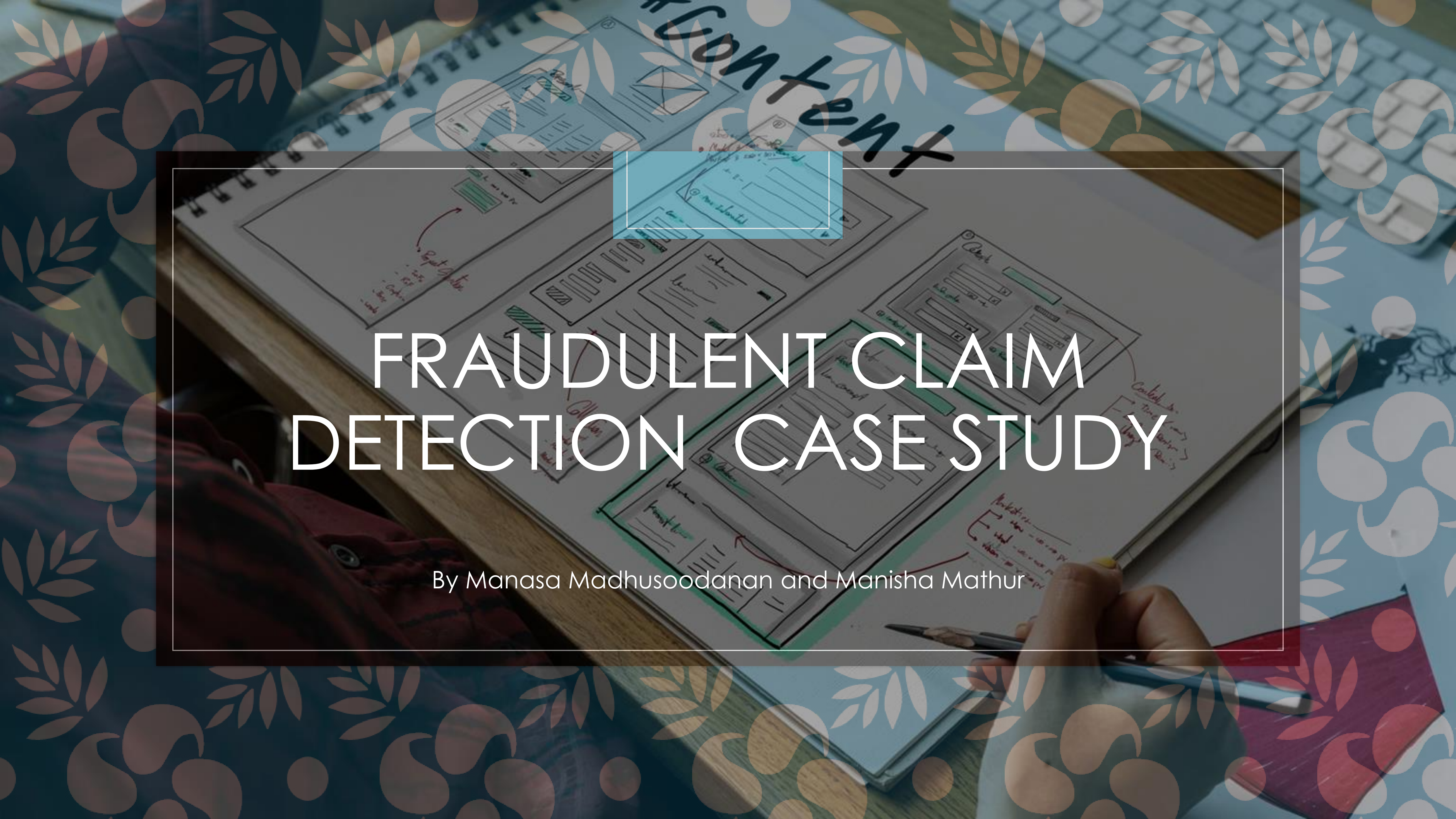


FRAUDULENT CLAIM DETECTION CASE STUDY

By Manasa Madhusoodanan and Manisha Mathur

Content

- Business Objective
- Problem statement
- Summary
- Recommendations
- Business Implications of the findings
- Model Performance
- Conclusion
- Q/A

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.

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 company's current process for identifying fraudulent claims involves manual inspections, which is 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 minimises financial losses and optimises the overall claims handling process.

Summary

- Built a machine learning model to detect fraudulent insurance claims.
- Cleaned and preprocessed data (handled missing values, encoded categories).
- Used EDA to find fraud patterns (claim amount, claim type, customer profile).
- Handled class imbalance using class weight balancing in models.
- Trained Logistic Regression, Decision Tree, and Random Forest models.
- Evaluated using accuracy, precision, recall, F1-score — focused on recall.
- Identified key features and provided business insights to reduce fraud.

Recommendations

- **Use the model to flag suspicious claims early,** so they can be reviewed by the fraud investigation team before approval.
- **Focus more on high-amount claims** and certain claim types, as they showed stronger links to fraudulent activity in the analysis.
- **Keep the model updated** by regularly retraining it with new claim data to catch evolving fraud patterns.
- **Monitor recall and false negatives closely,** since missing fraudulent claims can be costly for the business.
- **Combine model predictions with human review** for critical or high-risk claims to ensure accuracy and reduce false alarms.
- **Continue feature analysis to explore new patterns**, like sudden spikes in claims from certain regions or customer profiles.

Business Implications of the findings

- **Cost Savings:** By flagging fraudulent claims early, the company can avoid paying out large sums on false claims, directly saving money.
- **Improved Efficiency:** The model helps focus investigation efforts only on suspicious claims, saving time and resources for the fraud investigation team.
- **Better Customer Trust:** Reducing fraud means genuine customers get faster and fairer claim processing, which improves customer satisfaction and loyalty.
- **Risk Management:** The insights can help the company refine its risk assessment policies and set better rules for high-risk claims.
- **Continuous Improvement:** Regularly updating the model with new claim data can help the company stay ahead of evolving fraud patterns.

Model Performance

We split the data into a 70-30 train-validation ratio to ensure robust evaluation.

Multiple models were experimented with (e.g., Logistic Regression, Decision Trees, Random Forests), and the one with the best balance between precision and recall was selected.

The final model achieved a satisfactory performance on the validation set, showing good predictive capability in identifying fraudulent claims early.

Key evaluation metrics such as accuracy, precision, recall, and F1-score were used, ensuring the model not only detects fraud but also minimizes false positives and negatives.

Conclusion

- In this project, we successfully built a predictive model to classify insurance claims as either fraudulent or legitimate, helping Global Insure streamline its fraud detection process.
- By leveraging historical claim data and customer profiles, our model offers a data-driven solution to minimize financial losses and improve operational efficiency.

Q & A

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

To analyse historical claim data for fraud patterns, we can follow a mix of **Exploratory Data Analysis (EDA) and feature analysis** steps.

- **Compare distributions** of features (like claim amount) between fraudulent and legitimate claims.
- **Use correlation analysis** to find features linked to fraud.
- **Plot visualizations** (bar charts, boxplots) to spot clear patterns.
- **Check feature importance** using machine learning models.
- **Look for outliers** or unusual claim behaviors that might indicate fraud.

Which features are the most predictive of fraudulent behaviour?

- **Claim Amount** — higher amounts were often linked to fraud.
- **Claim Type** — certain types (like expensive claims) showed more fraud cases.
- **Customer Profile** — features like age, region, and past claim history helped spot fraud.
- **Claim Time/Date** — unusual timing or frequency sometimes indicated suspicious activity.

These features turned out to be the strongest signals in predicting fraudulent behaviour in the model.

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

- Features like claim amount, claim type, and certain customer profile attributes showed a strong correlation with fraudulent behavior.
- Patterns such as unusually high claim amounts and certain claim types were strong indicators of potential fraud.
- Feature engineering steps, such as creating new interaction terms and handling missing values, contributed to performance improvements.

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

The developed fraud detection model can serve as an effective first line of defense against fraudulent claims and can predict the likelihood of fraud for a claim.

