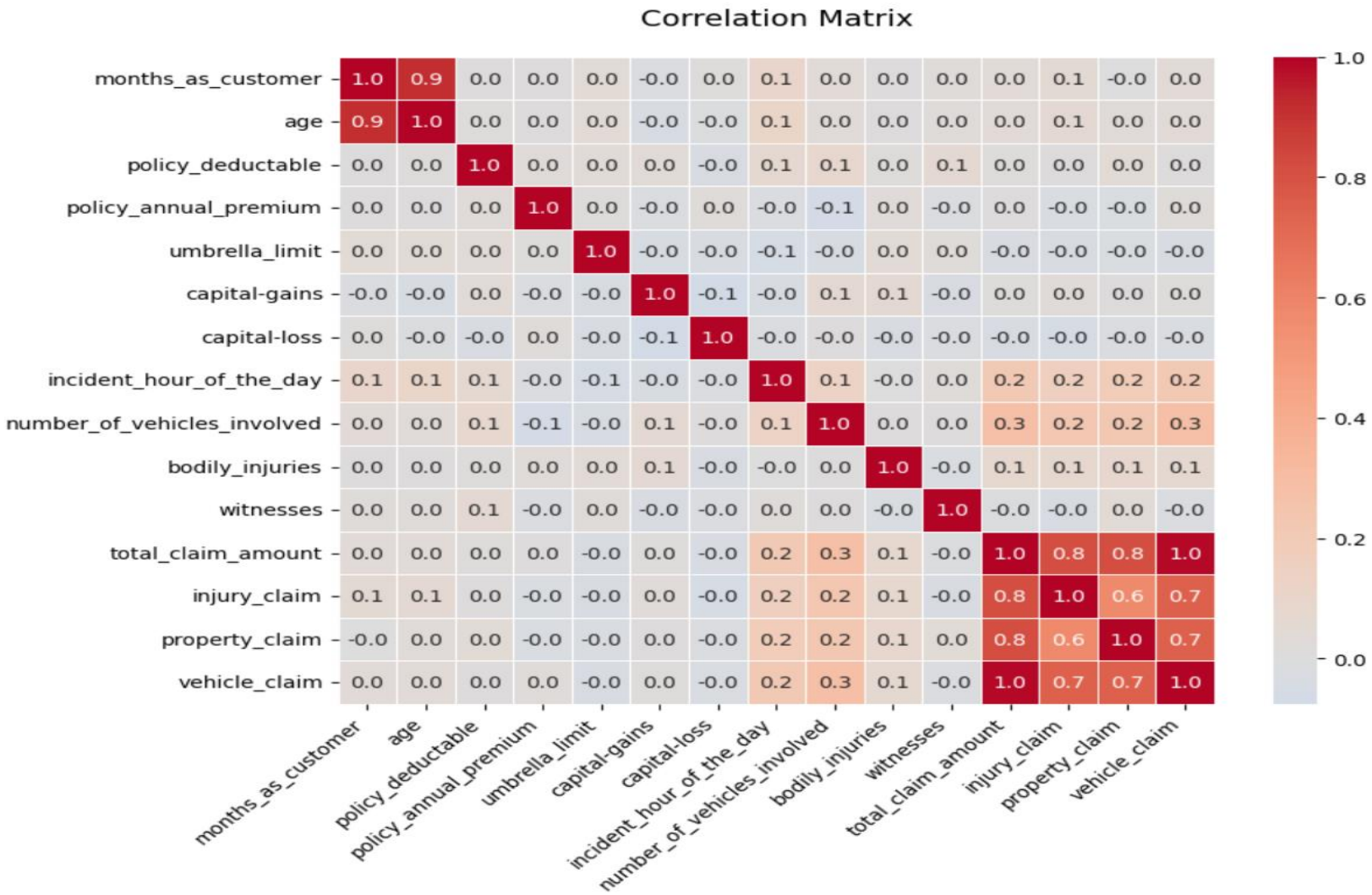
- Number of vehicles involved, bodily injuries, witnesses are few values so no normal distribution, which is an expected behaviour

-Policy_annual premium is normal to some extent

- Rest are left skewed due to one lower value

EDA on training data contd…

Correlation Matrix - Investigating the relationships between numerical features to identify potential multicollinearity or dependencies. Visualising the correlation structure using the heatmap to gain insights into feature relationships.

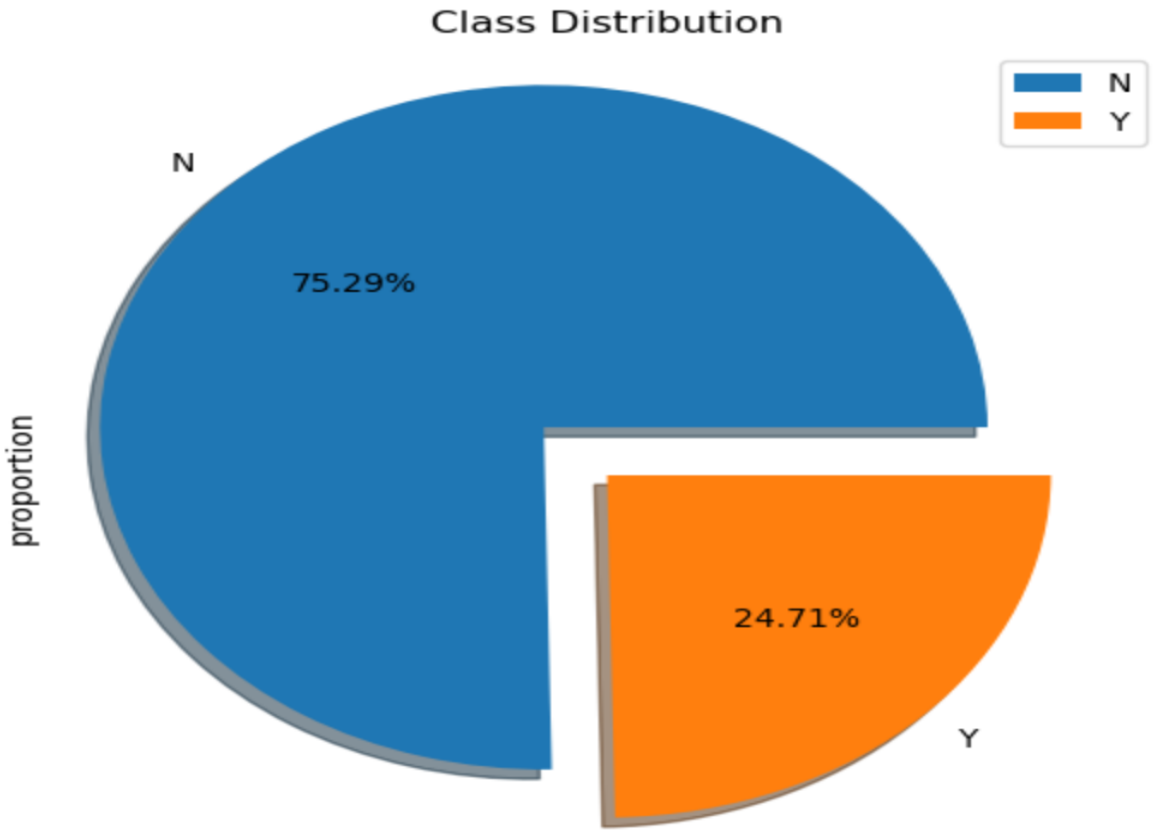


Insight from the graph :

- Claim amounts are highly corelated so only one should be retained and rest can be dropped

4. EDA on training data contd…

Check class balance - Examining the distribution of target variable to identify potential class imbalances using visualisation for better understanding.



Insight from the graph :

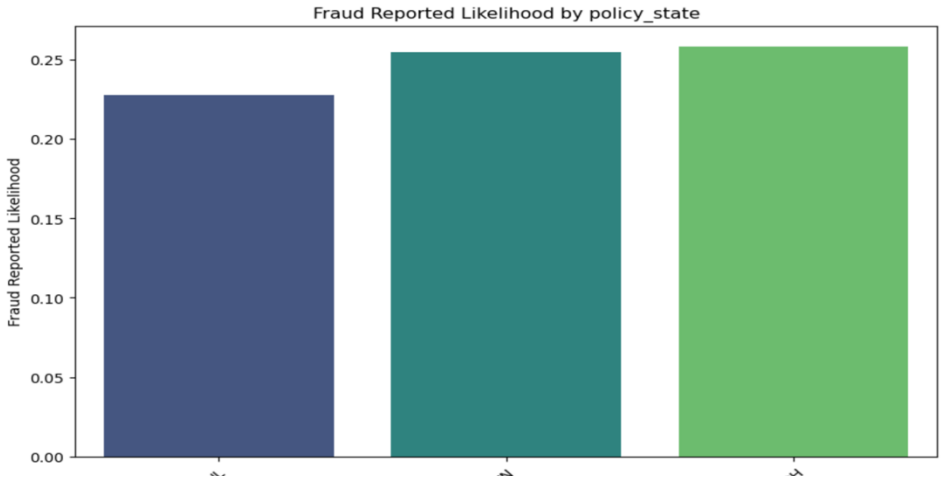
fraud_reported N
75.285714
Y 24.714286

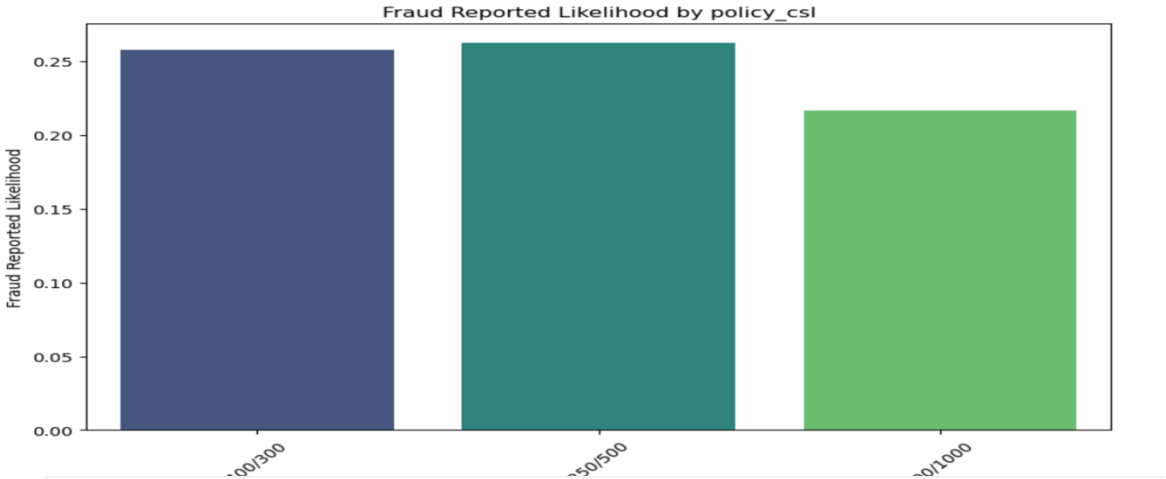
- Class imbalance not that much for a fraud detection data set.

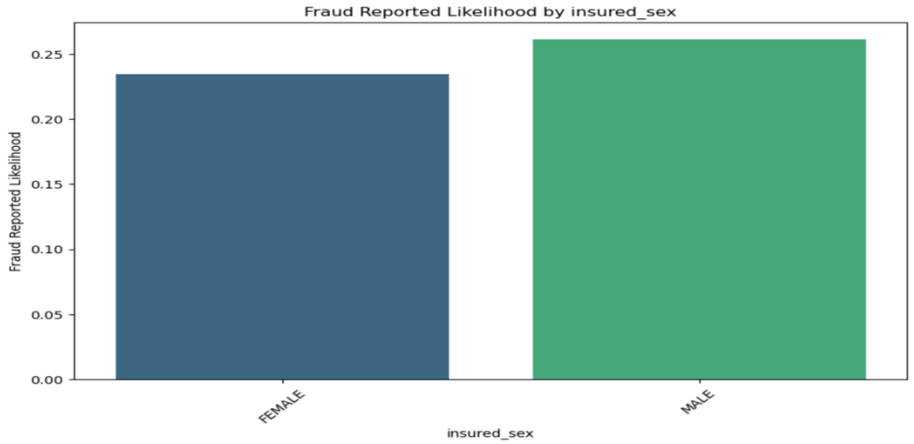
- Will be handled later in Random Sampler

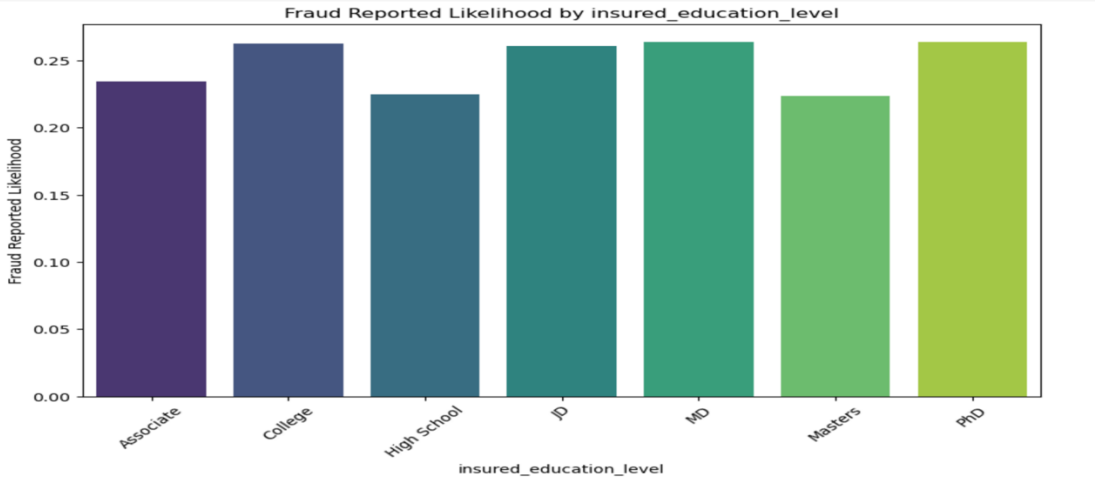
4. EDA on training data contd…
Bivariate analysis - Train Dataset

Target likelihood analysis for categorical variables - Investigating the relationships between categorical features and the target variable by analysing the target event likelihood (for the 'Y' event) for each level of every relevant categorical feature. Through this analysis, we are identifying categorical features that do not contribute much in explaining the variation in the target variable.