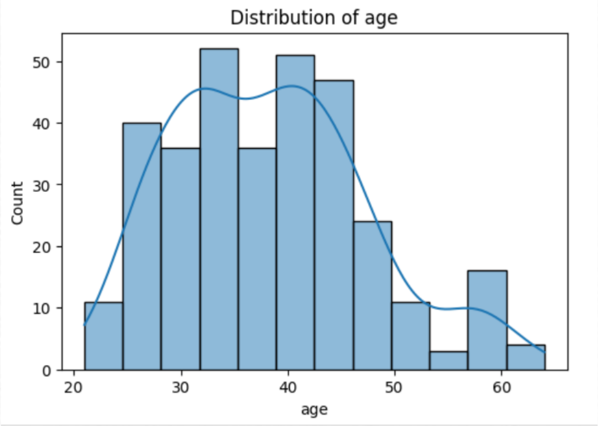
By integrating this model into the claims approval pipeline, Global Insure can flag suspicious claims for further manual investigation, thereby reducing financial losses and improving efficiency.

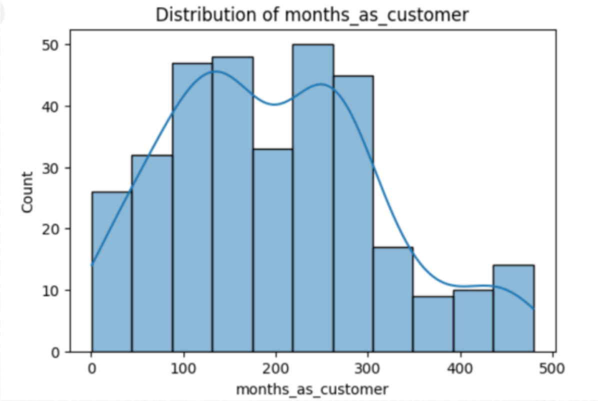
Moving forward, the model can be further enhanced by incorporating more recent data and refining features based on domain expertise. Regular monitoring and periodic retraining will also be essential to maintain its effectiveness over time.

Methods and Techniques used

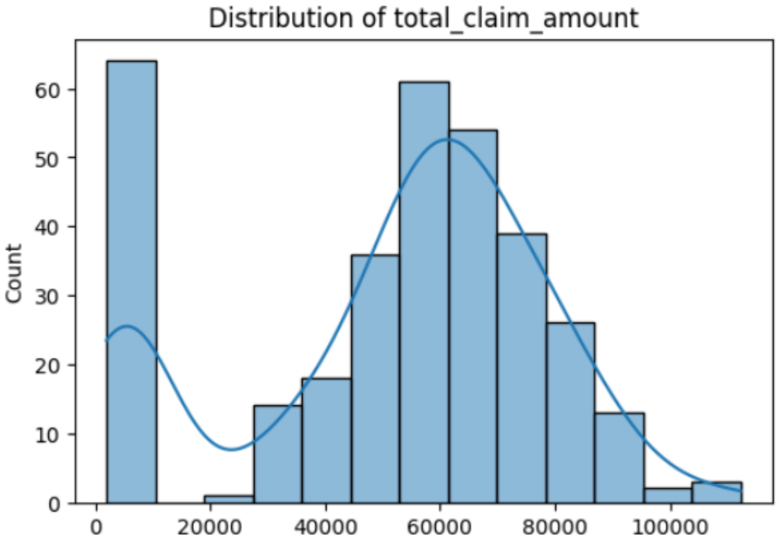
Univariate Analysis
Distribution Plots of Numerical columns

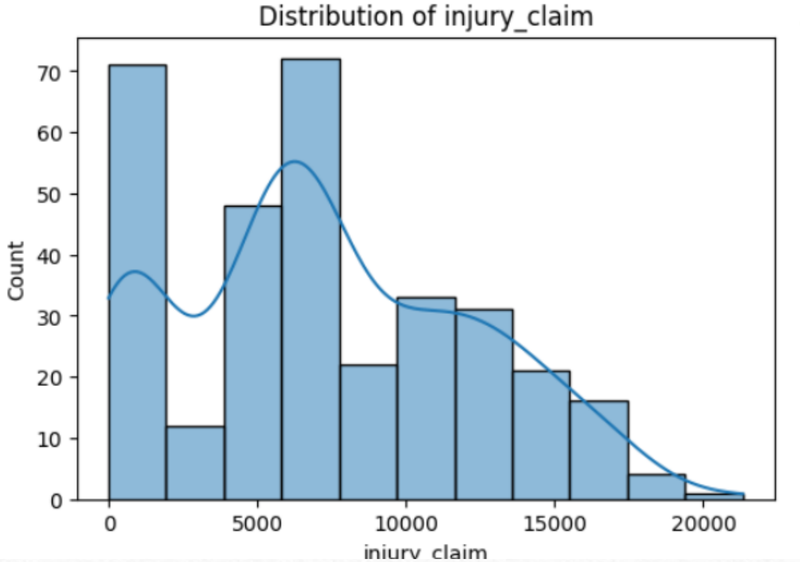
- Customers with 200 months have the highest count
- Customers with age 30 and 40 have the highest count