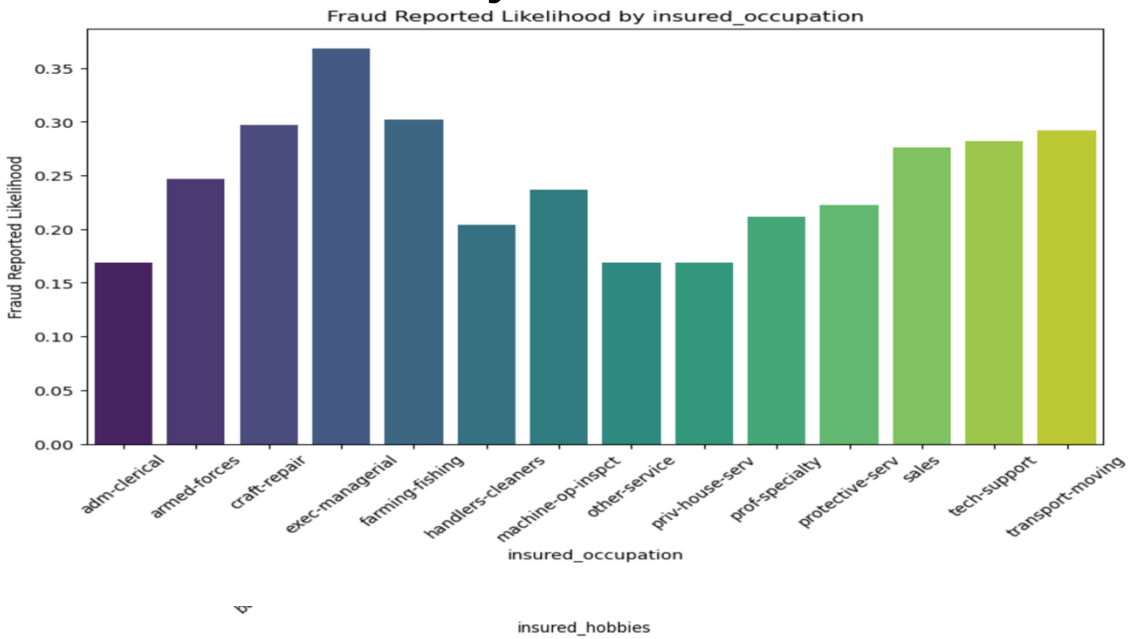


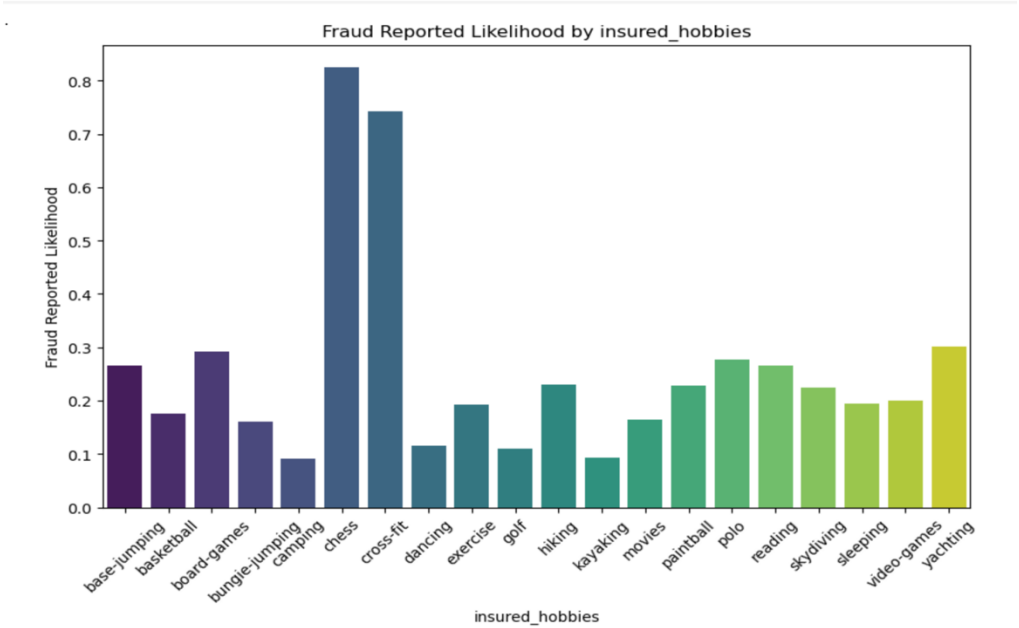


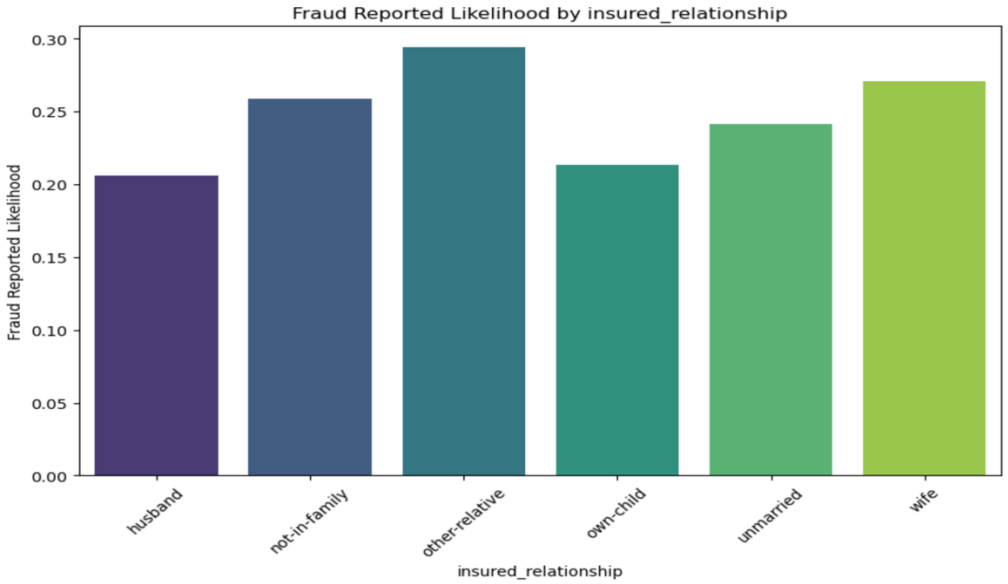


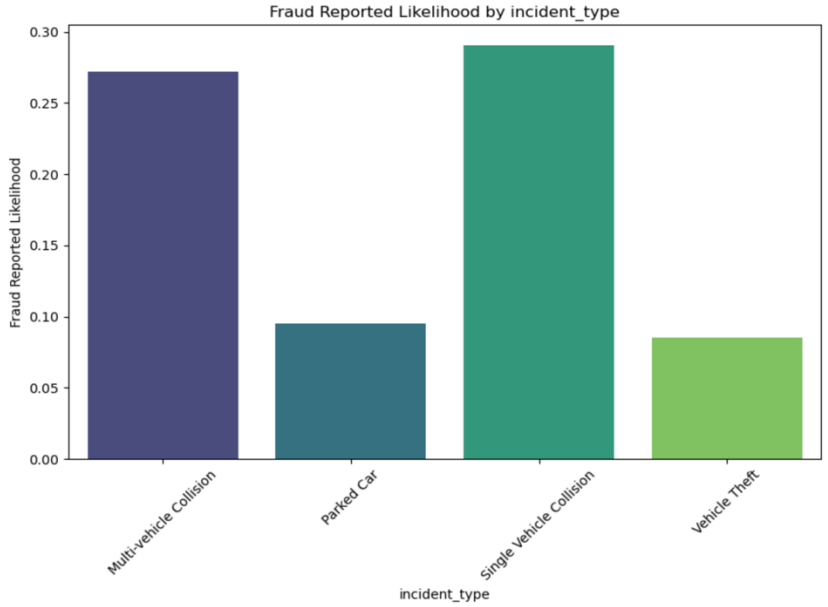


4. EDA on training data contd…
Bivariate analysis - Train Dataset



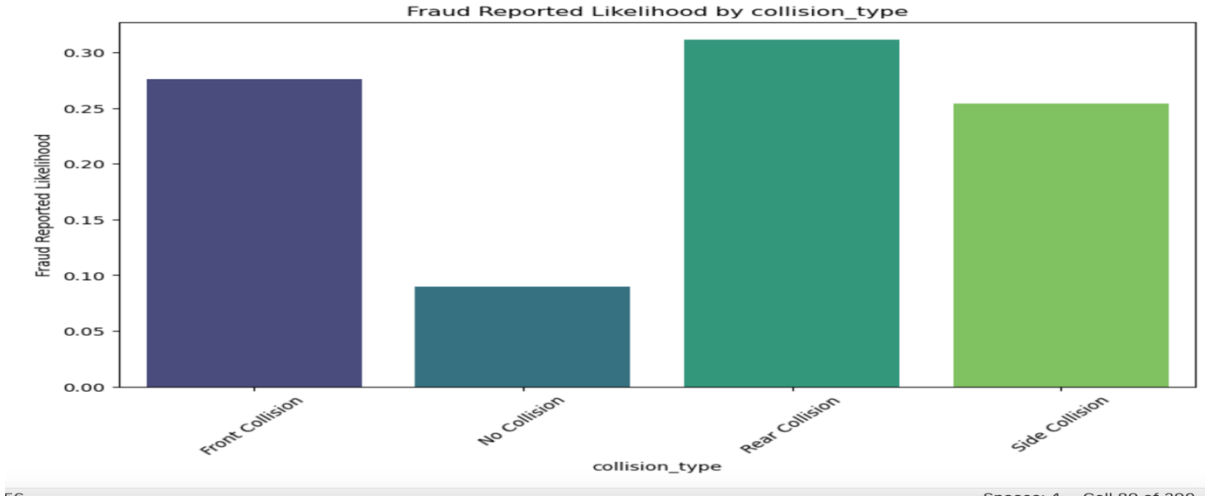


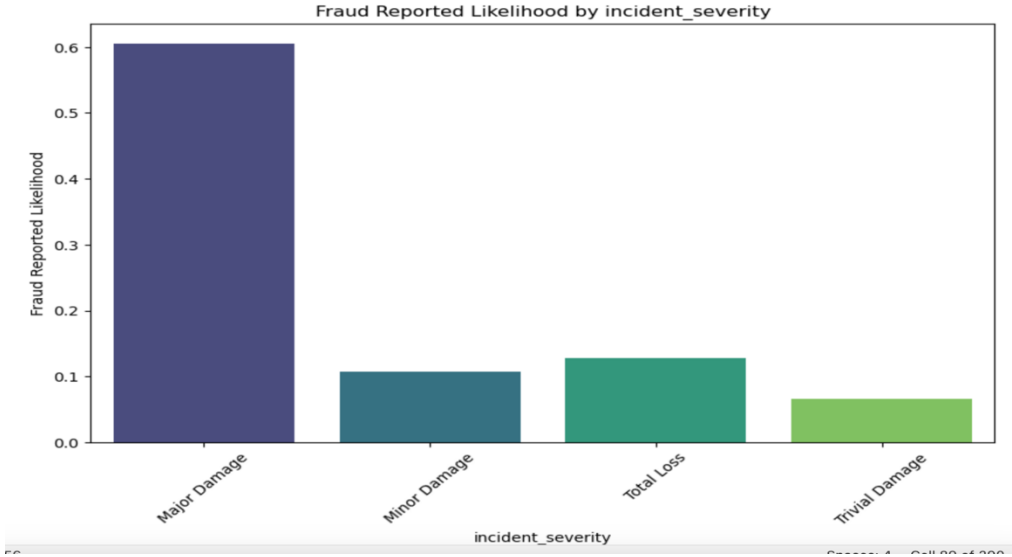


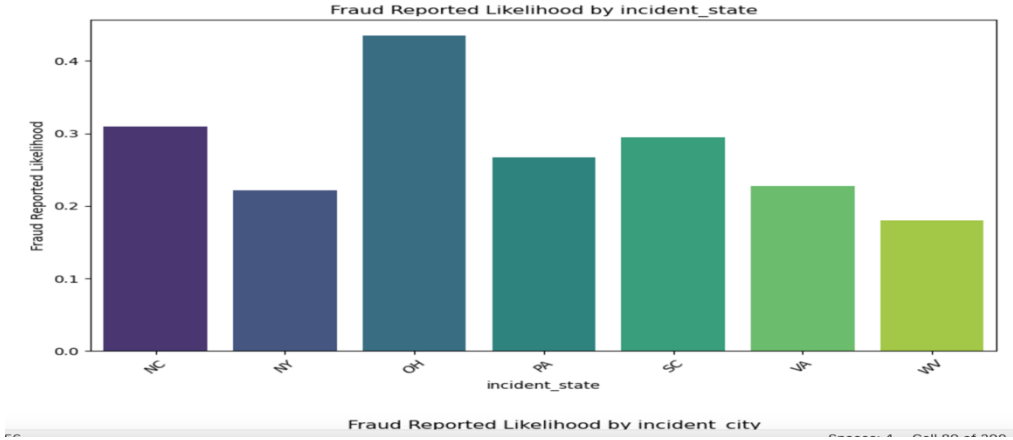


4. EDA on training data contd…

Bivariate analysis - Train Dataset

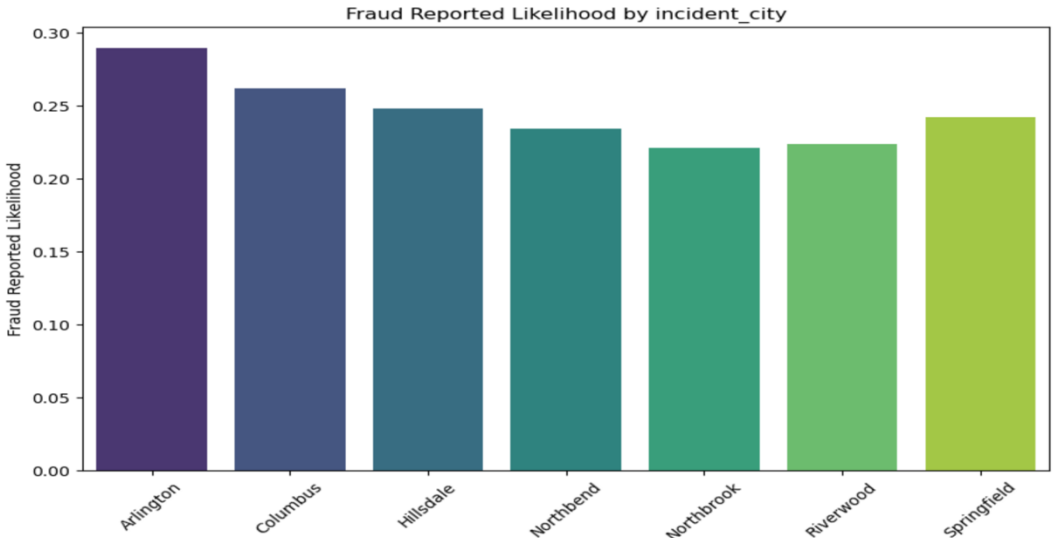


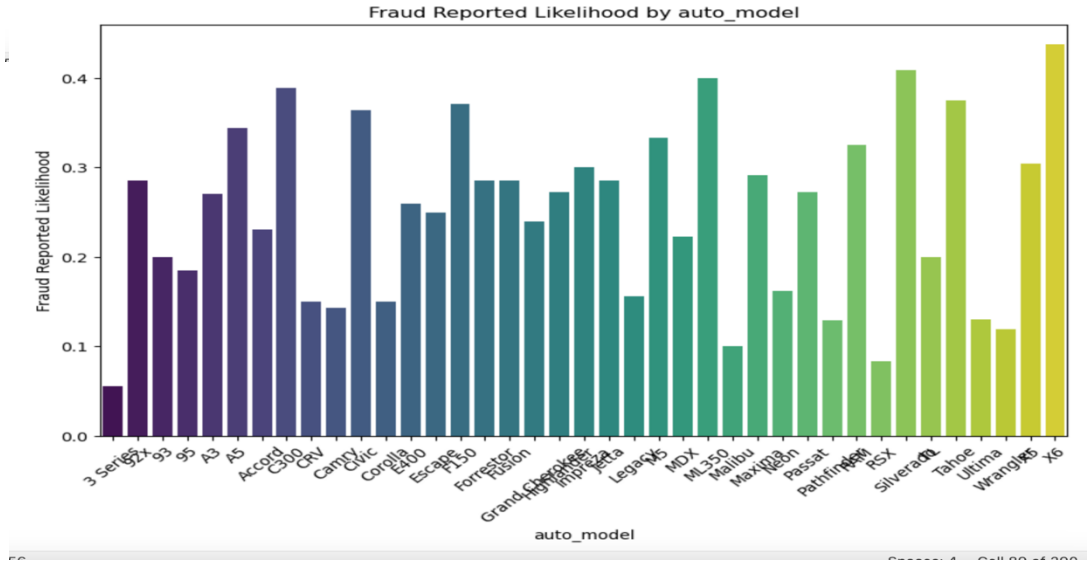


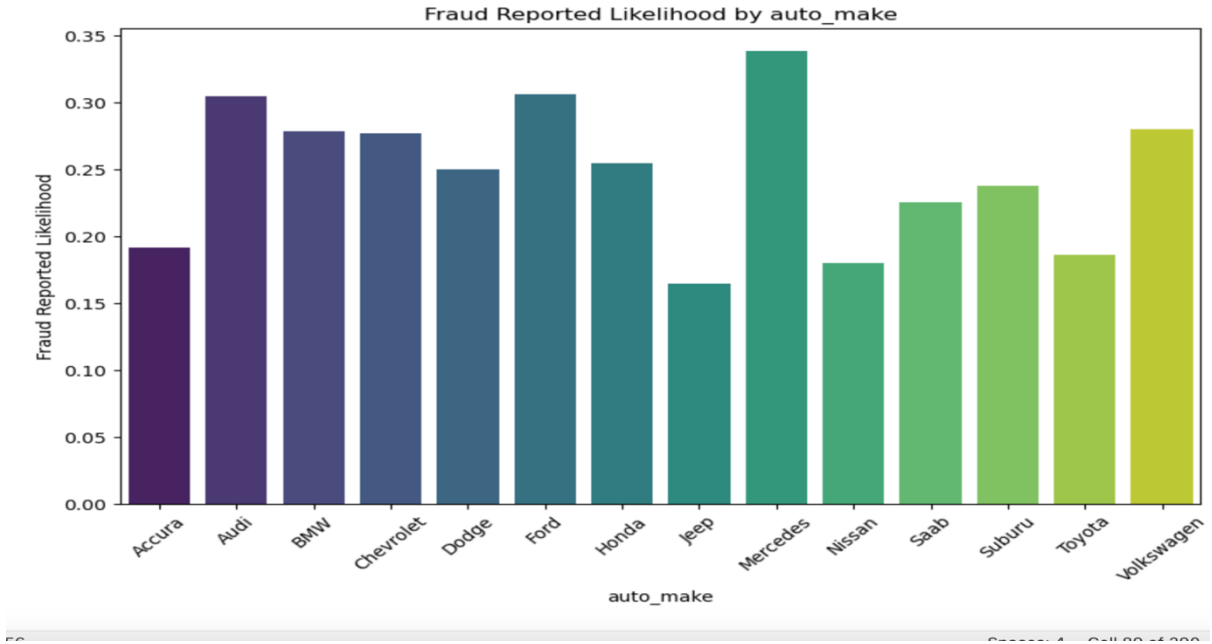


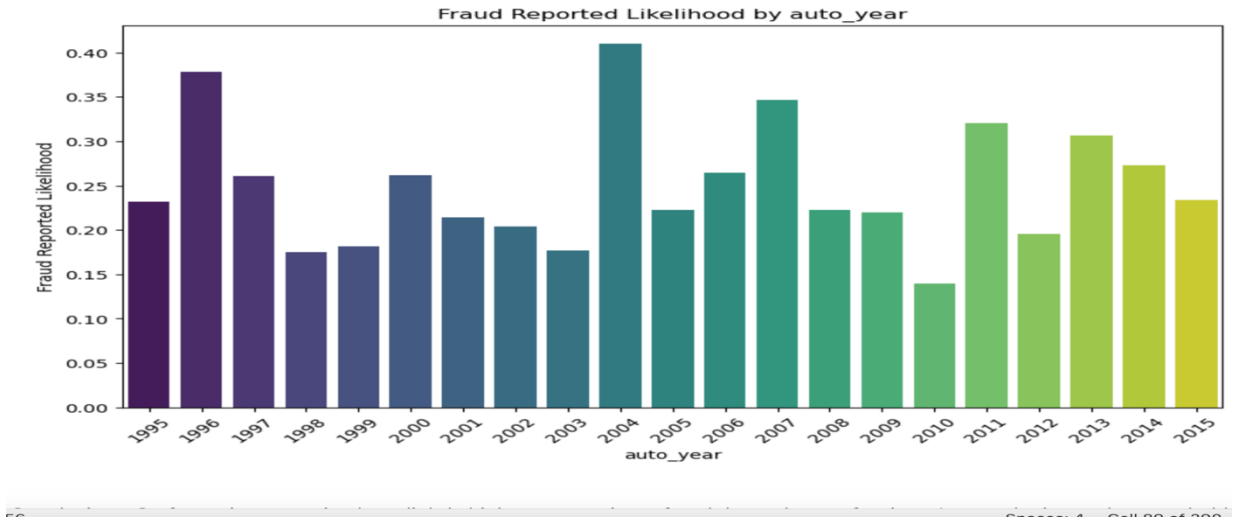
4. EDA on training data contd…

Bivariate analysis - Train Dataset



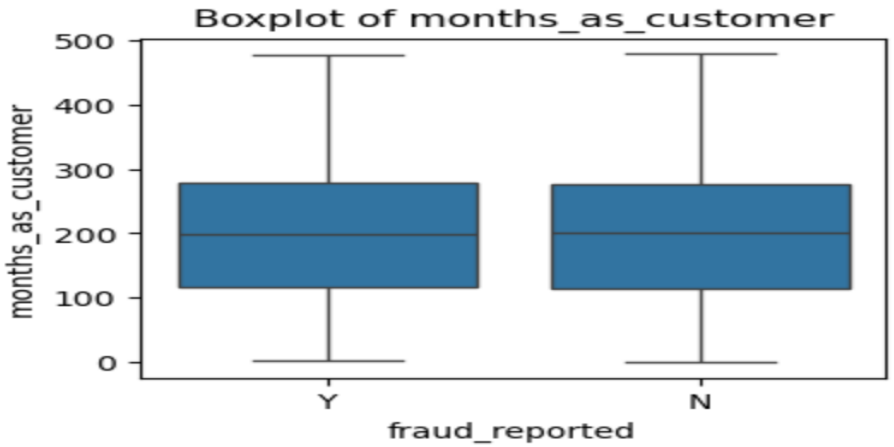


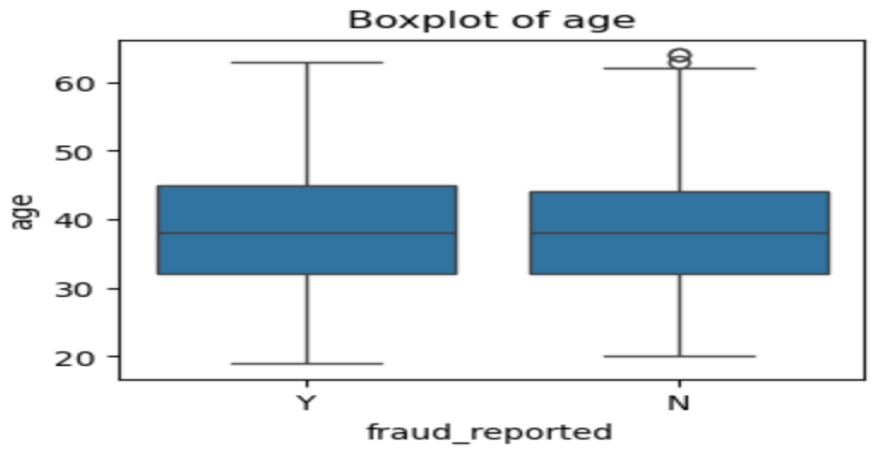


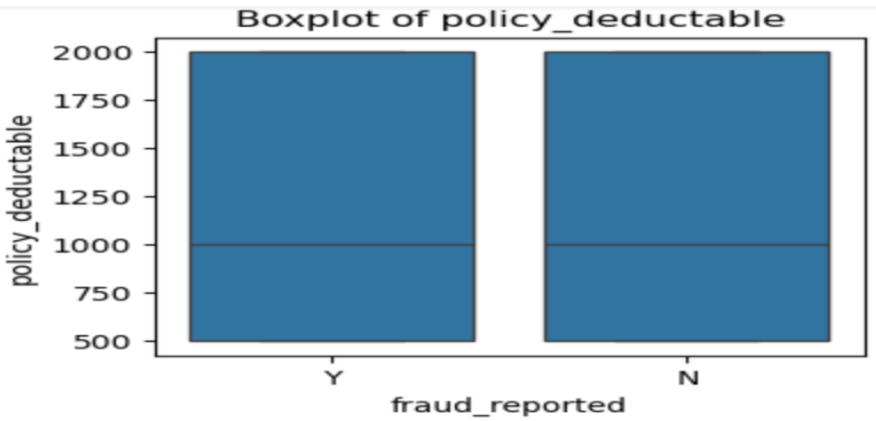


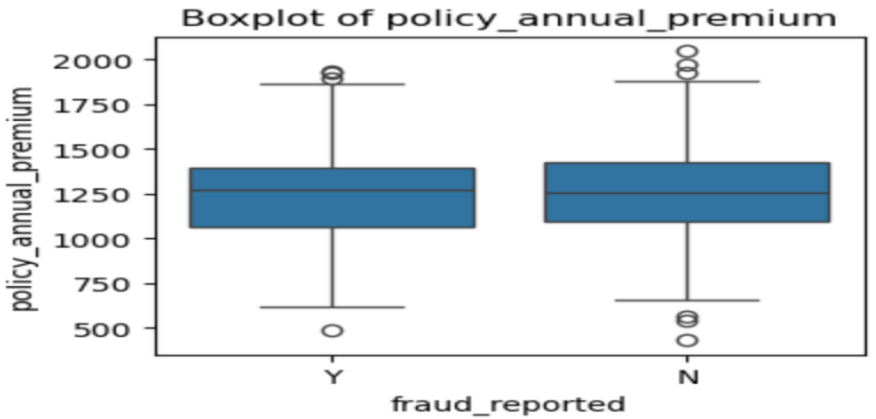
4. EDA on training data contd…

Visualise the relationship between numerical features and the target variable to understand their impact on the target outcome

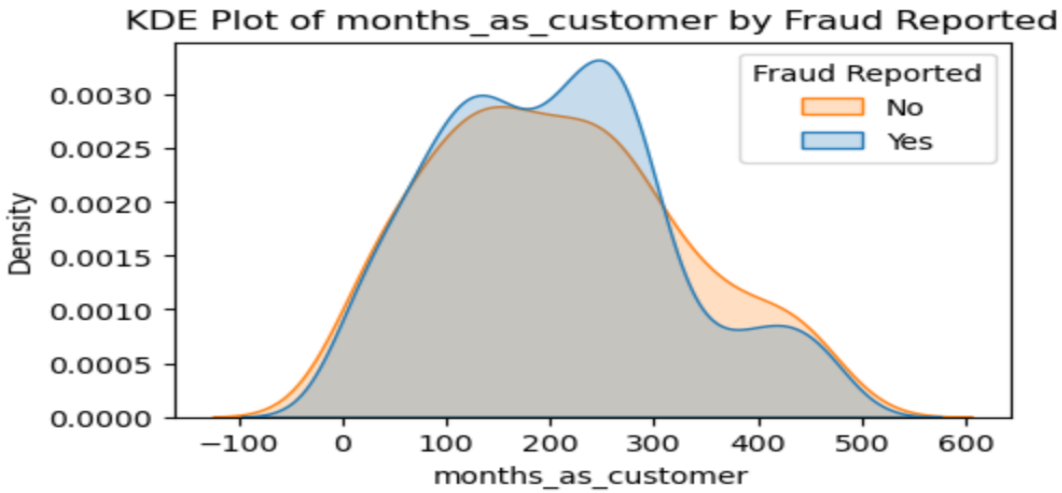


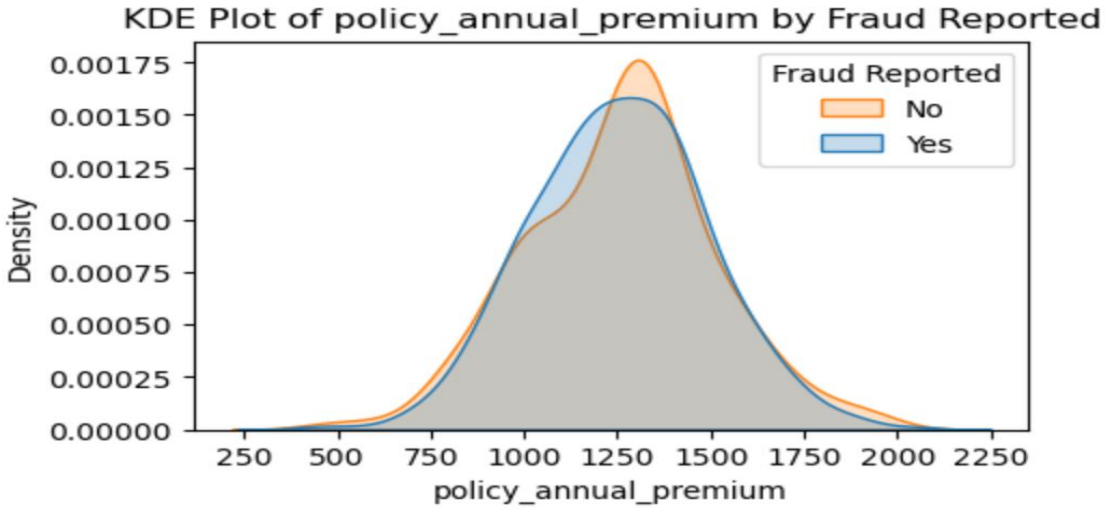


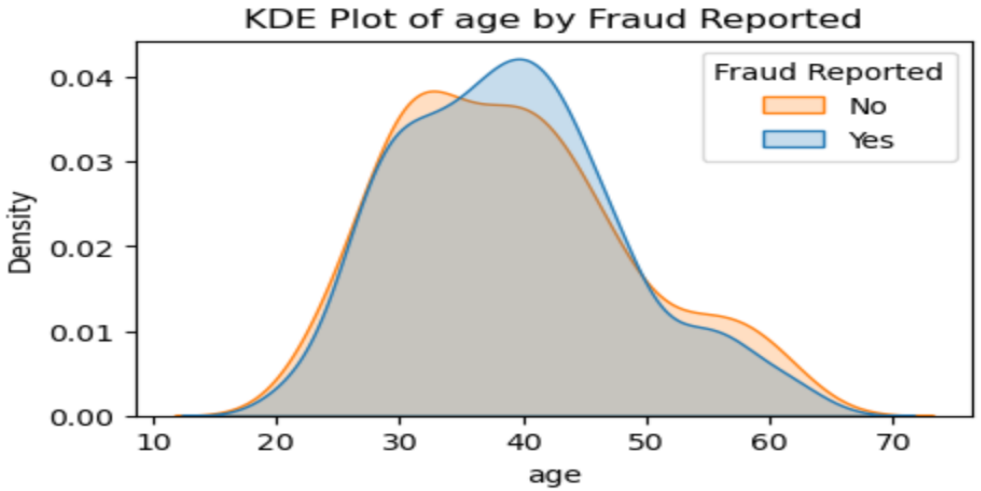


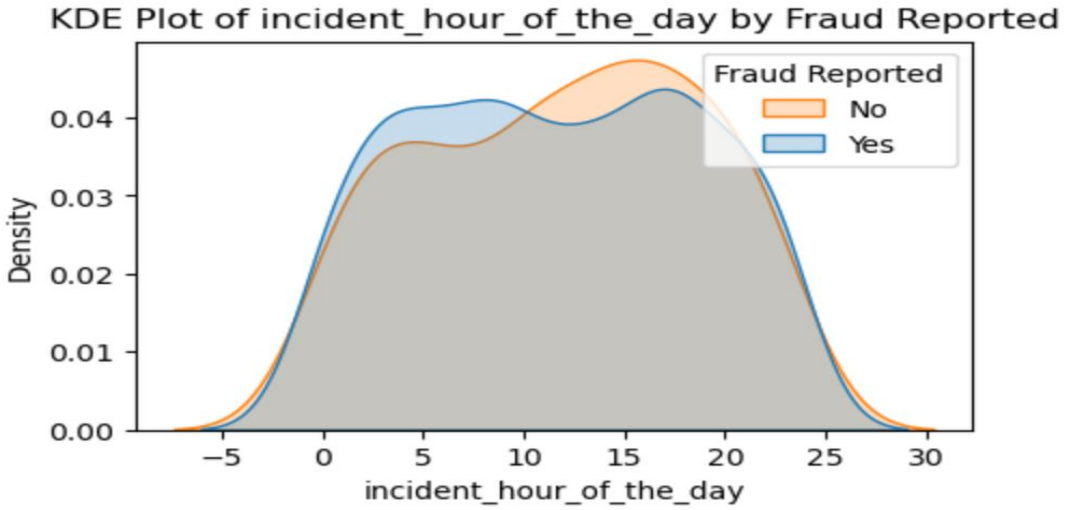


4. EDA on training data contd…
Bivariate analysis - Train Dataset

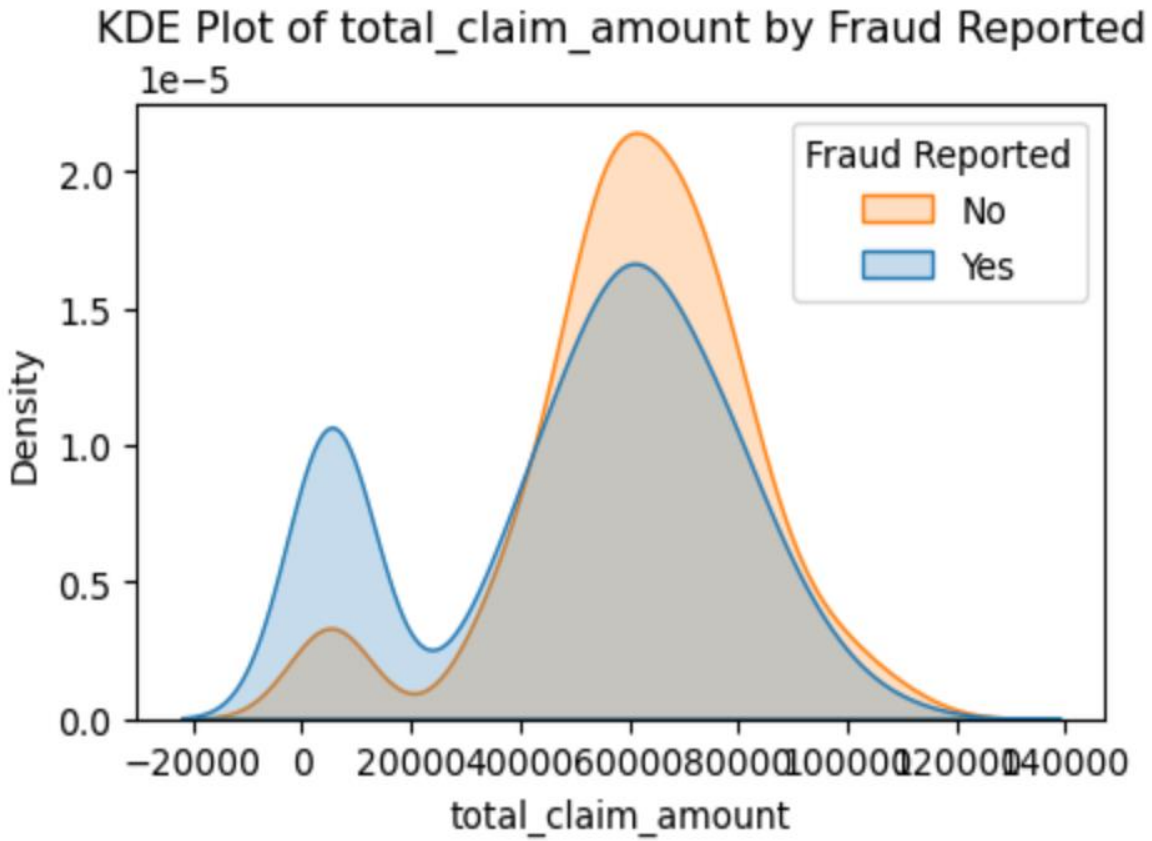








4. EDA on training data contd…
Bivariate analysis - Train Dataset

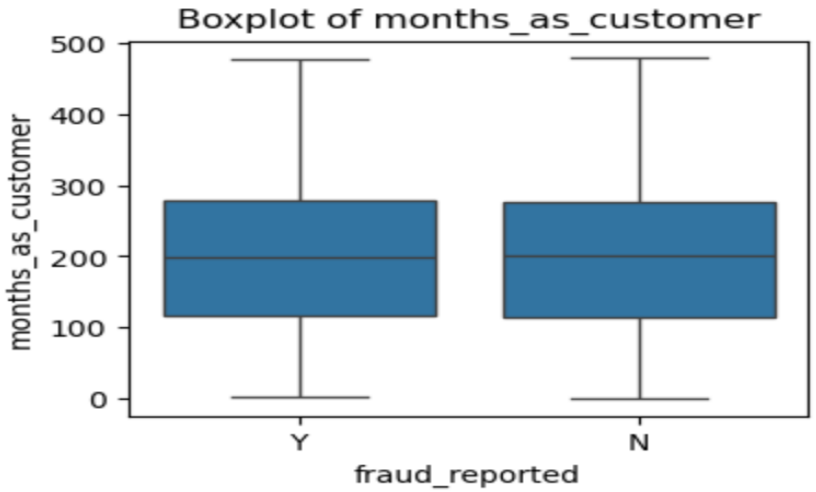


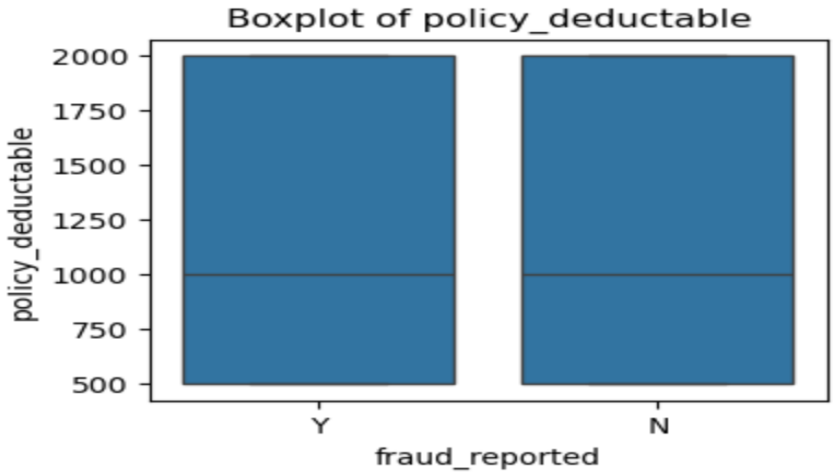
Insights from Graph :

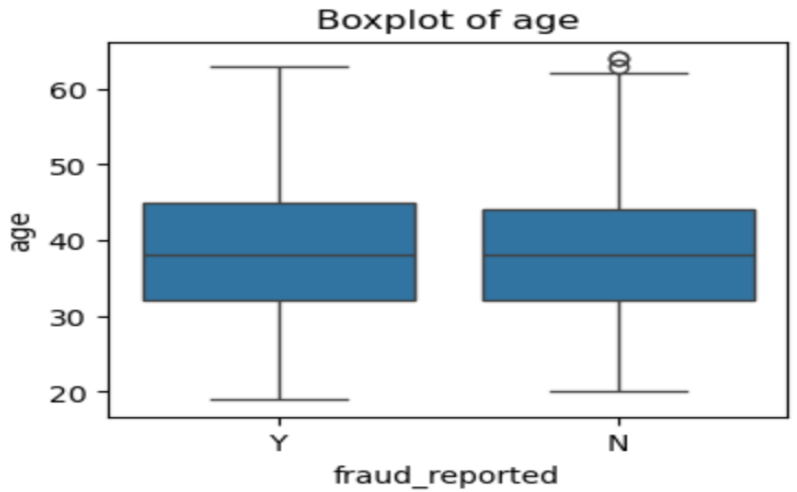
- Craft repair occupation has slightly higher propensity to fraud than other professions
- Among the insured person hobbies -people interested in Chess and Cross fit have observed to have more likelihood to fraud
- When the insured relative is other relative the fraud rate is higher
- Vehicle collisions are more resulting in frauds that theft
- Major damage has a much higher fraud likelihood than minor damages or total loss
- When No authorities were reported the number of frauds was also less implying fraudsters don't mind reporting to authorities appear authentic
- OH state has reported more frauds than rest