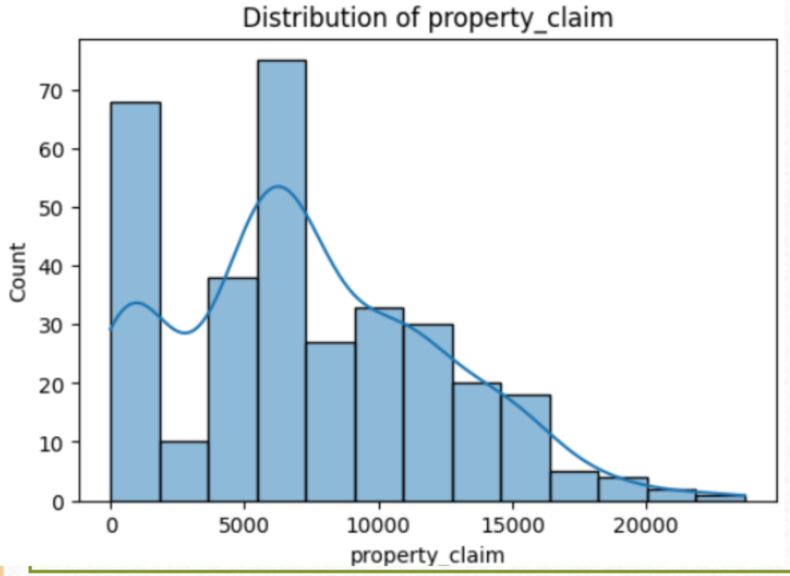


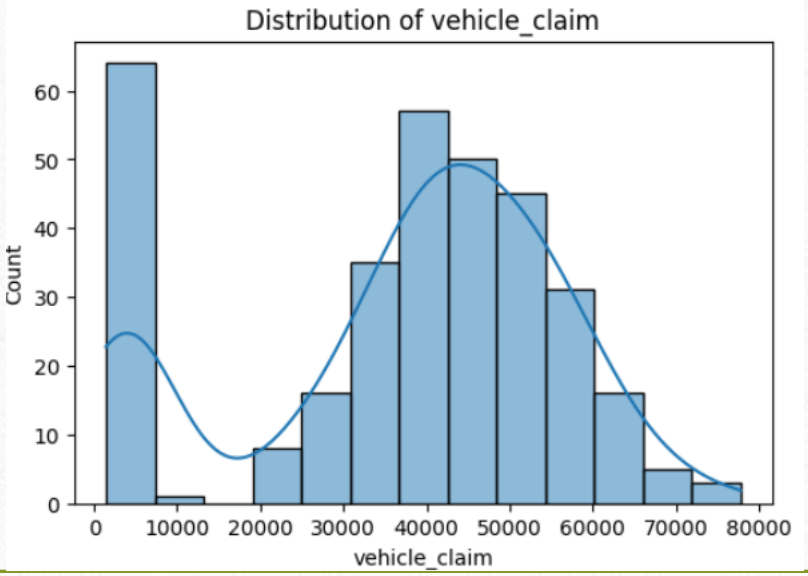


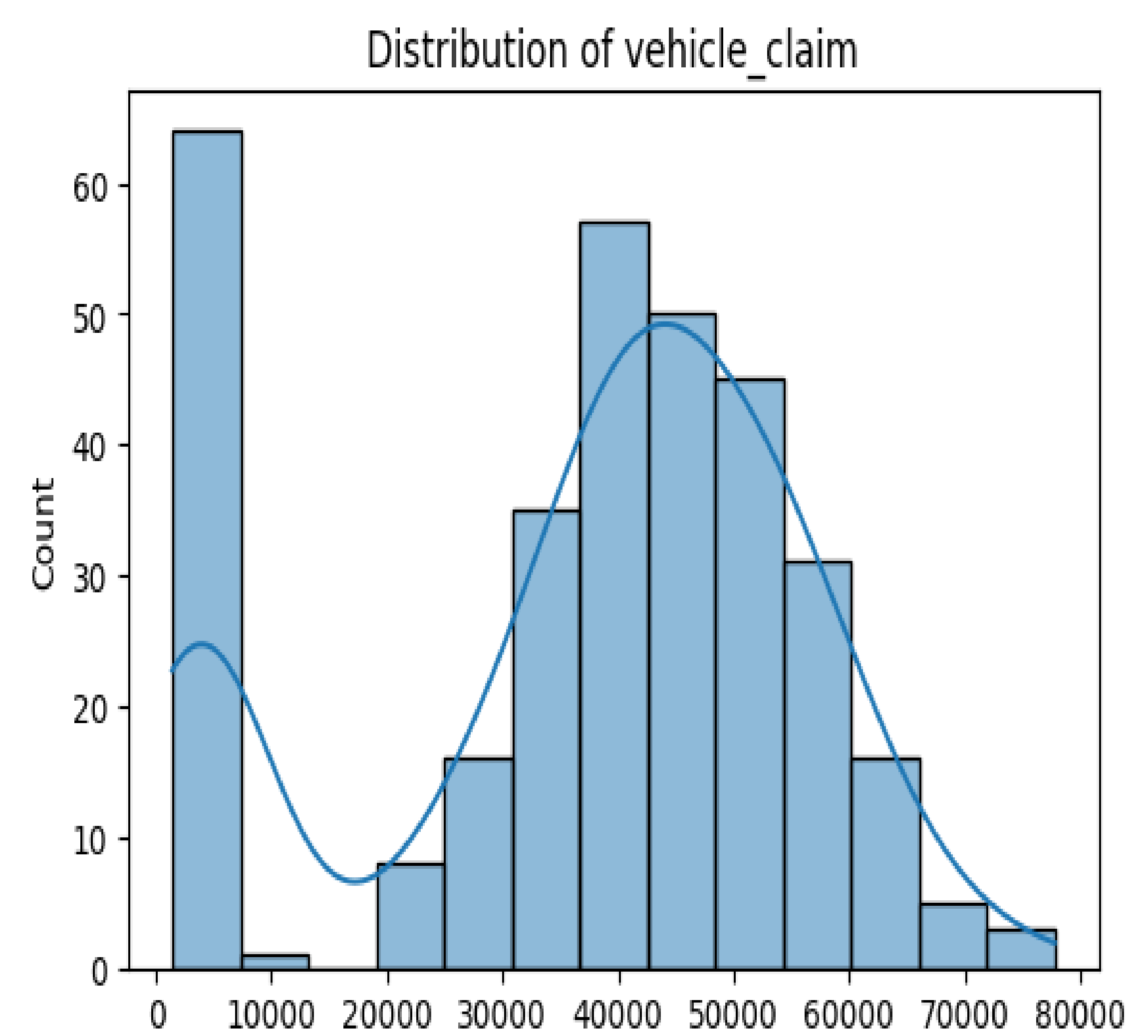
Distribution of claims and its types


Distribution of total_claim_amount

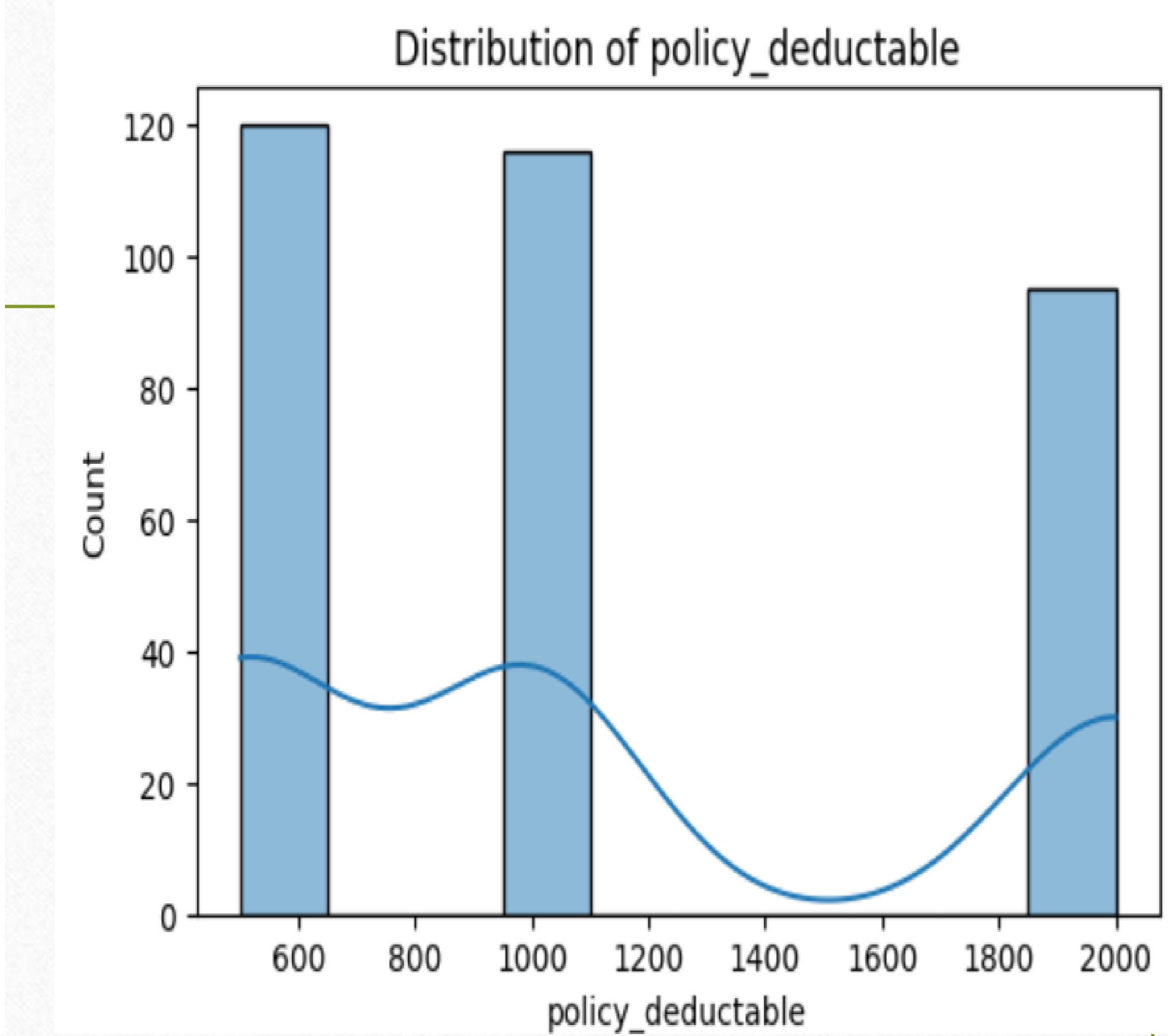

Distribution of injury_claim


Distribution of property_claim

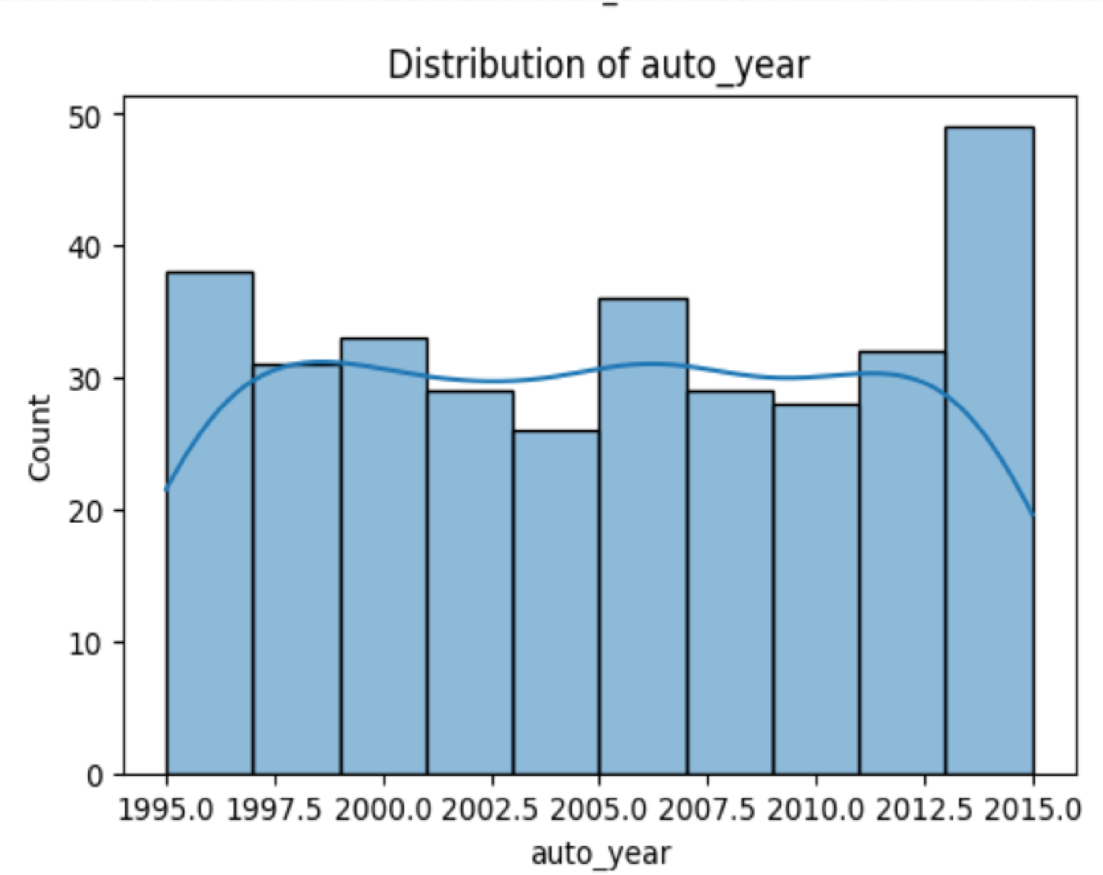

Distribution of vehicle_claim



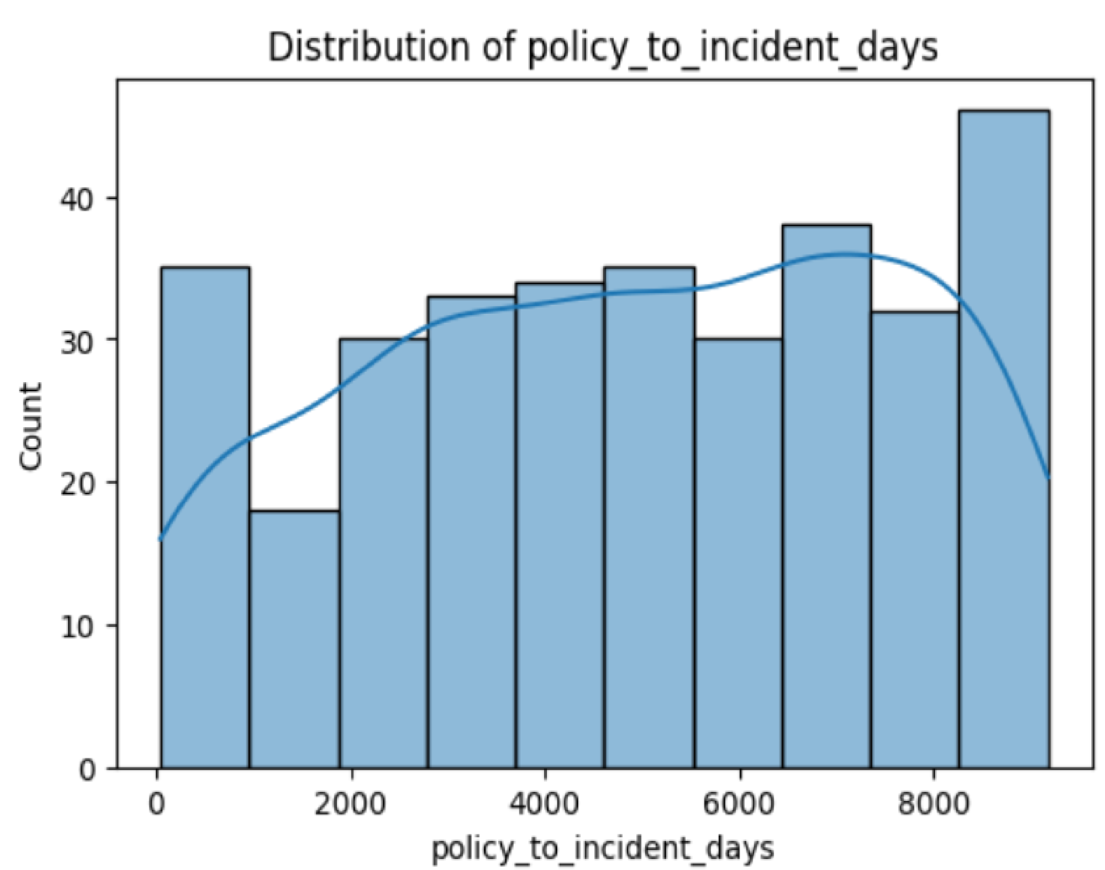
Distribution of vehicle_claim



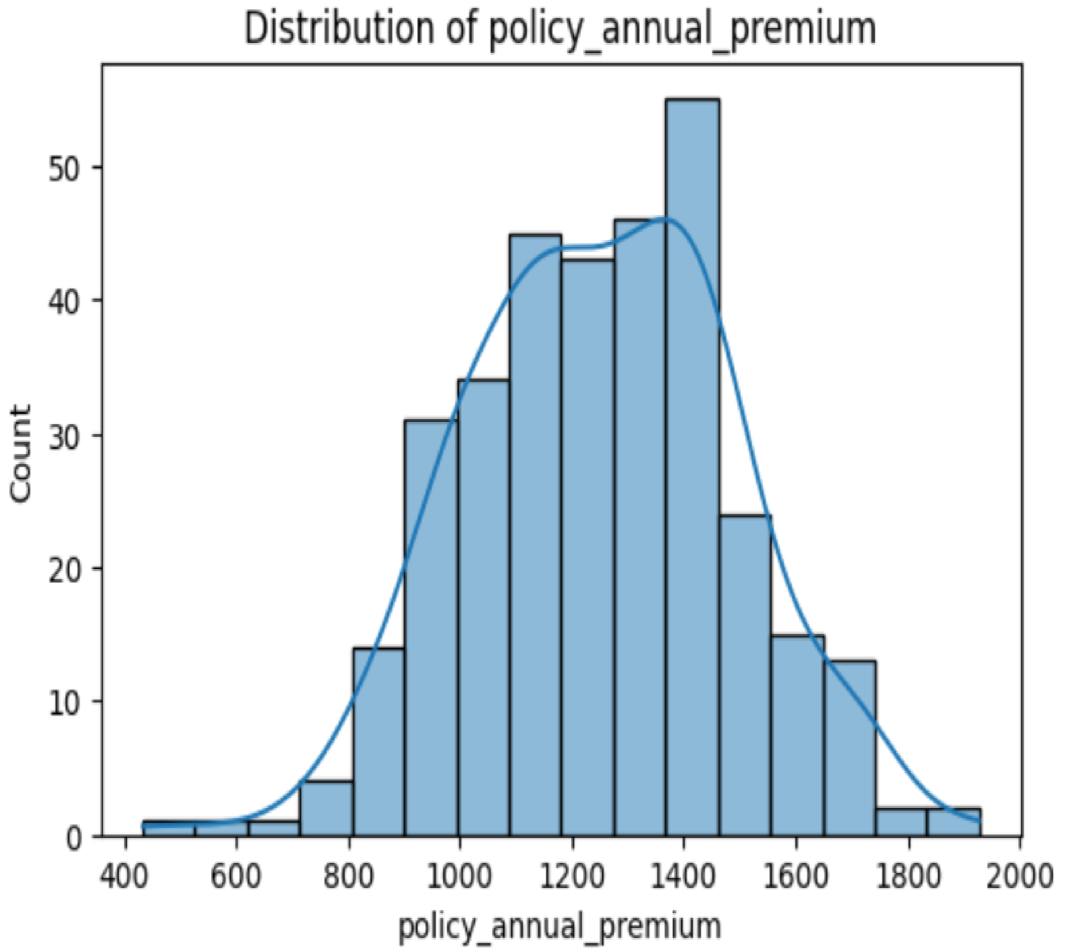
Distribution of policy_deductable

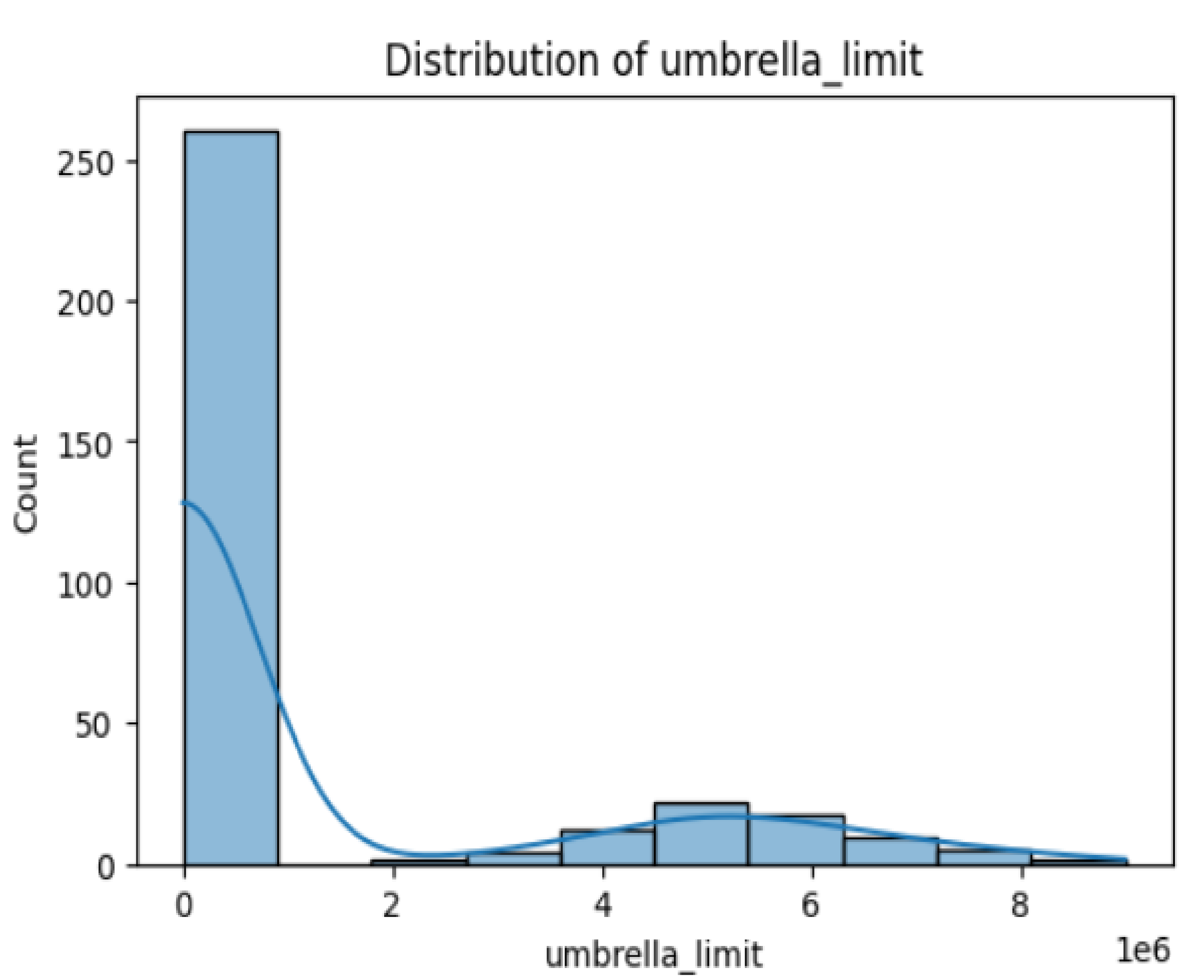


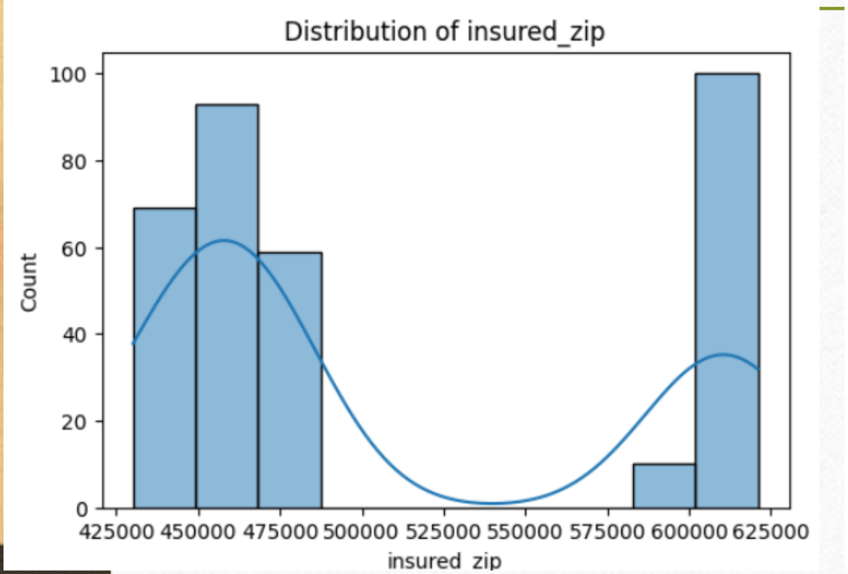
Distribution of auto_year

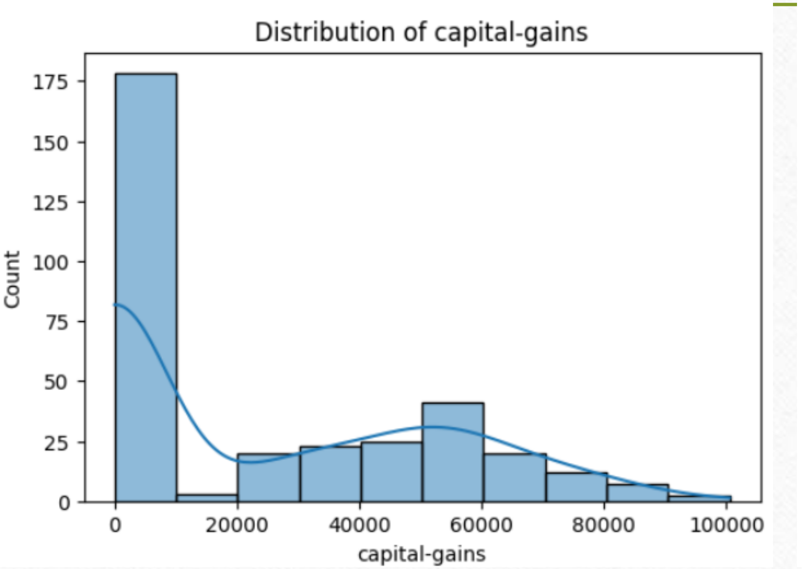