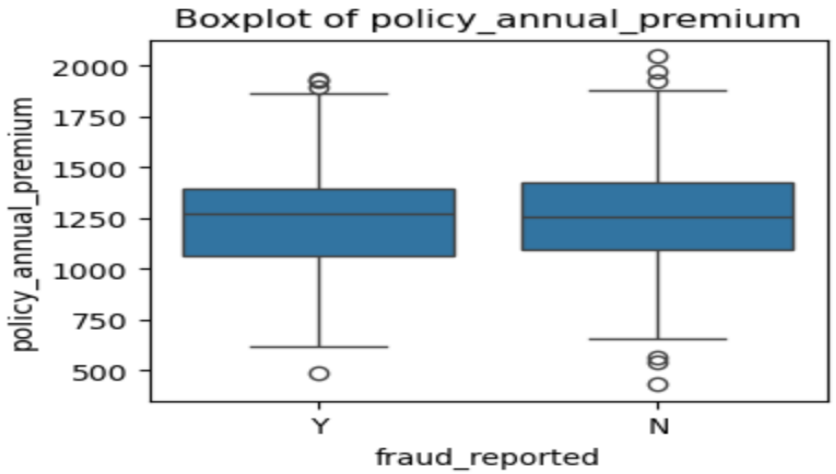
Conclusion : From KDE analysis we find that frauds are higher in lower claim amounts