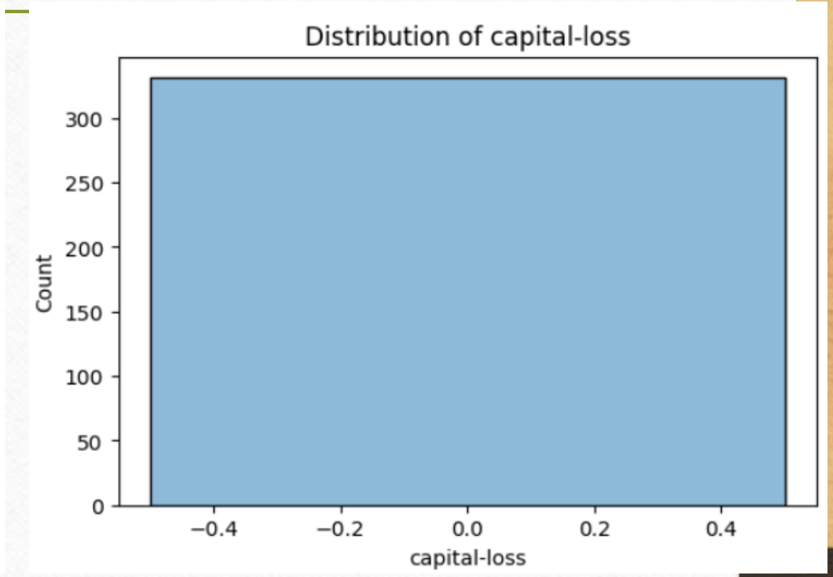
Distribution of policy_to_incident_days

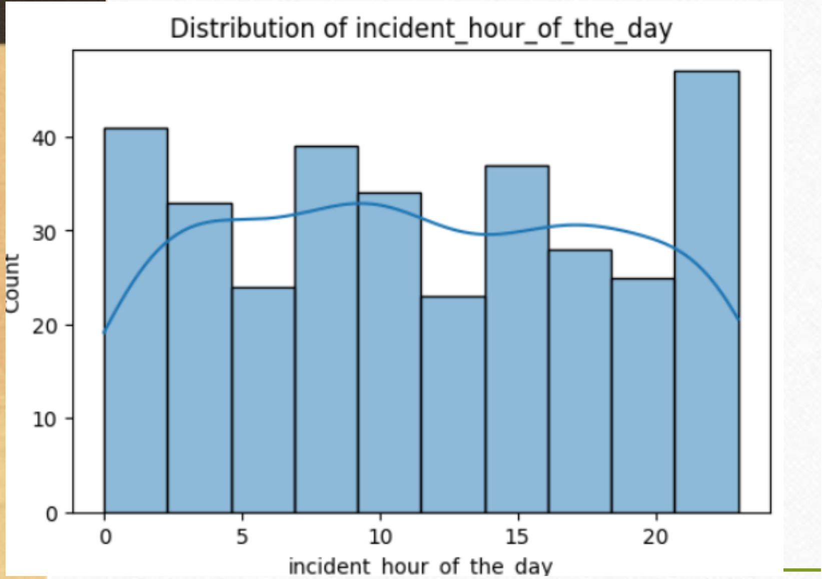

Distribution of policy_annual_premium
Count
0
10
20
30
40
50
400
600
800
1000
1200
1400
1600
1800
2000
policy_annual_premium

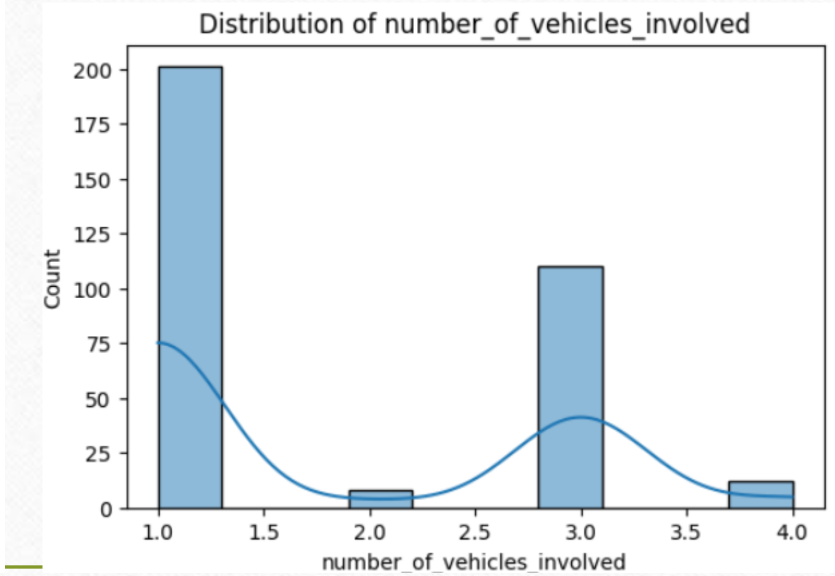

Distribution of umbrella_limit
Count
0
50
100
150
200
250
0
2
4
6
8
1e6
umbrella_limit

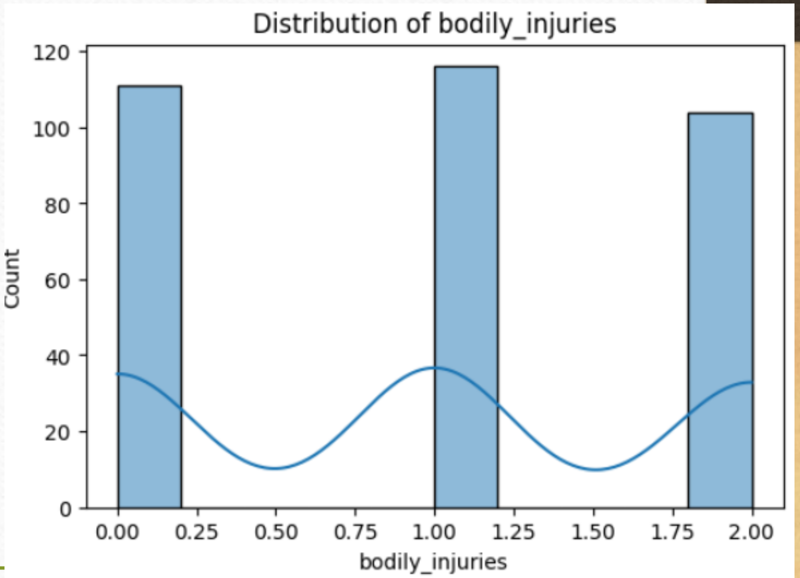

Distribution of insured_zip
Count
0
20
40
60
80
100
425000
450000
475000
500000
525000
550000
575000
600000
625000
insured_zip


Distribution of capital-gains
Count
0
25
50
75
100
125
150
175
0
20000
40000
60000
80000
100000
capital-gains