4. EDA on training data contd…

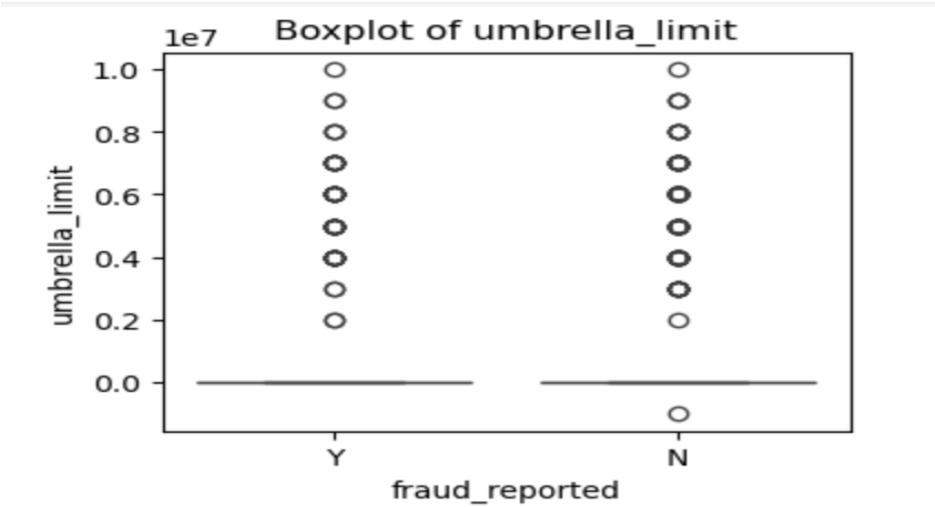


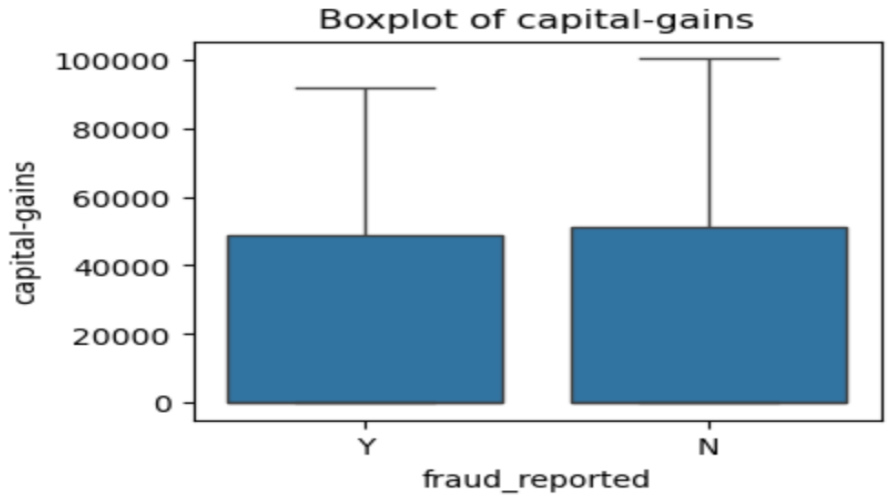


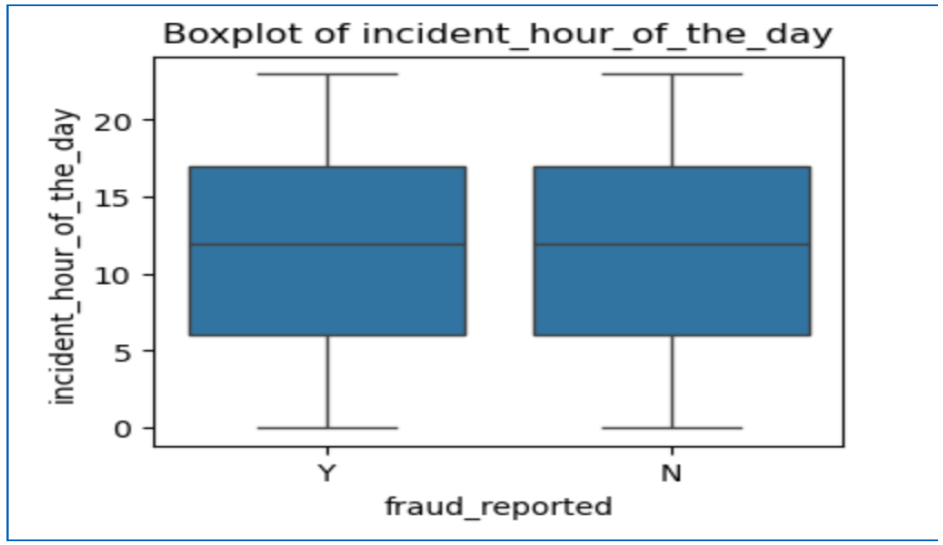




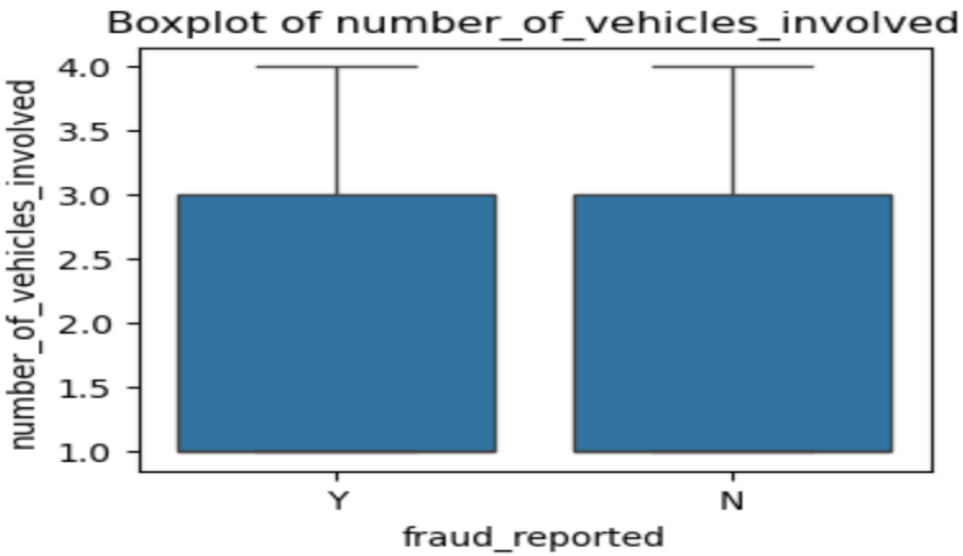
4. EDA on training data contd…

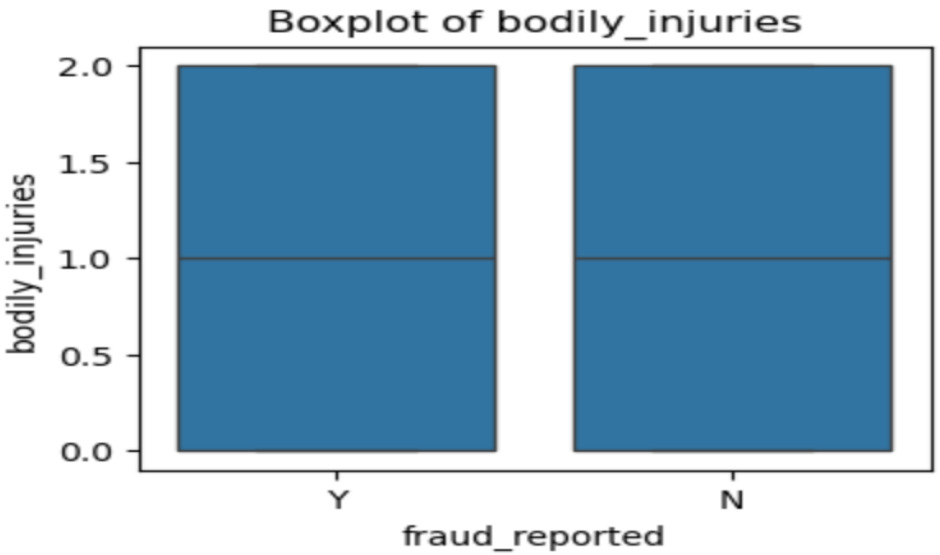


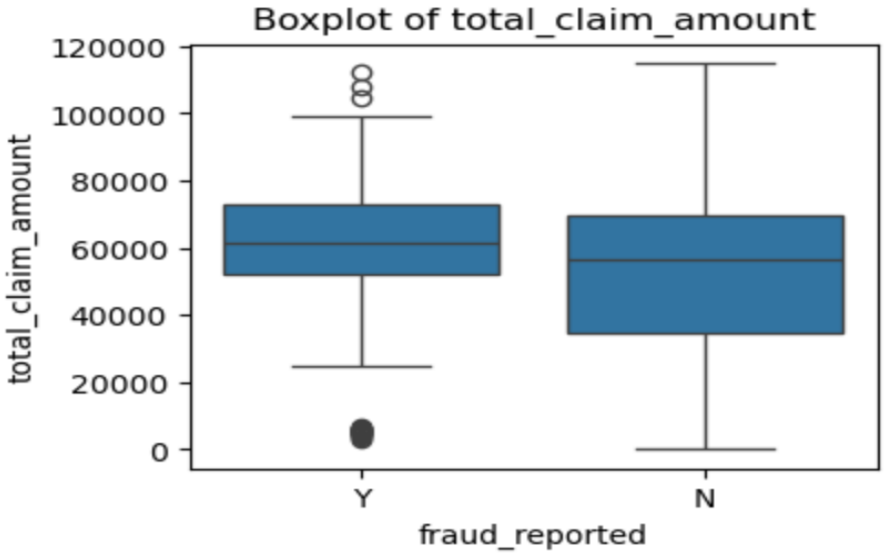




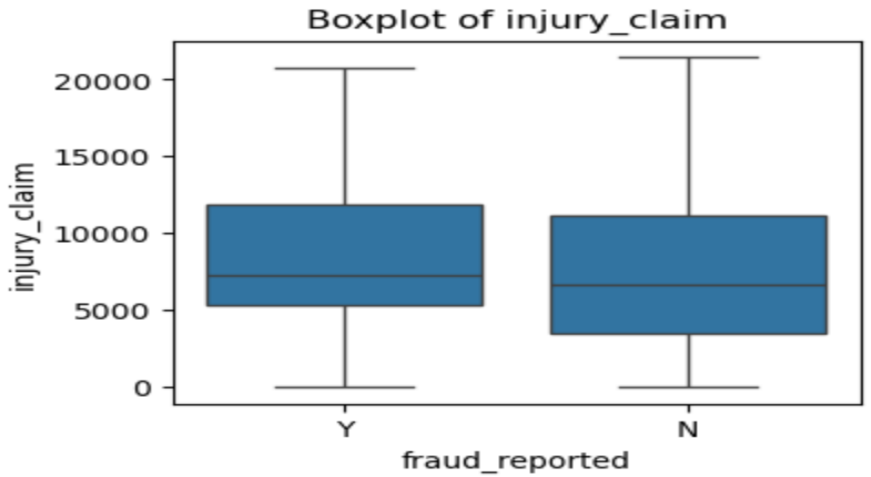
4. EDA on training data contd…

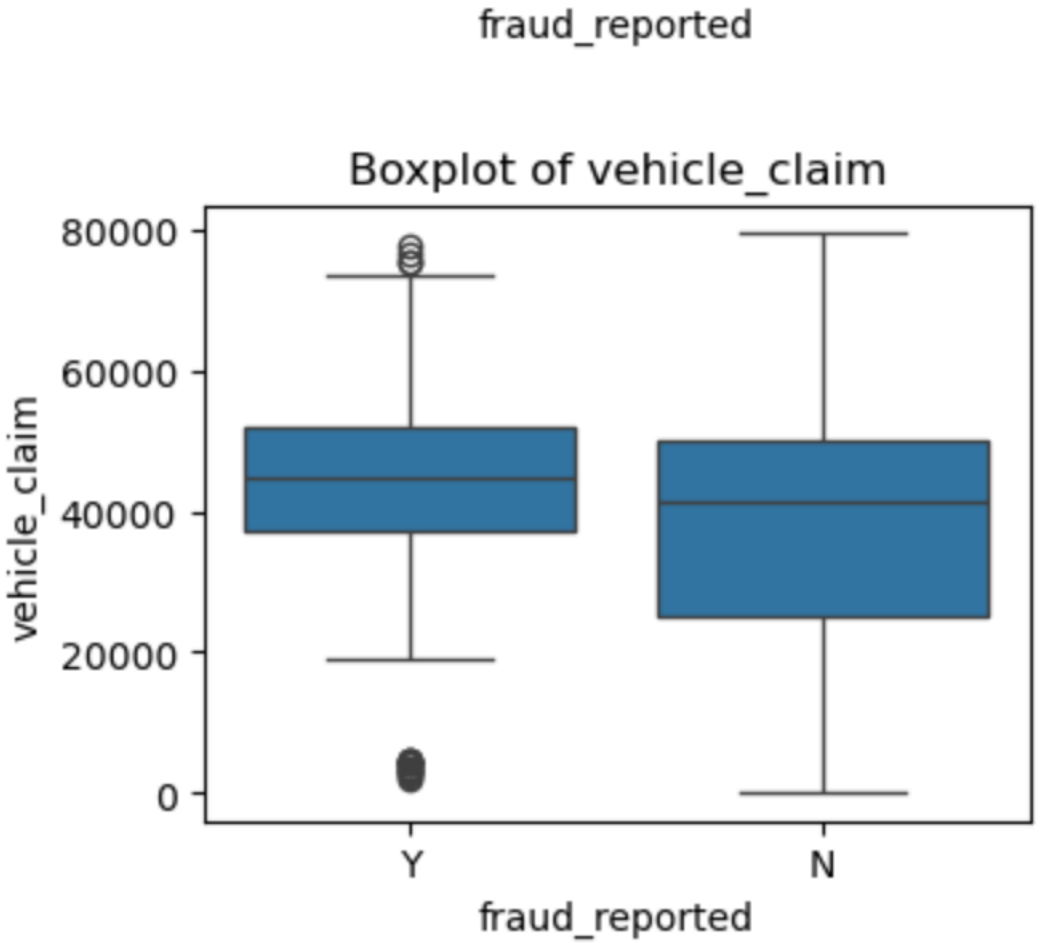






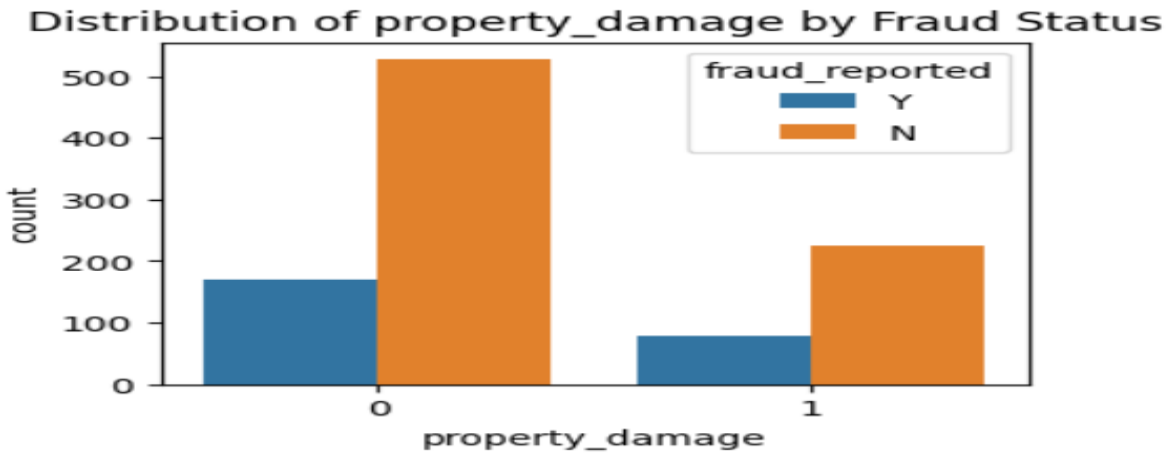
4. EDA on training data contd…

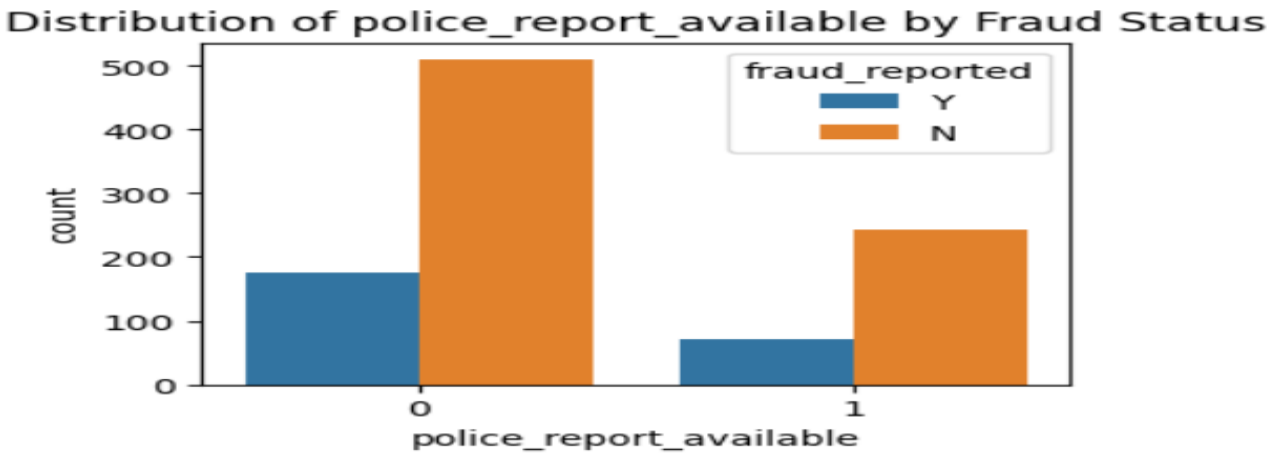




4. EDA on training data contd…

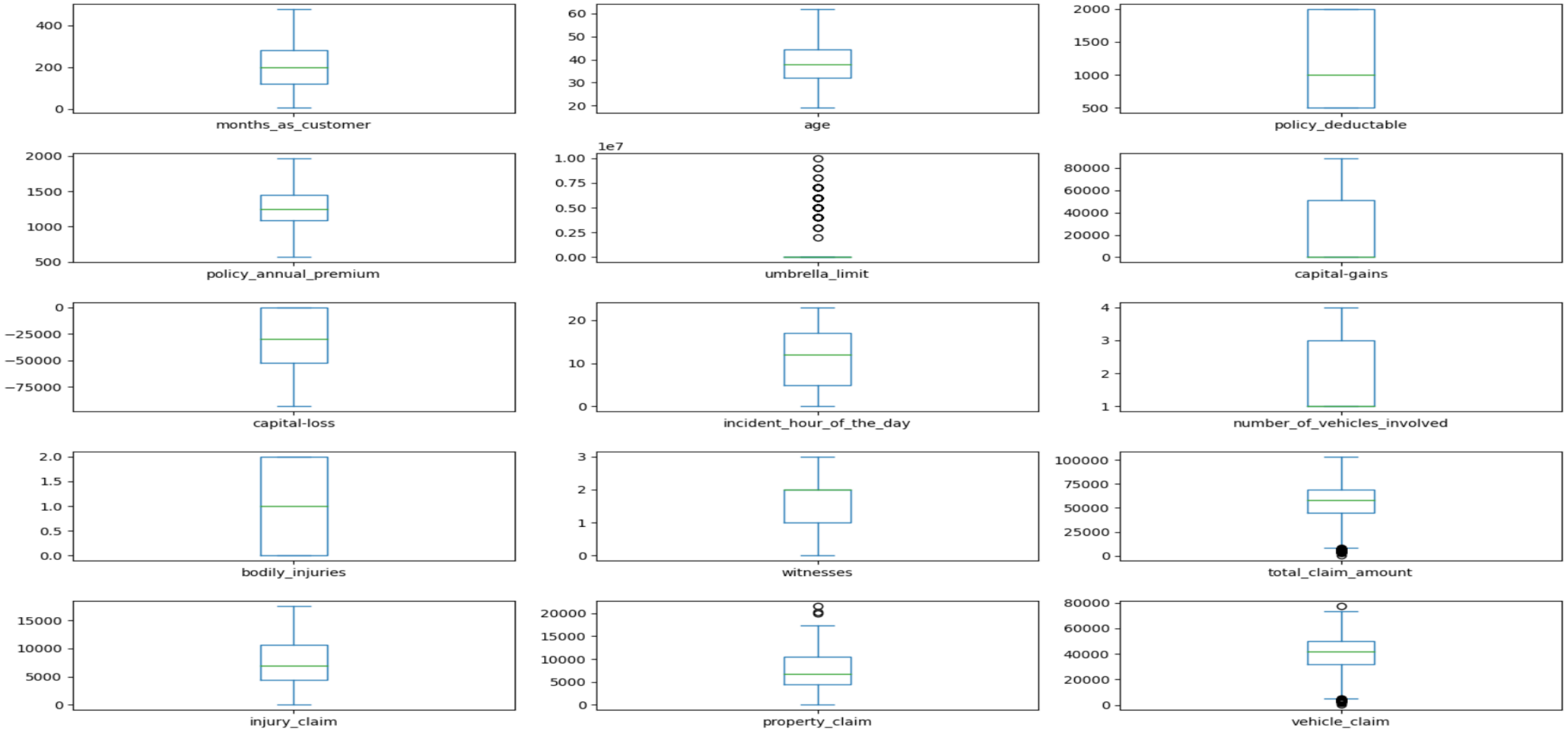
Distribution of property damage by fraud status





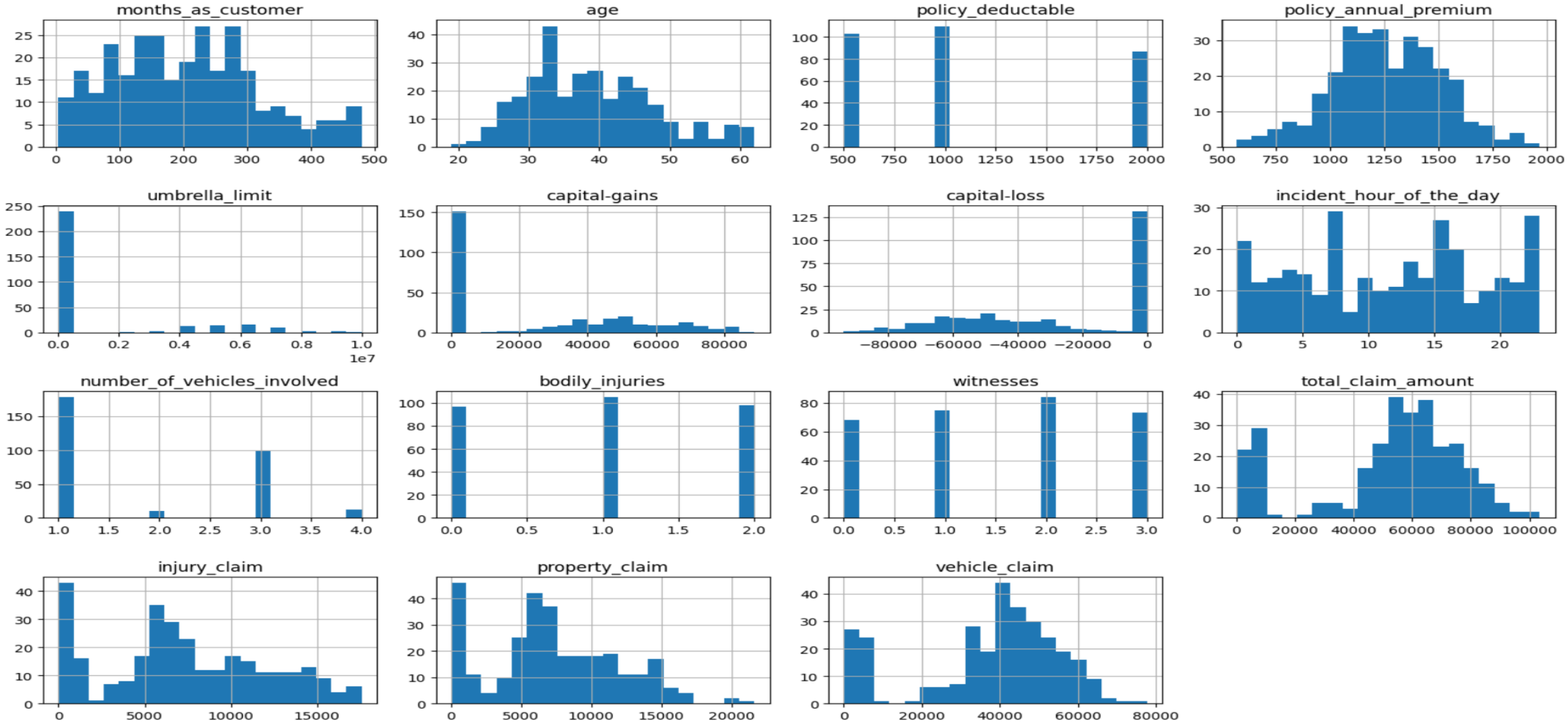
5 . EDA on validation data

Performing Univariate Analysis - Visualise the distribution of selected numerical features using appropriate plots to understand their characteristics
- Plotting all the numerical columns to understand their distribution
- Boxplots to check spread of data and outliers



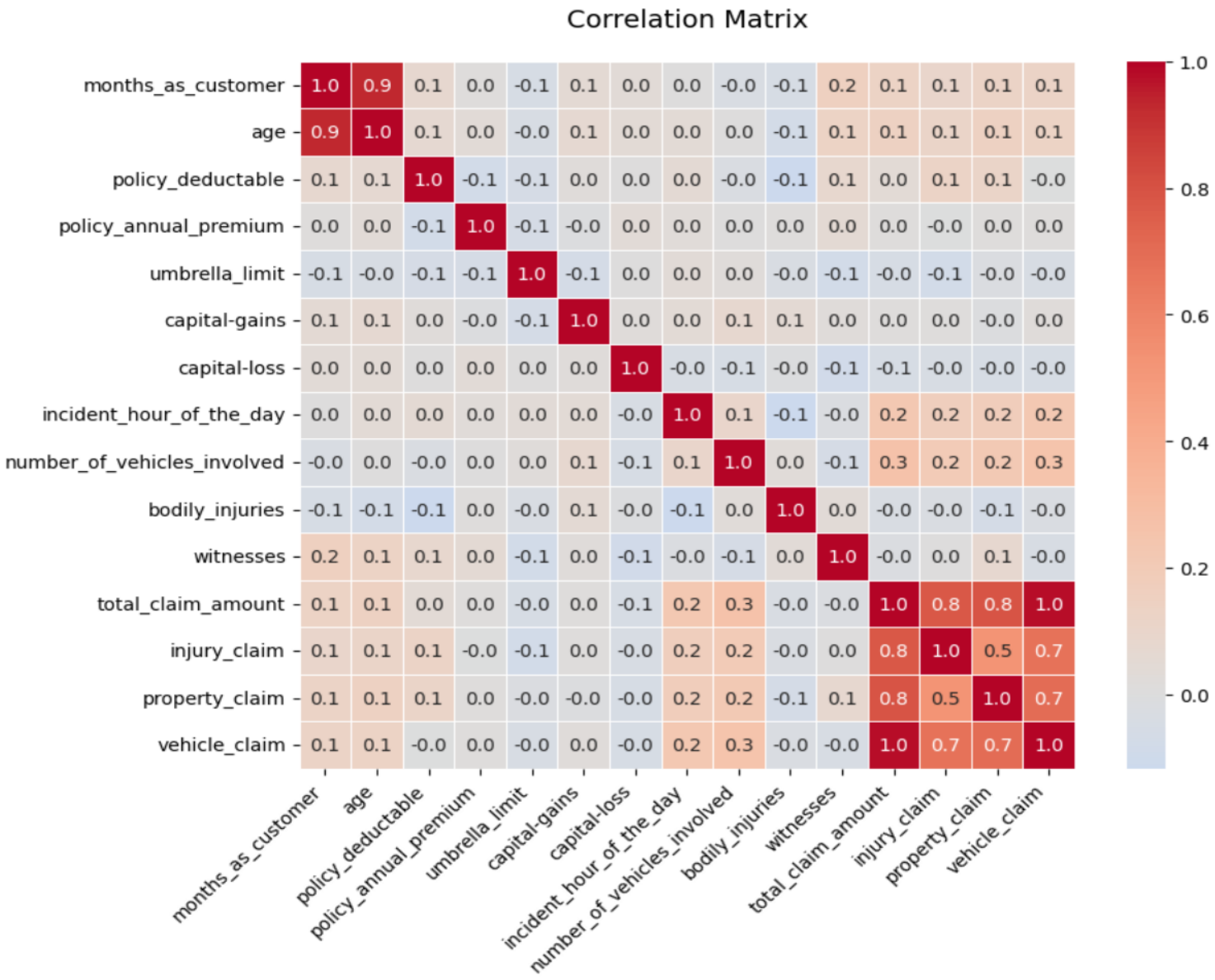
5 . EDA on validation data contd…

Histograms - For univariate distributions



5 . EDA on validation data contd…

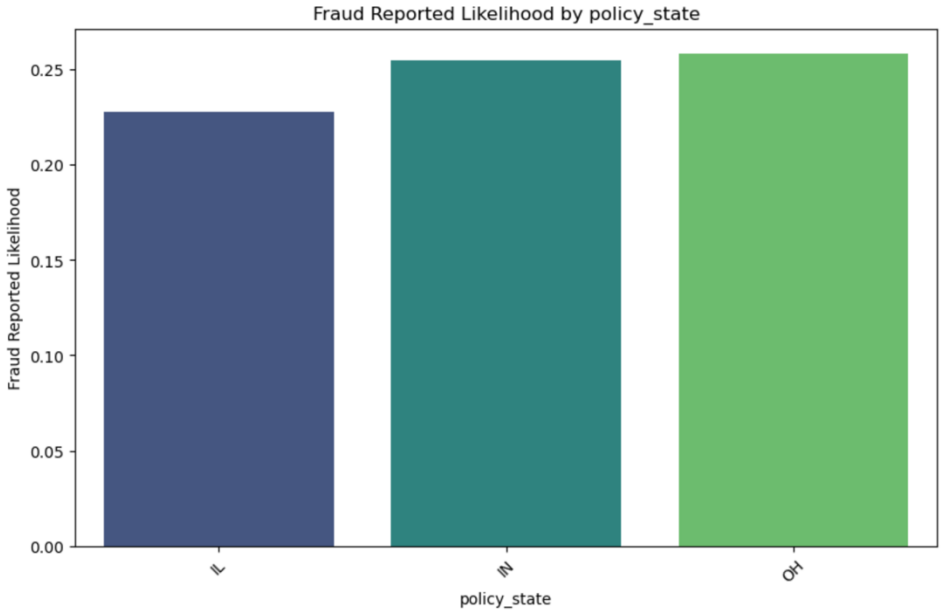
Pair plot to visualize the correlation

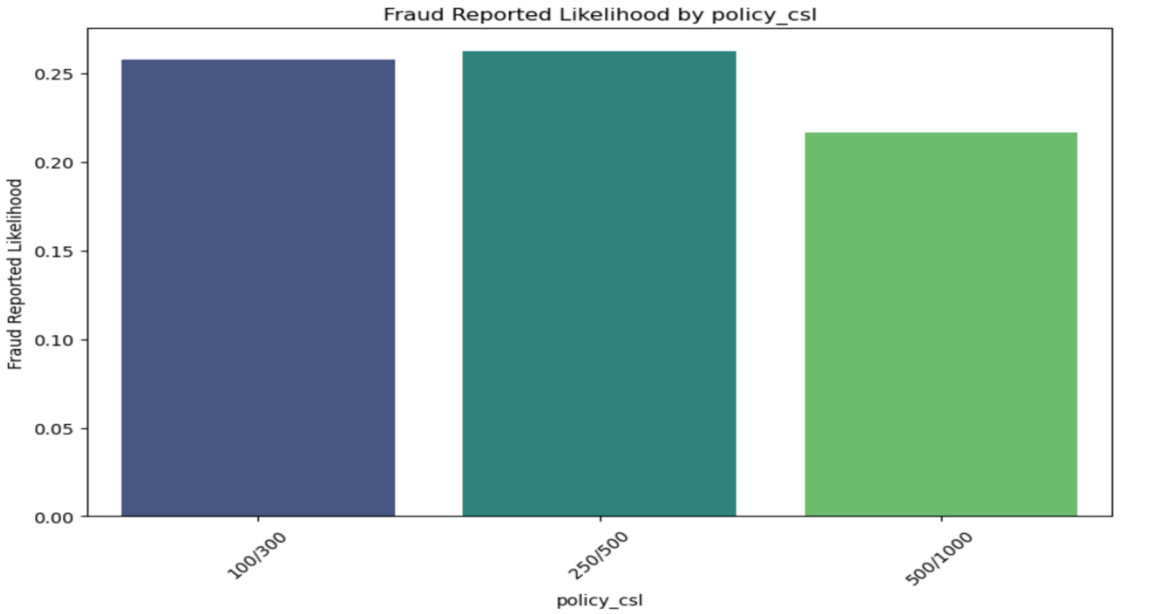


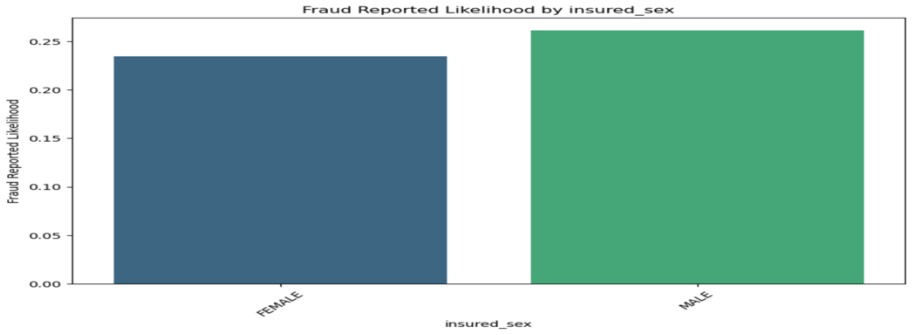
Insight from graph : Same insights as identified in train set, one claim should be retained rest should be dropped as they are highly corelated