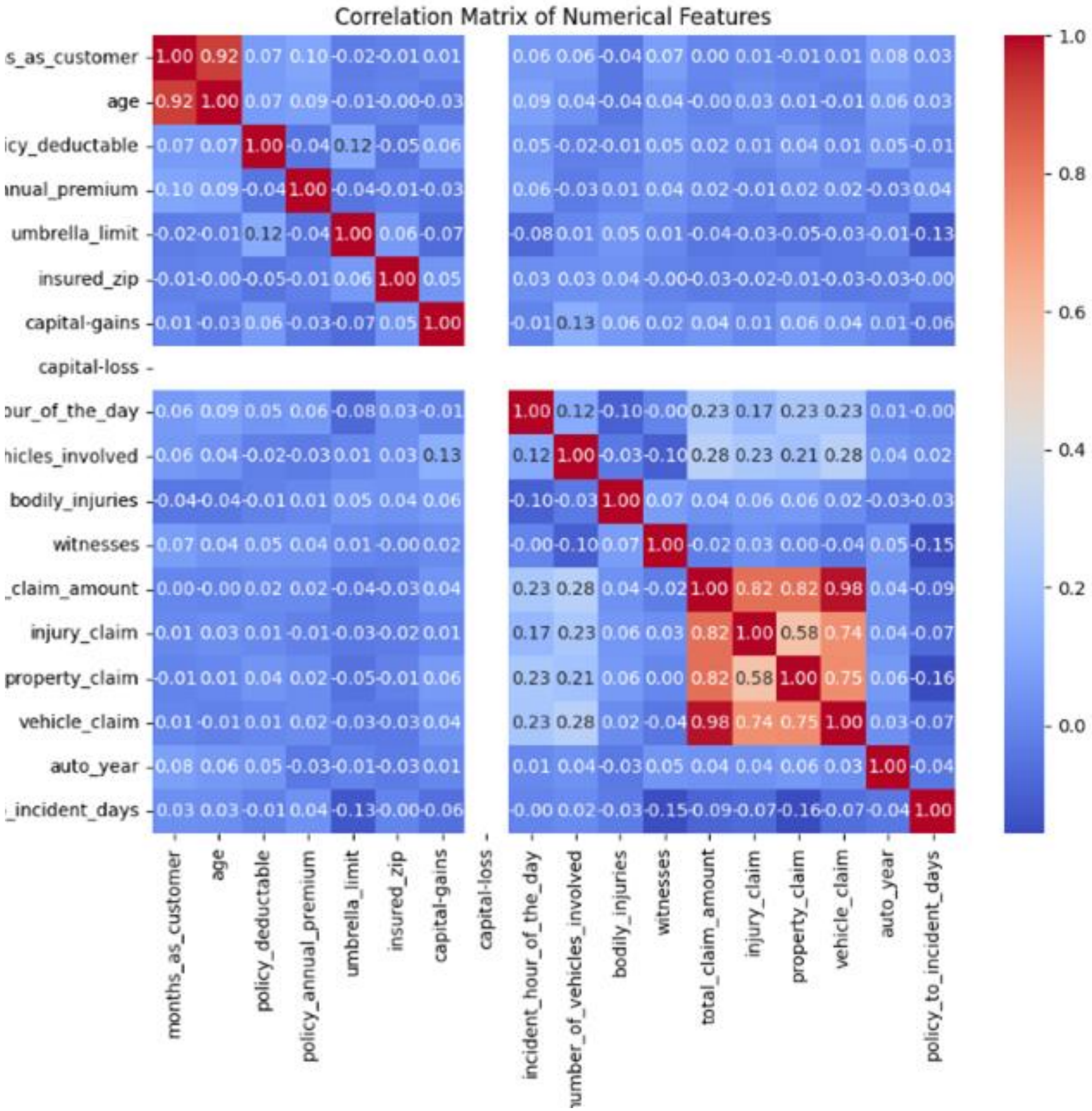

Distribution of capital-loss
Count
0
50
100
150
200
250
300
−0.4
−0.2
0.0
0.2
0.4
capital-loss


Distribution of incident_hour_of_the_day
Count
0
10
20
30
40
0
5
10
15
20
incident_hour_of_the_day


Distribution of number_of_vehicles_involved
Count
0
25
50
75
100
125
150
175
200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
number_of_vehicles_involved


Distribution of bodily_injuries
Count
0
20
40
60
80
100
120
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
bodily_injuries

Correlation Matrix of Numerical features - Heatmap



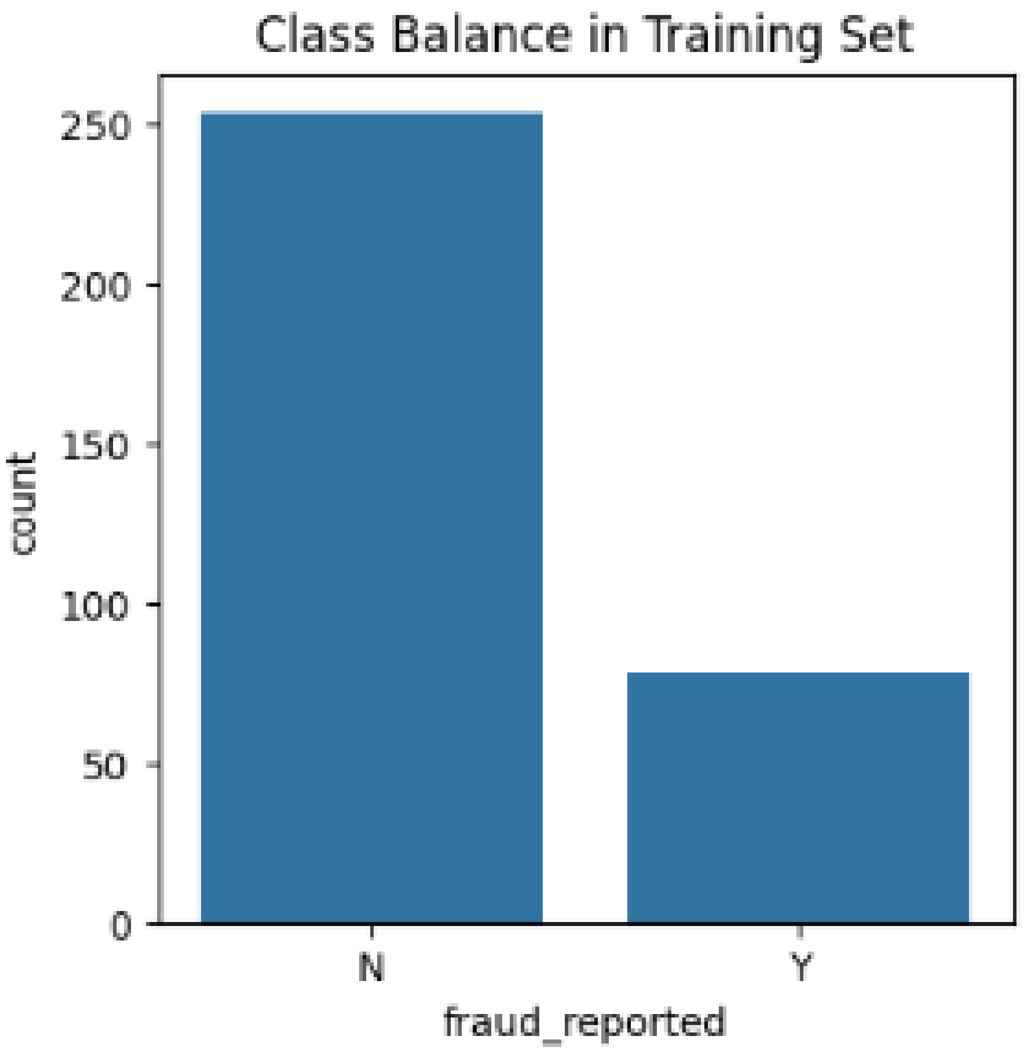
	months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	insured_zip	capital-gains	capital-loss	incident_hour_of_the_day	umber_of_vehicles_involved	bodily_injuries	witnesses	total_claim_amount	injury_claim	property_claim	vehicle_claim	auto_year	policy_to_incident_days
s_as_customer	1.00	0.92	0.07	0.10	-0.02	-0.01	0.01		0.06	0.06	-0.04	0.07	0.00	0.01	-0.01	0.01	0.08	0.03
age	0.92	1.00	0.07	0.09	-0.01	-0.00	-0.03		0.09	0.04	-0.04	0.04	-0.00	0.03	0.01	-0.01	0.06	0.03
icy_deductable	0.07	0.07	1.00	-0.04	0.12	-0.05	0.06		0.05	-0.02	-0.01	0.05	0.02	0.01	0.04	0.01	0.05	-0.01
nual_premium	0.10	0.09	-0.04	1.00	-0.04	-0.01	-0.03		0.06	-0.03	0.01	0.04	0.02	-0.01	0.02	0.02	-0.03	0.04
umbrella_limit	-0.02	-0.01	0.12	-0.04	1.00	0.06	-0.07		-0.08	0.01	0.05	0.01	-0.04	-0.03	-0.05	-0.03	-0.01	-0.13
insured_zip	-0.01	-0.00	-0.05	-0.01	0.06	1.00	0.05		0.03	0.03	0.04	-0.00	-0.03	-0.02	-0.01	-0.03	-0.03	-0.00
capital-gains	0.01	-0.03	0.06	-0.03	-0.07	0.05	1.00		-0.01	0.13	0.06	0.02	0.04	0.01	0.06	0.04	0.01	-0.06
capital-loss																		
our_of_the_day	0.06	0.09	0.05	0.06	-0.08	0.03	-0.01		1.00	0.12	-0.10	-0.00	0.23	0.17	0.23	0.23	0.01	-0.00
hicles_involved	0.06	0.04	-0.02	-0.03	0.01	0.03	0.13		0.12	1.00	-0.03	-0.10	0.28	0.23	0.21	0.28	0.04	0.02
bodily_injuries	-0.04	-0.04	-0.01	0.01	0.05	0.04	0.06		-0.10	-0.03	1.00	0.07	0.04	0.06	0.06	0.02	-0.03	-0.03
witnesses	0.07	0.04	0.05	0.04	0.01	-0.00	0.02		-0.00	-0.10	0.07	1.00	-0.02	0.03	0.00	-0.04	0.05	-0.15
_claim_amount	0.00	-0.00	0.02	0.02	-0.04	-0.03	0.04		0.23	0.28	0.04	-0.02	1.00	0.82	0.82	0.98	0.04	-0.09
injury_claim	0.01	0.03	0.01	-0.01	-0.03	-0.02	0.01		0.17	0.23	0.06	0.03	0.82	1.00	0.58	0.74	0.04	-0.07
property_claim	-0.01	0.01	0.04	0.02	-0.05	-0.01	0.06		0.23	0.21	0.06	0.00	0.82	0.58	1.00	0.75	0.06	-0.16
vehicle_claim	0.01	-0.01	0.01	0.02	-0.03	-0.03	0.04		0.23	0.28	0.02	-0.04	0.98	0.74	0.75	1.00	0.03	-0.07
auto_year	0.08	0.06	0.05	-0.03	-0.01	-0.03	0.01		0.01	0.04	-0.03	0.05	0.04	0.04	0.06	0.03	1.00	-0.04
_incident_days	0.03	0.03	-0.01	0.04	-0.13	-0.00	-0.06		-0.00	0.02	-0.03	-0.15	-0.09	-0.07	-0.16	-0.07	-0.04	1.00