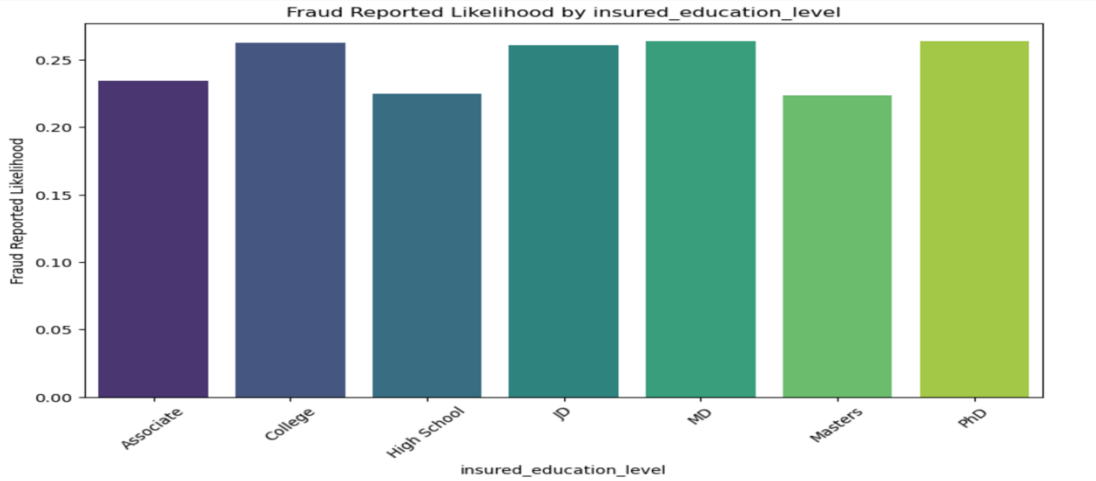
5 . EDA on validation data contd…

Bivariate Analysis



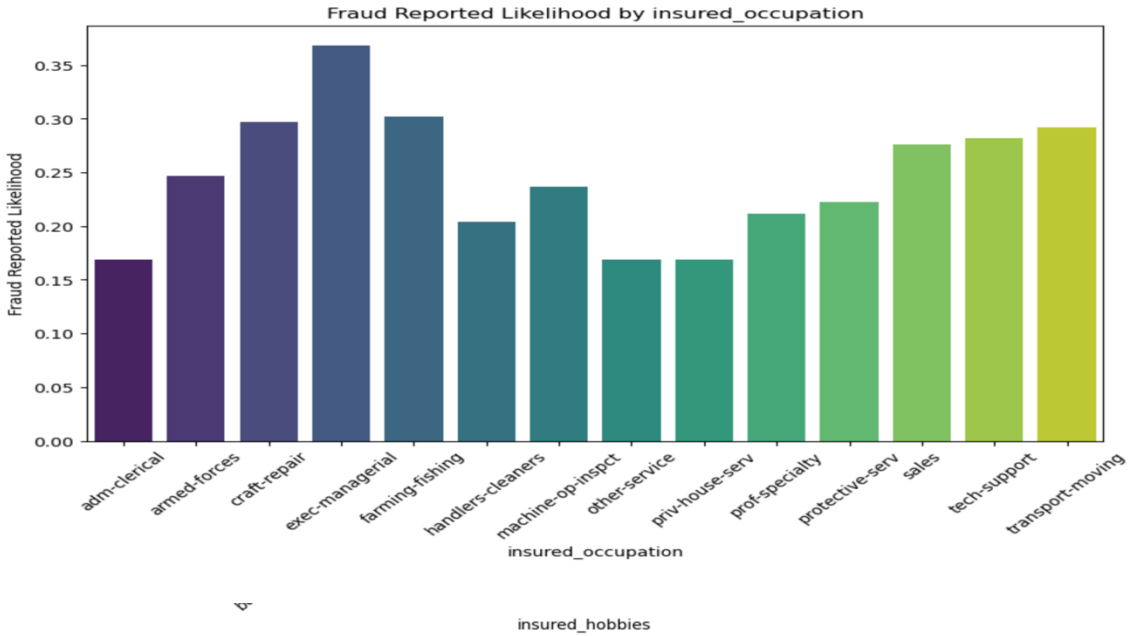


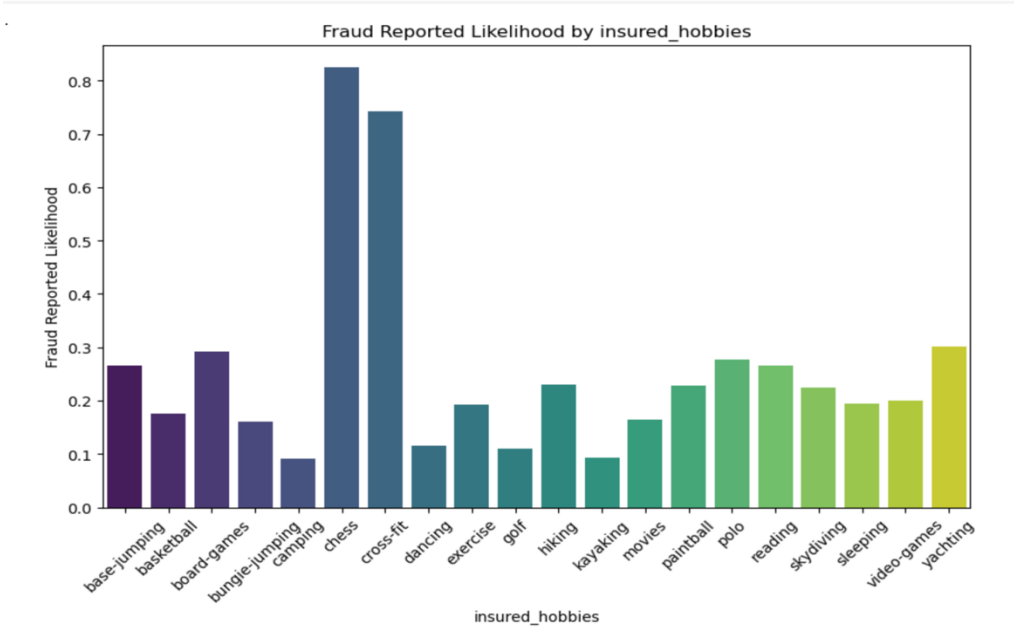


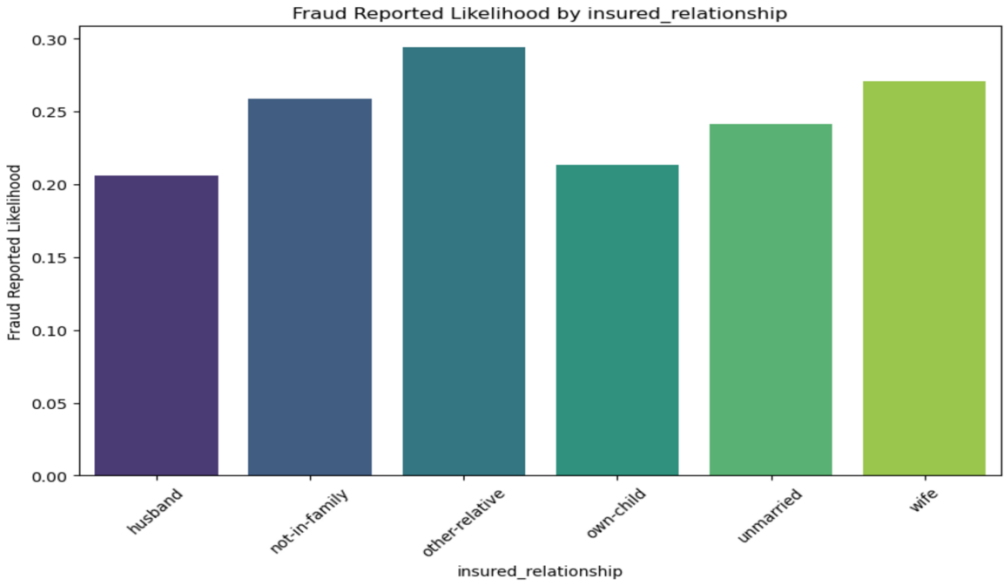


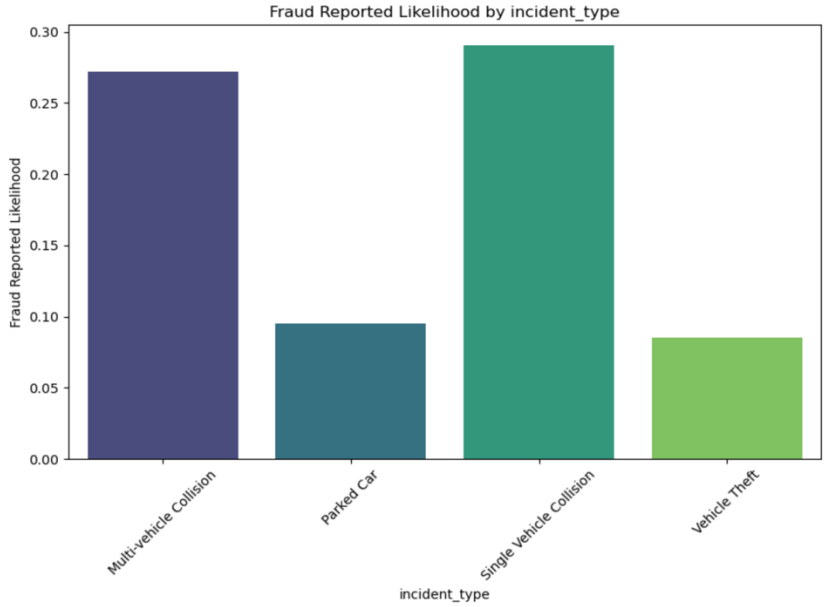
5 . EDA on validation data contd…

Bivariate Analysis



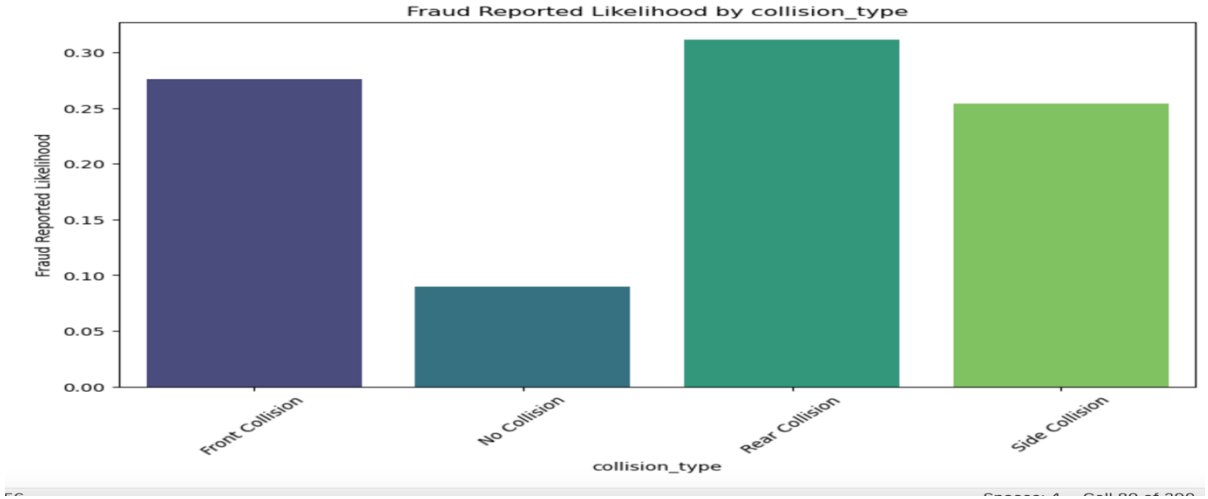


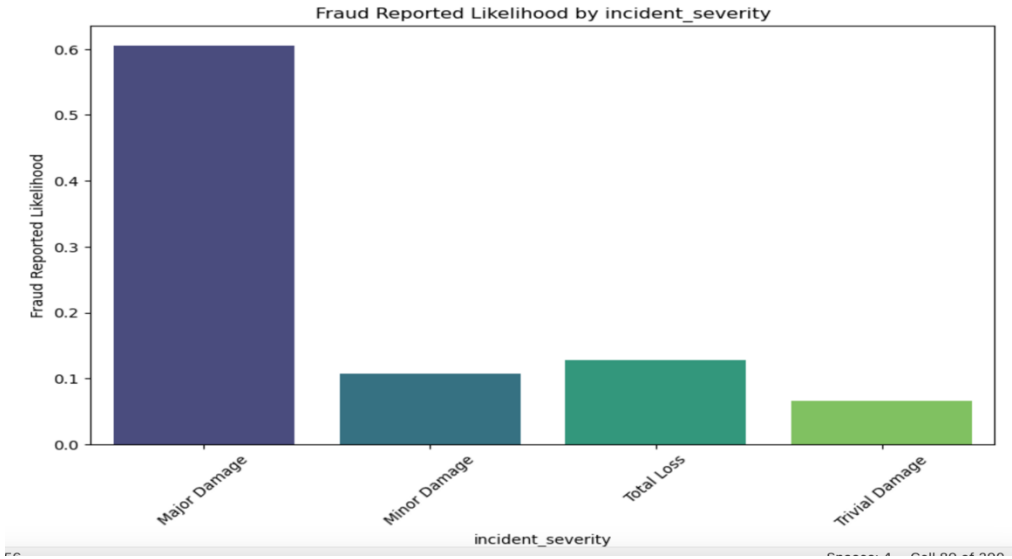




5 . EDA on validation data contd…

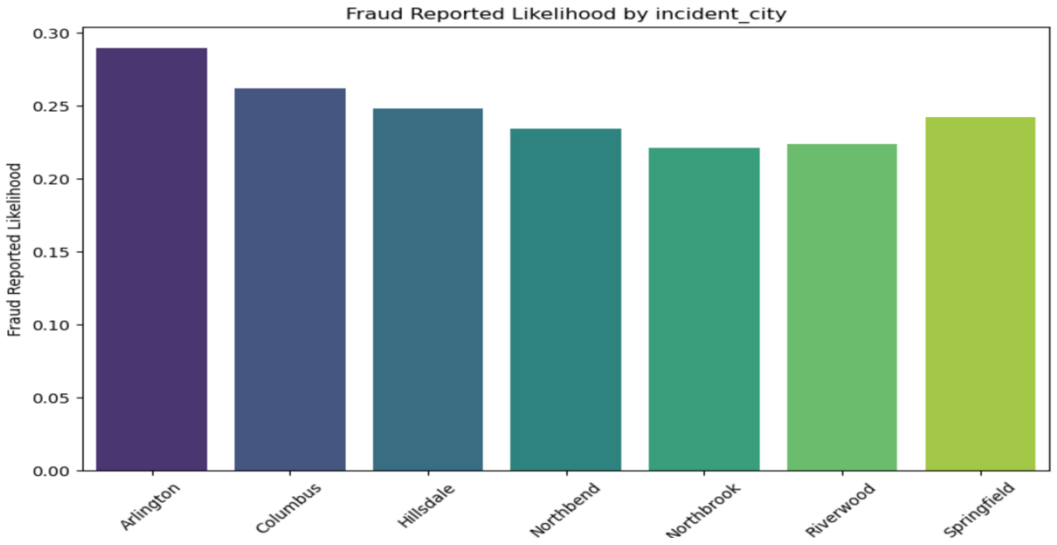
Bivariate Analysis

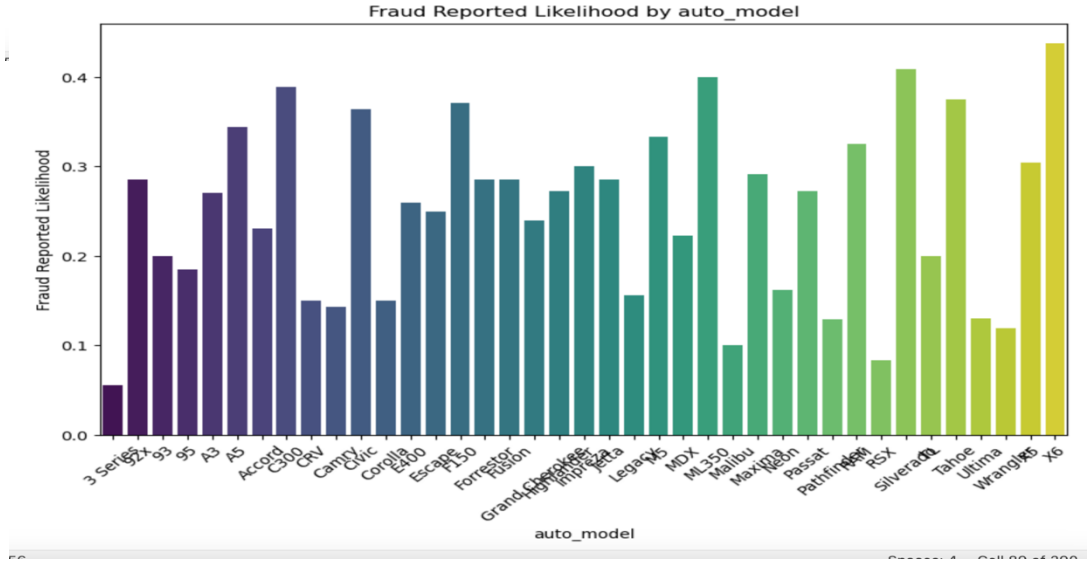


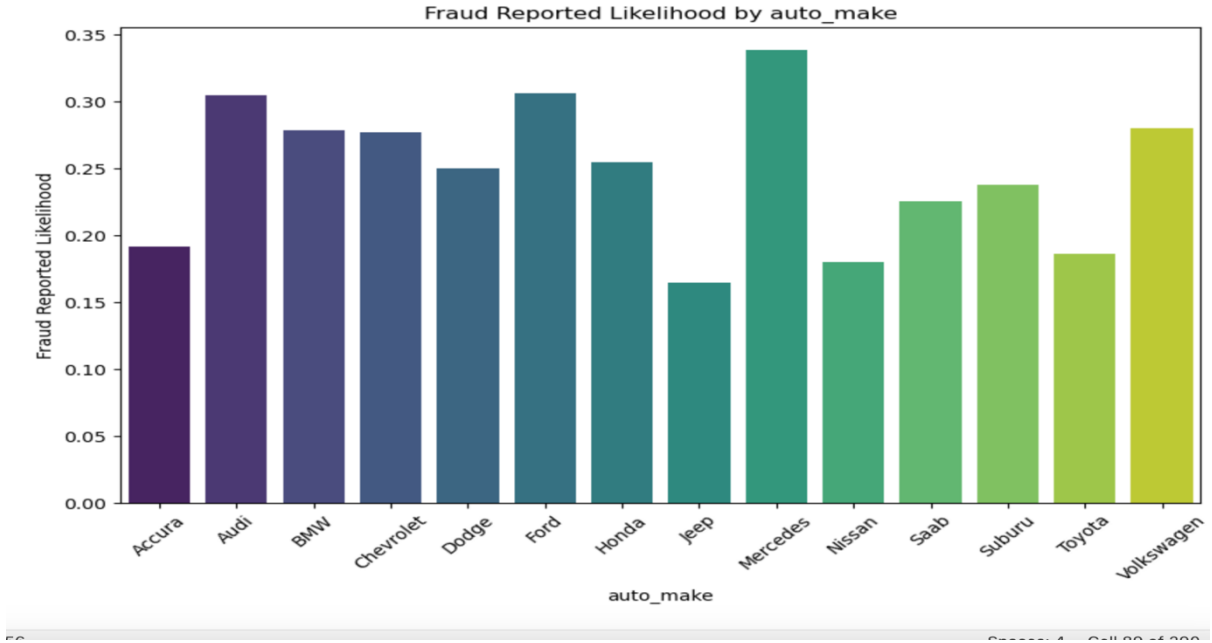


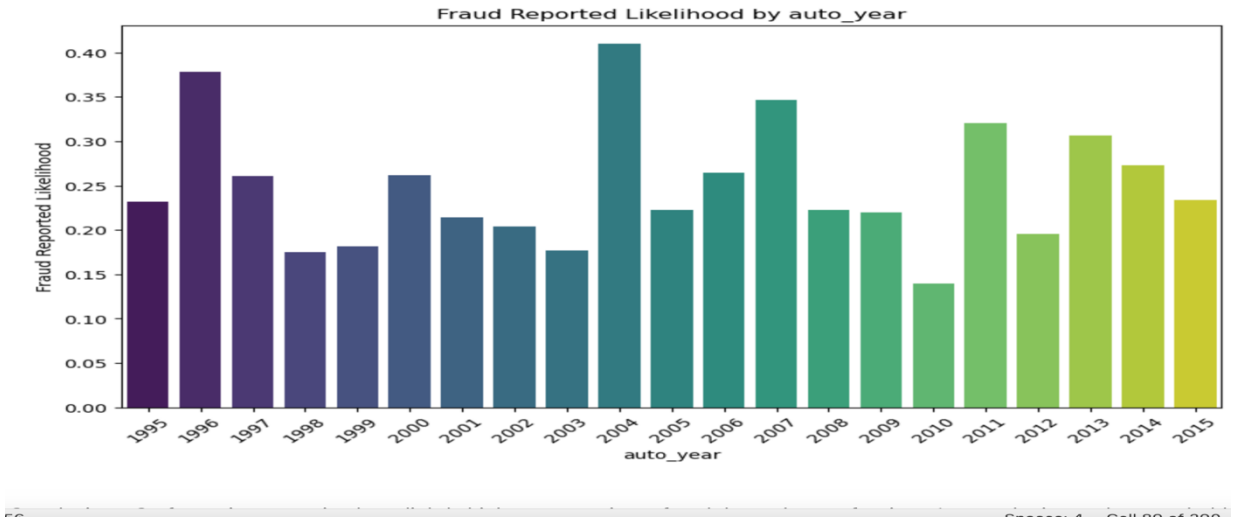
5 . EDA on validation data contd…

Bivariate Analysis









5 . EDA on validation data contd…

Bivariate Analysis

Insights from graphs :

- Craft repair occupation has slightly higher propensity to fraud than other professions

- Among the insured person hobbies -people instructed in Chess and Cross fit have observed to have more likelihood to fraud

- When the insured relative is other relative the fraud rate is higher

- Vehicle collisions are more resulting in frauds that theft

- Major damage has a much higher fraud likelihood than minor damages or total loss

- When No authorities were reported the number of frauds was also less. Implying fraudsters don't mind reporting to authorities appear authentic.

- OH state has reported more frauds than rest

6. Feature Engineering

Perform resampling : Handling class imbalance in the training data by applying resampling technique
We used the below code to perform this task:

```
# Import RandomOverSampler from imblearn library
from imblearn.over_sampling import RandomOverSampler

# Perform resampling on training data
X_train_resampled, y_train_resampled = RandomOverSampler(random_state=42).fit_resample(X_train, y_train)
```

Results :

fraud_reported
N 50.0
Y 50.0

Feature Creation :

```
# Variance of categorical features
for column in categorical_cols_train:
print(X_train_resampled[column].value_counts())
print('---------------------------------')
#There are no categorical features with high cardinality
```

We also performed other steps like combing values in categorical columns,
handled dummy variable creation.

6. Feature Engineering contd…

Feature scaling

Feature scaling was done using below code:

```
 # Importing the necessary scaling tool from scikit-learn
from sklearn.preprocessing import MinMaxScaler
# Initialize the scaler
scaler = MinMaxScaler()
# # Scale the numeric features present in the training data
scaler.fit(X_train_dummies[numerical_cols_train])
X_train_dummies[numerical_cols_train] =
scaler.fit_transform(X_train_dummies[numerical_cols_train])
X_train_dummies.head()



# Scaling the numeric features present in the validation data
X_test_dummies[numerical_cols_train]= scaler.transform(X_test_dummies[numerical_cols_train])

X_test_dummies.head()
```

7. Model Building

- We have done feature selection, retained the selected features, selected relevant features and added constant to training data and built a Logistic Regression Model

Model 1

```
print(logit_model1.summary())
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:          fraud_reported   No. Observations:                 1054
Model:                           Logit   Df Residuals:                     1003
Method:                            MLE   Df Model:                           50
Date:                 Sun, 06 Apr 2025   Pseudo R-squ.:                 0.5280
Time:                         17:38:15   Log-Likelihood:                -344.82
converged:                       True   LL-Null:                       -730.58
Covariance Type:             nonrobust   LLR p-value:                5.941e-130
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            0.2943      1.190      0.247      0.805      -2.038       2.627
months_as_customer               1.9760      1.182      1.671      0.095      -0.341       4.293
insured_hobbies_cross-fit        4.5131      0.605      7.466      0.000       3.328       5.698
insured_hobbies_golf            -0.3399      0.526     -0.647      0.518      -1.370       0.691
insured_relationship_not-in-family  0.8484  0.375      2.265      0.024       0.114       1.583
insured_relationship_other-relative 0.5975  0.365      1.639      0.101      -0.117       1.312
insured_relationship_own-child   0.2321      0.387      0.600      0.548      -0.525       0.990
insured_relationship_unmarried   0.8026      0.375      2.139      0.032       0.067       1.538
insured_relationship_wife        0.6137      0.387      1.588      0.112      -0.144       1.371
incident_type_Single Vehicle Collision  -0.7552  0.818  -0.923     0.356      -2.359       0.849
collision_type_No Collision      0.4963      1.038      0.478      0.633      -1.538       2.531
collision_type_Rear Collision    0.2239      0.280      0.799      0.424      -0.325       0.773
collision_type_Side Collision   -0.6069      0.289     -2.102      0.036      -1.173      -0.041
...
policy_annual_premium            0.3965      0.728      0.545      0.586      -1.029       1.822
number_of_vehicles_involved     -1.4009      1.169     -1.198      0.231      -3.693       0.891
policy_csl_250/500               0.6869      0.257      2.671      0.008       0.183       1.191
================================================================================================
```

Model Interpretation

- The output summary table provides the features used for building model along with coefficient of each of the feature and their p-value. Using p-value in a logistic regression model we assessed the statistical significance of each coefficient. As we can see lesser the p-value, more significant the feature is in the model.

7. Model Building contd…

Model 2

```
[211]  print(logit_model2.summary())
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:          fraud_reported   No. Observations:                 1054
Model:                          Logit   Df Residuals:                     1004
Method:                           MLE   Df Model:                           49
Date:                Sun, 06 Apr 2025   Pseudo R-squ.:                  0.5274
Time:                        17:38:35   Log-Likelihood:                -345.24
converged:                       True   LL-Null:                       -730.58
Covariance Type:            nonrobust   LLR p-value:                2.215e-130
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         -0.3937      0.929     -0.424      0.672      -2.215       1.428
months_as_customer             1.9982      1.179      1.694      0.090      -0.314       4.310
insured_hobbies_cross-fit      4.4569      0.597      7.463      0.000       3.286       5.627
insured_hobbies_golf          -0.3417      0.525     -0.651      0.515      -1.371       0.688
insured_relationship_not-in-family  0.8497  0.375     2.269      0.023       0.116       1.584
insured_relationship_other-relative 0.5709  0.363     1.575      0.115      -0.140       1.281
insured_relationship_own-child 0.2502      0.385      0.650      0.516      -0.505       1.005
insured_relationship_unmarried 0.8182      0.374      2.186      0.029       0.085       1.552
insured_relationship_wife      0.5947      0.386      1.542      0.123      -0.161       1.351
collision_type_No Collision    1.2421      0.656      1.893      0.058      -0.044       2.528
collision_type_Rear Collision  0.2221      0.280      0.794      0.427      -0.326       0.770
collision_type_Side Collision -0.5994      0.288     -2.081      0.037      -1.164      -0.035
incident_severity_Minor Damage -4.4099     0.348    -12.675      0.000      -5.092      -3.728
...
policy_annual_premium          0.3440      0.725      0.475      0.635      -1.077       1.765
number_of_vehicles_involved   -0.3660      0.336     -1.090      0.276      -1.024       0.292
policy_csl_250/500             0.6868      0.257      2.673      0.008       0.183       1.190
==============================================================================================
```

7. Model Building contd…

Model 3

```
print(logit_model3.summary())
```

```
         Current function value: 0.329027
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:         fraud_reported   No. Observations:                 1054
Model:                          Logit   Df Residuals:                     1005
Method:                           MLE   Df Model:                           48
Date:                Sun, 06 Apr 2025   Pseudo R-squ.:                 0.5253
Time:                        17:41:51   Log-Likelihood:                -346.79
converged:                       True   LL-Null:                       -730.58
Covariance Type:            nonrobust   LLR p-value:                2.363e-130
==========================================================================================================
                                             coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                     -0.6250      0.917     -0.682      0.495      -2.422       1.172
months_as_customer                         0.1168      0.452      0.259      0.796      -0.769       1.002
insured_hobbies_cross-fit                  4.4087      0.595      7.405      0.000       3.242       5.576
insured_hobbies_golf                      -0.3096      0.524     -0.591      0.555      -1.337       0.718
insured_relationship_not-in-family         0.8140      0.372      2.189      0.029       0.085       1.543
insured_relationship_other-relative        0.5810      0.360      1.613      0.107      -0.125       1.287
insured_relationship_own-child             0.2744      0.383      0.716      0.474      -0.477       1.025
insured_relationship_unmarried             0.7937      0.372      2.132      0.033       0.064       1.523
insured_relationship_wife                  0.5802      0.386      1.505      0.132      -0.176       1.336
collision_type_No Collision                1.1997      0.653      1.837      0.066      -0.080       2.480
collision_type_Rear Collision              0.2378      0.278      0.854      0.393      -0.308       0.783
collision_type_Side Collision             -0.5834      0.287     -2.031      0.042      -1.146      -0.020
incident_severity_Minor Damage            -4.3937      0.346    -12.683      0.000      -5.073      -3.715
incident_severity_Total Loss              -3.8377      0.301    -12.741      0.000      -4.428      -3.247
authorities_contacted_Fire                 0.6088      0.310      1.966      0.049       0.002       1.216
authorities_contacted_Other                0.3734      0.318      1.174      0.240      -0.250       0.997
```

```
print(logit_model3.summary())
```

```
incident_state_WV                         -0.8893      0.314     -2.828      0.005      -1.506      -0.273
incident_city_Columbus                    -0.1321      0.299     -0.442      0.658      -0.718       0.454
incident_city_Springfield                  0.3203      0.293      1.094      0.274      -0.254       0.894
auto_make_Dodge                            0.0617      0.414      0.149      0.881      -0.749       0.872
auto_make_Mercedes                        -0.0760      0.423     -0.180      0.857      -0.904       0.752
insured_hobbies_chess                      5.6489      0.593      9.520      0.000       4.486       6.812
insured_hobbies_camping                   -1.2797      0.565     -2.266      0.023      -2.387      -0.173
auto_year_Old                             -0.3499      0.241     -1.453      0.146      -0.822       0.122
witnesses                                  0.7129      0.298      2.392      0.017       0.129       1.297
insured_sex_MALE                          -0.1815      0.218     -0.832      0.405      -0.609       0.246
policy_deductable                          0.4899      0.257      1.909      0.056      -0.013       0.993
policy_state_OH                            0.1624      0.255      0.636      0.525      -0.338       0.663
insured_education_level_JD                 0.4390      0.291      1.506      0.132      -0.132       1.010
policy_state_IN                            0.0488      0.267      0.183      0.855      -0.474       0.572
total_claim_amount                         1.2825      0.824      1.556      0.120      -0.332       2.898
police_report_available                   -0.1013      0.232     -0.437      0.662      -0.556       0.353
insured_occupation_Service/Specialized    -0.2320      0.265     -0.876      0.381      -0.751       0.287
bodily_injuries                           -0.1588      0.256     -0.621      0.535      -0.660       0.343
policy_csl_500/1000                       -0.1320      0.275     -0.479      0.632      -0.672       0.408
property_damage                            0.4712      0.224      2.105      0.035       0.033       0.910
incident_hour_of_the_day                  -0.2631      0.373     -0.705      0.481      -0.995       0.468
capital-loss                              -0.2548      0.400     -0.637      0.524      -1.039       0.529
insured_education_level_MD                 0.5453      0.327      1.669      0.095      -0.095       1.186
capital-gains                              0.6415      0.389      1.651      0.099      -0.120       1.403
umbrella_limit                             1.0454      0.497      2.105      0.035       0.072       2.019
insured_education_level_Masters            0.0712      0.307      0.232      0.817      -0.531       0.673
insured_occupation_Engineering            -0.0690      0.243     -0.284      0.776      -0.544       0.406
policy_annual_premium                      0.3916      0.725      0.540      0.589      -1.030       1.813
number_of_vehicles_involved               -0.3418      0.335     -1.020      0.308      -0.999       0.315
policy_csl_250/500                         0.6819      0.256      2.659      0.008       0.179       1.184
```

7. Model Building contd…

Model performance evaluation :

Evaluating VIFs of features to assess multicollinearity – As per the data we see that VIF is less than 5 so we will use logit_model3
-Accuracy of the model = 0.8785578747628083
-Confusion matrix based on the predictions made on the training data
[[453 74]
[54 473]]

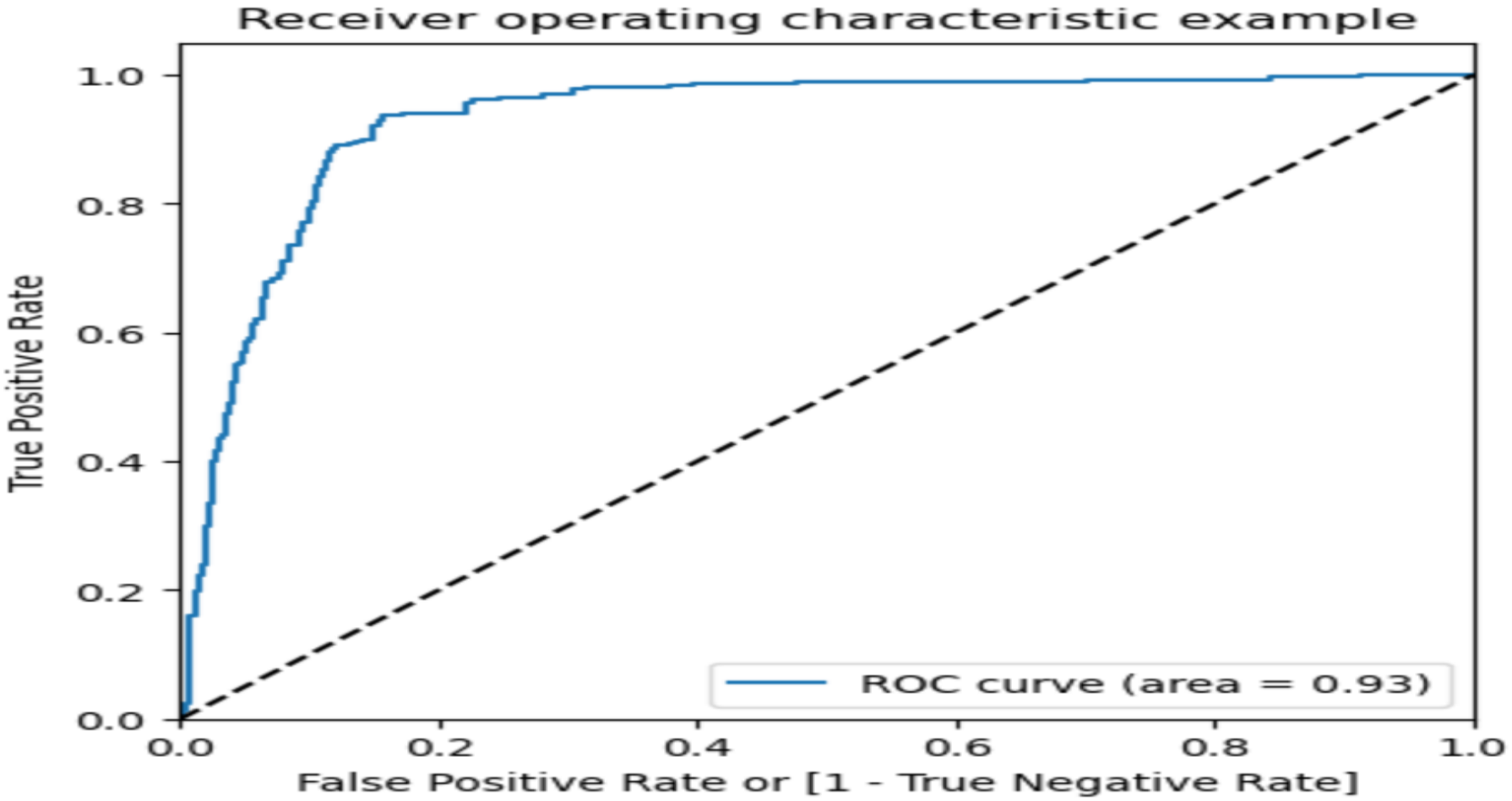
-Variables for true positive, true negative, false positive and false negative
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

-Calculating sensitivity, specificity, precision, recall and F1-score
Sensitivity 0.8975332068311196
Specificity 0.8595825426944972
Precision 0.8647166361974405
Recall 0.8975332068311196
F1 Score 0.8808193668528864