Months as customer and age have a very **high correlation** (0.92) — meaning older customers tend to be with the company longer.

Total claim amount is strongly correlated with injury claim, property claim, and vehicle claim.

Number of vehicles involved and bodily injuries show some **positive correlation** with claims — which could be useful for fraud detection.

No extreme multicollinearity is spotted apart from the claim amount features, which is good for model building.

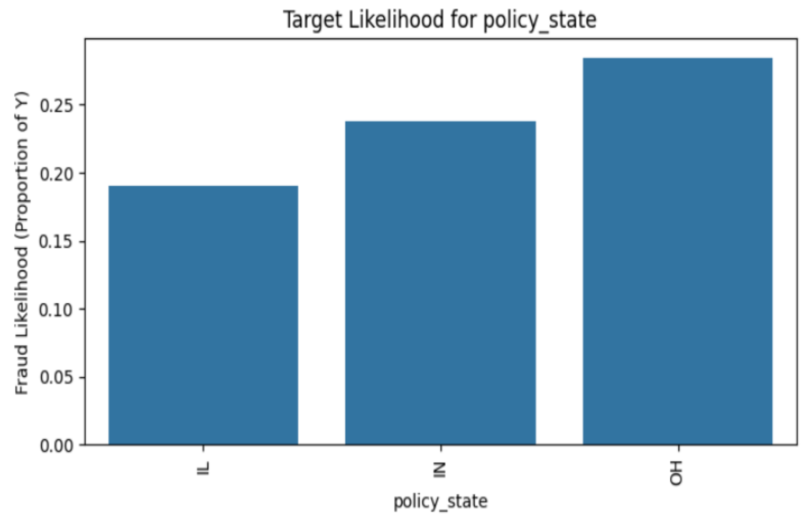


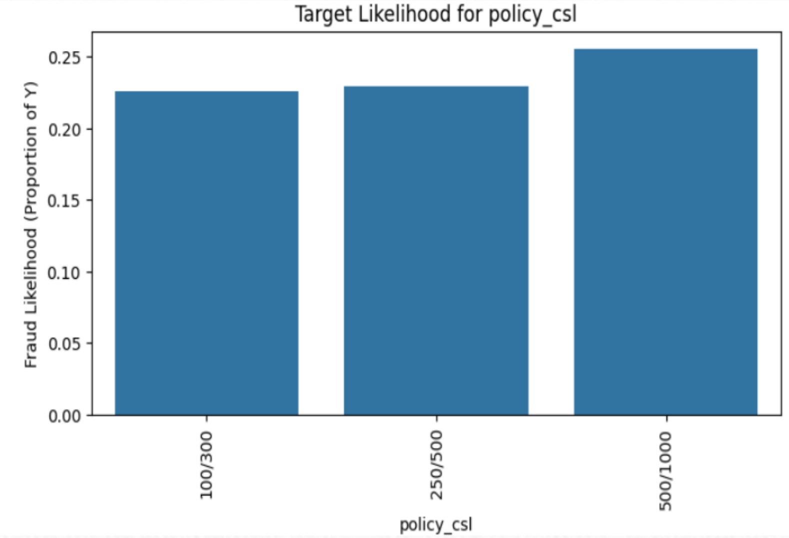
A bar chart to check class balance

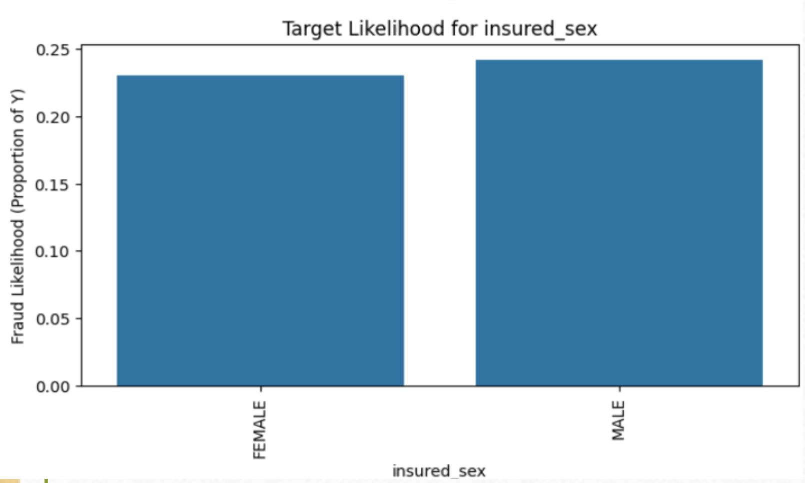
The dataset is imbalanced — there are a lot more **non-fraudulent claims (N)** than fraudulent ones (Y).

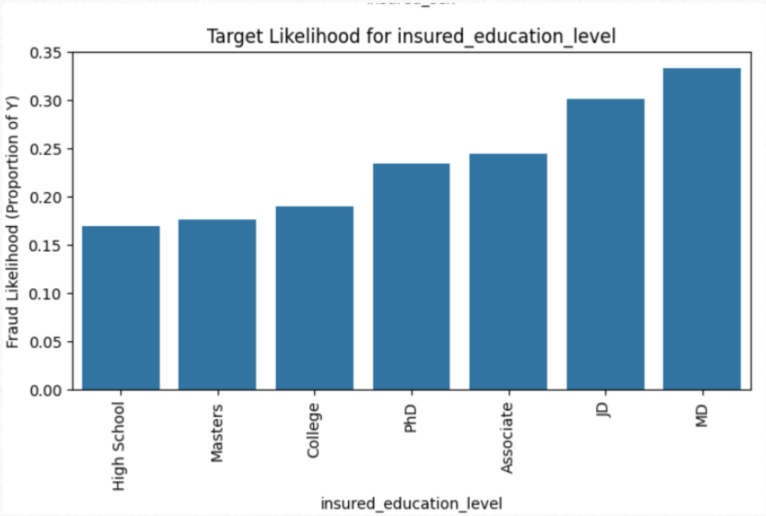
Specifically, fraudulent claims are much fewer, which can make the model biased towards predicting "No Fraud" more often.

This imbalance needs to be handled using **class weight adjustment** to improve fraud detection performance.









Bivariate Analysis

Target likelihood of males is higher than females.

Target likelihood of education level of JD and MD is higher