7. Model Building contd...

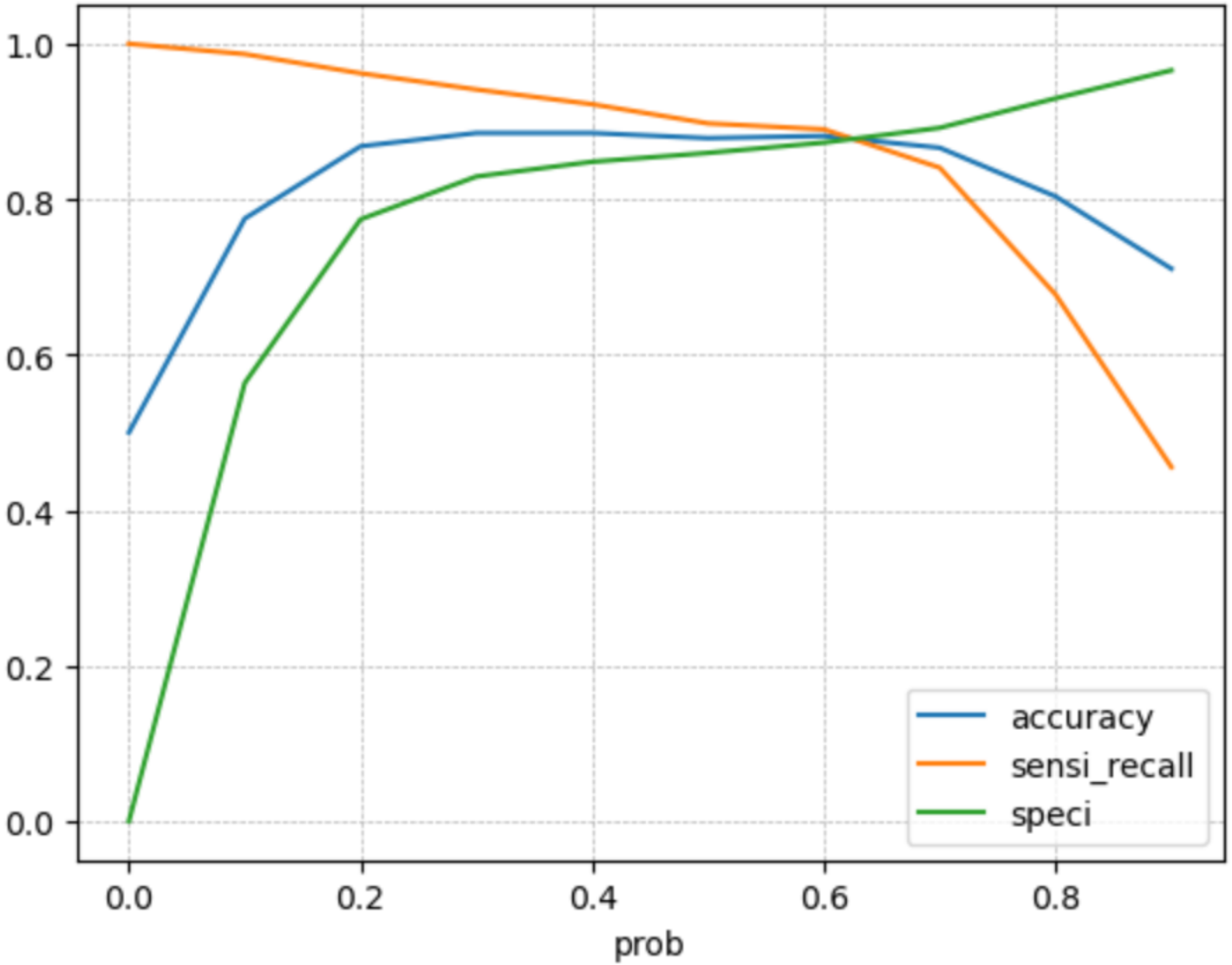
Finding the Optimal Cutoff

Plotting ROC Curve to visualise the trade-off between true positive rate and false positive rate across different classification thresholds



7. Model Building contd...

Plotting accuracy, sensitivity, and specificity at different values of probability cutoffs



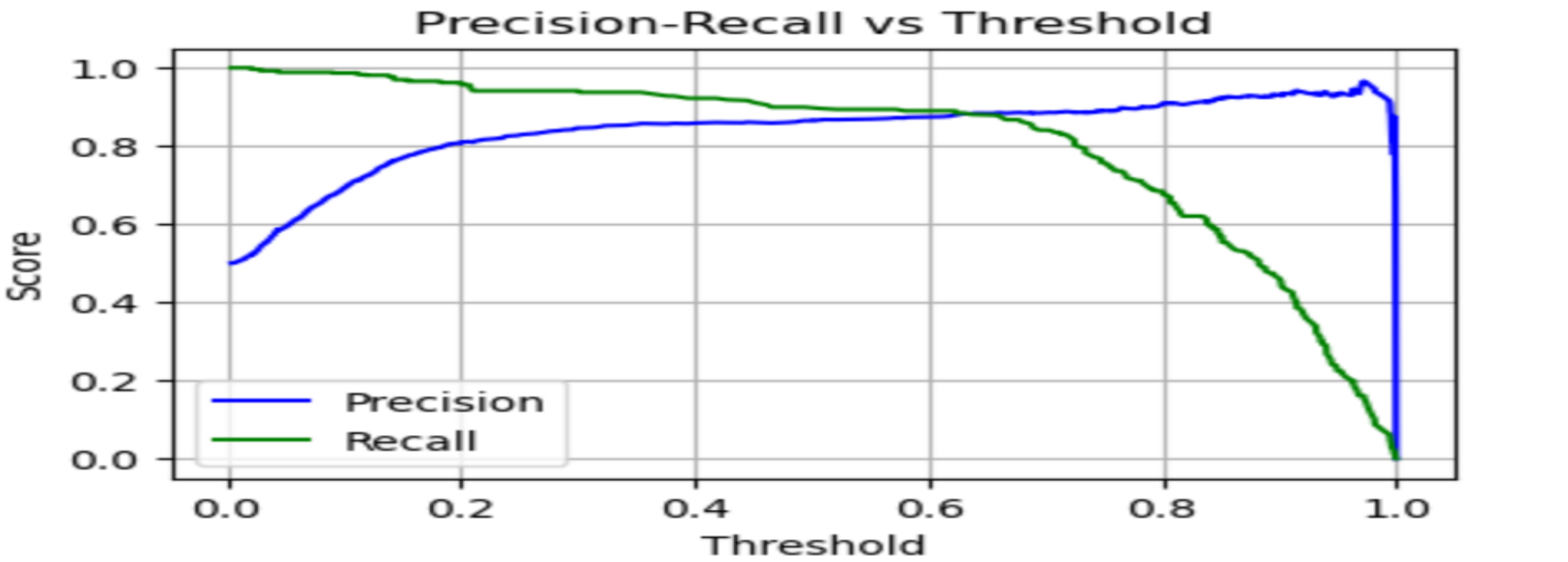
	prob	accuracy	sensi_recall	speci	precision
0.0	0.0	0.500000	1.000000	0.000000	0.500000
0.1	0.1	0.775142	0.986717	0.563567	0.693333
0.2	0.2	0.868121	0.962049	0.774194	0.809904
0.3	0.3	0.885199	0.941176	0.829222	0.846416
0.4	0.4	0.885199	0.922201	0.848197	0.858657
0.5	0.5	0.878558	0.897533	0.859583	0.864717
0.6	0.6	0.881404	0.889943	0.872865	0.875000
0.7	0.7	0.866224	0.840607	0.891841	0.886000
0.8	0.8	0.803605	0.677419	0.929791	0.906091
0.9	0.9	0.710626	0.455408	0.965844	0.930233

7. Model Building contd...
Final prediction based on optimal cutoff

- Accuracy = 0.881404174573055

- Creating the confusion matrix once again
[[460 67]
[58 469]]

- Plotting precision Recall curve



7. Model Building contd...

Building Random Forest Model

After performing certain analysis we find that :

Accuracy on training data = 1.0

Confusion matrix is
[[527 0]
[0 527]]

Sensitivity 1.0 Specificity 1.0 Precision 1.0 Recall 1.0 F1 Score 1.0

Cross-validation scores: [0.93838863 0.92890995 0.93838863 0.93364929 0.91428571]

Mean cross-validation score: 0.9307244414353419

Standard Deviation: 0.0089

	precision	recall	f1-score	support
0	1.00	1.00	1.00	527
1	1.00	1.00	1.00	527
accuracy			1.00	1054
macro avg	1.00	1.00	1.00	1054
weighted avg	1.00	1.00	1.00	1054

ROC AUC score: 1.0

Model is overfitting as we see ROC AUC Score as 1 and all there is no false negative and false positive

7. Model Building contd...

Hyperparameter Tuning

By making predictions on the training data we get :

Best Model Accuracy: 0.9269

Confusion matrix

[[471 56]
[21 506]]

Calculating sensitivity, specificity, precision, recall and F1-score of the model

Sensitivity 0.9601518026565465
Specificity 0.8937381404174574
Precision 0.900355871886121
Recall 0.9601518026565465
F1 Score 0.9292929292929293

8 . Prediction and Model Evaluation

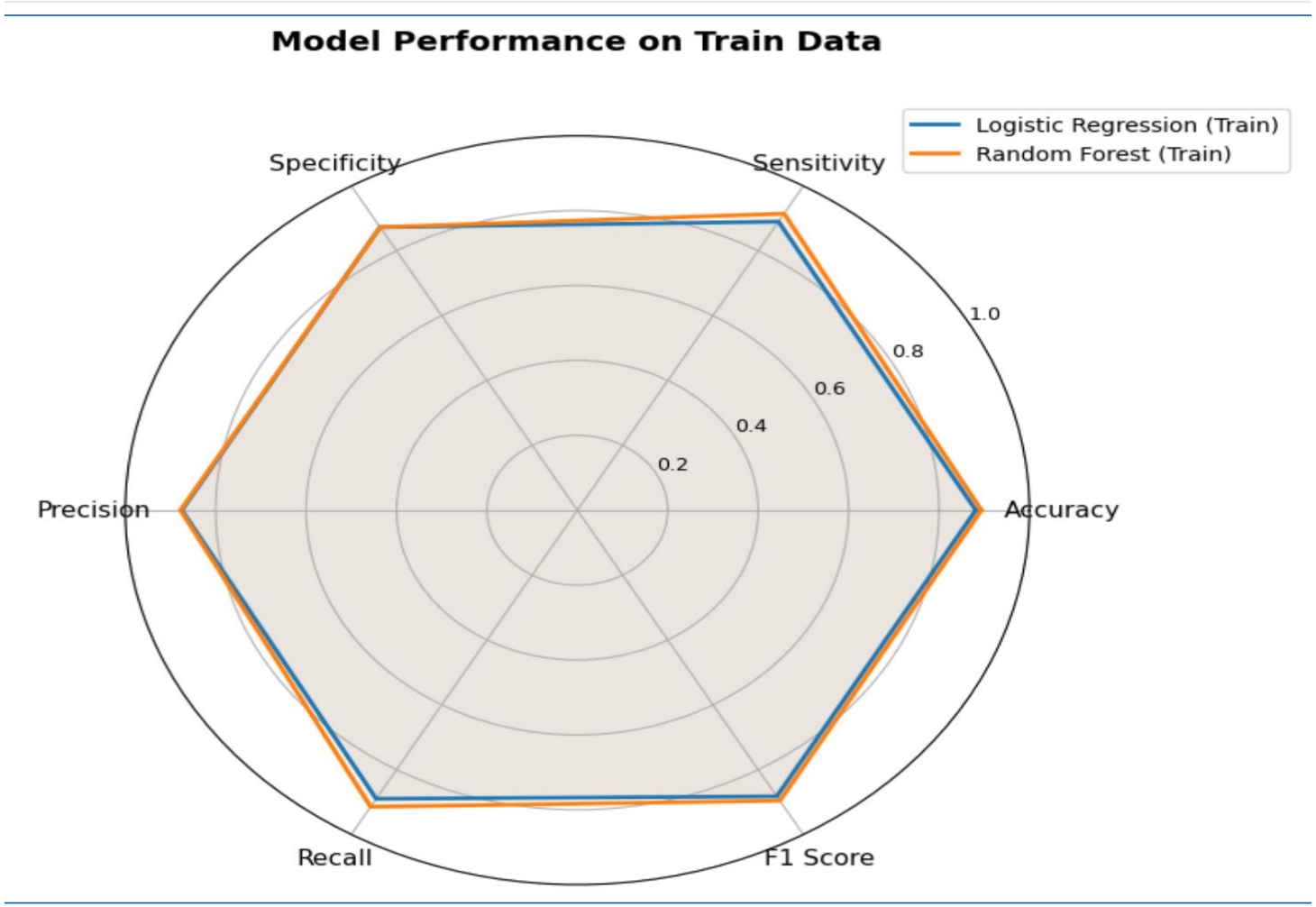
Make predictions over validation data using logistic regression model
accuracy = 0.8533333333333334
confusion matrix
[[201 25]
[19 55]]

Calculating sensitivity, specificity, precision, recall and f1 score of the model
Sensitivity 0.7432432432432432
Specificity 0.8893805309734514
Precision 0.6875
Recall 0.7432432432432432
F1 Score 0.7142857142857143

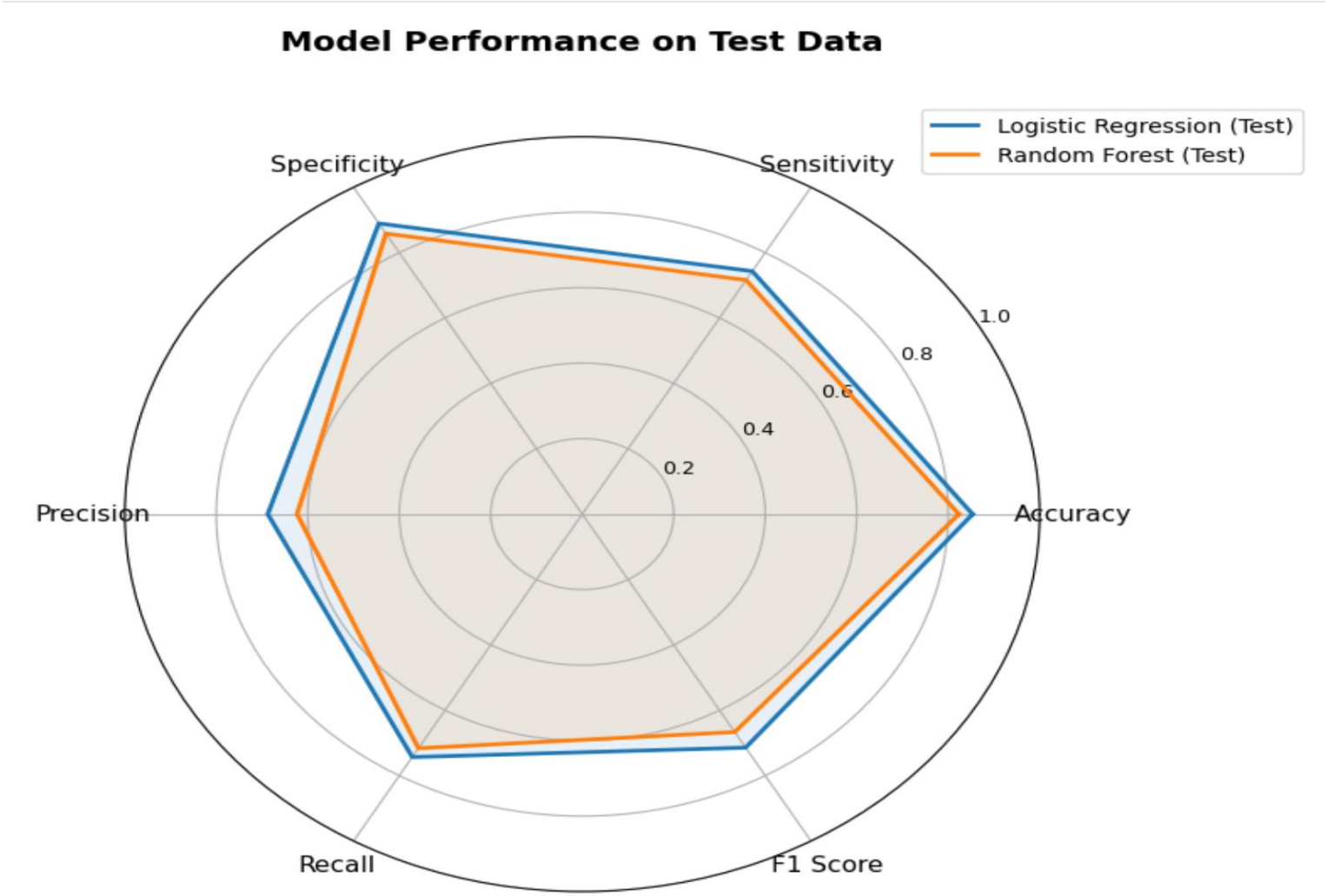
Making predictions over validation data using random forest model
Accuracy on test data: 0.8233333333333334
confusion matrix
[[194 32]
[21 53]]

Calculating sensitivity, specificity, precision, recall and F1-score of the model
Sensitivity 0.7162162162162162
Specificity 0.8584070796460177
Precision 0.6235294117647059
Recall 0.7162162162162162
F1 Score 0.6666666666666666

8 . Prediction and Model Evaluation contd…



8 . Prediction and Model Evaluation contd…



Evaluation and Conclusion

Findings from Exploratory Data Analysis (EDA)

- Individuals employed in craft repair occupations exhibit a slightly higher propensity for engaging in fraudulent activities compared to other professions.

- Among the insured individuals, those with hobbies such as chess and crossfit show a higher likelihood of fraud, suggesting a possible correlation between certain recreational interests and fraudulent behaviour.

- When the insured person's relationship to the claimant is listed as "other relative," the incidence of fraud is notably higher.

- Vehicle collisions are associated with a higher frequency of fraud compared to vehicle thefts.

- Claims involving major damage demonstrate a significantly greater likelihood of being fraudulent than those classified as minor damage or total loss.

- Interestingly, when no authorities were reported, the number of frauds was also lower. This may indicate that fraudsters often choose to involve authorities to appear legitimate and avoid suspicion.

- The state of Ohio (OH) reported a higher number of fraudulent claims compared to other states.

Evaluation and Conclusion contd..

Findings from Modelling :

-The performance of both logistic regression and random forest are equally good.

- The balance between precision and recall is crucial in fraud detection datasets.

- While high recall ensures that most fraudulent activities are identified, high precision minimizes false positives, thereby reducing the cost and effort associated with investigating non-fraudulent alerts.