Table Of Content

Problem Statement

Problem:

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 minimise financial losses and optimise the overall claims handling.

Problem Statement

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.

Overall Approach

- Learn about the business domain of the enterprise through the terms in the Data Dictionary.
- Based on the acquired knowledge, predict the factors that may influence the fraudulent claim rate.
- Use EDA and visualization to understand and confirm our first thoughts
- Build models from simple to complex and evaluate their effectiveness.
- Choose the most suitable solution for our business problem.

Data Analysis

Data Preparation and Cleaning:

- Understanding and Cleaning:

- **Drop null column(s):** *_c39*
- **Delete rows containing null values** (in *‘authorities_contacted’* columns)
- **Replace** ‘?’ **with** ‘UNKNOWN’
- **Remove columns where a large proportion of the values are unique**: *'policy_number', 'insured_zip', 'incident_location’*
- **Change data type** (datetime): ‘*policy_bind_date’, ‘incident_date’*
- **Drop rows where features have illogical or invalid values (column: ‘***umbrella_limit’***)**

Data Analysis

Data Preparation and Cleaning:

- Understanding and Cleaning:

- The initial data set (1000, 40) was reduced to (908, 36)
- Key features for deeper analysis have been classified as below:

Numerical Features

'months_as_customer', 'age', 'policy_deductable', 'policy_annual_premium', 'umbrella_limit', 'capital-gains', 'capital-loss', 'incident_hour_of_the_day', 'number_of_vehicles_involved', 'bodily_injuries', 'witnesses', 'total_claim_amount', 'injury_claim', 'property_claim', 'vehicle_claim'

Categorical Features

'policy_state', 'policy_csl', 'insured_sex', 'insured_education_level', 'insured_occupation', 'insured_hobbies', 'insured_relationship', 'incident_type', 'collision_type', 'incident_severity', 'authorities_contacted', 'incident_state', 'incident_city', 'property_damage', 'police_report_available', 'auto_make', 'auto_model', 'auto_year'

Date Features

'policy_bind_date', 'incident_date'

Target Feature

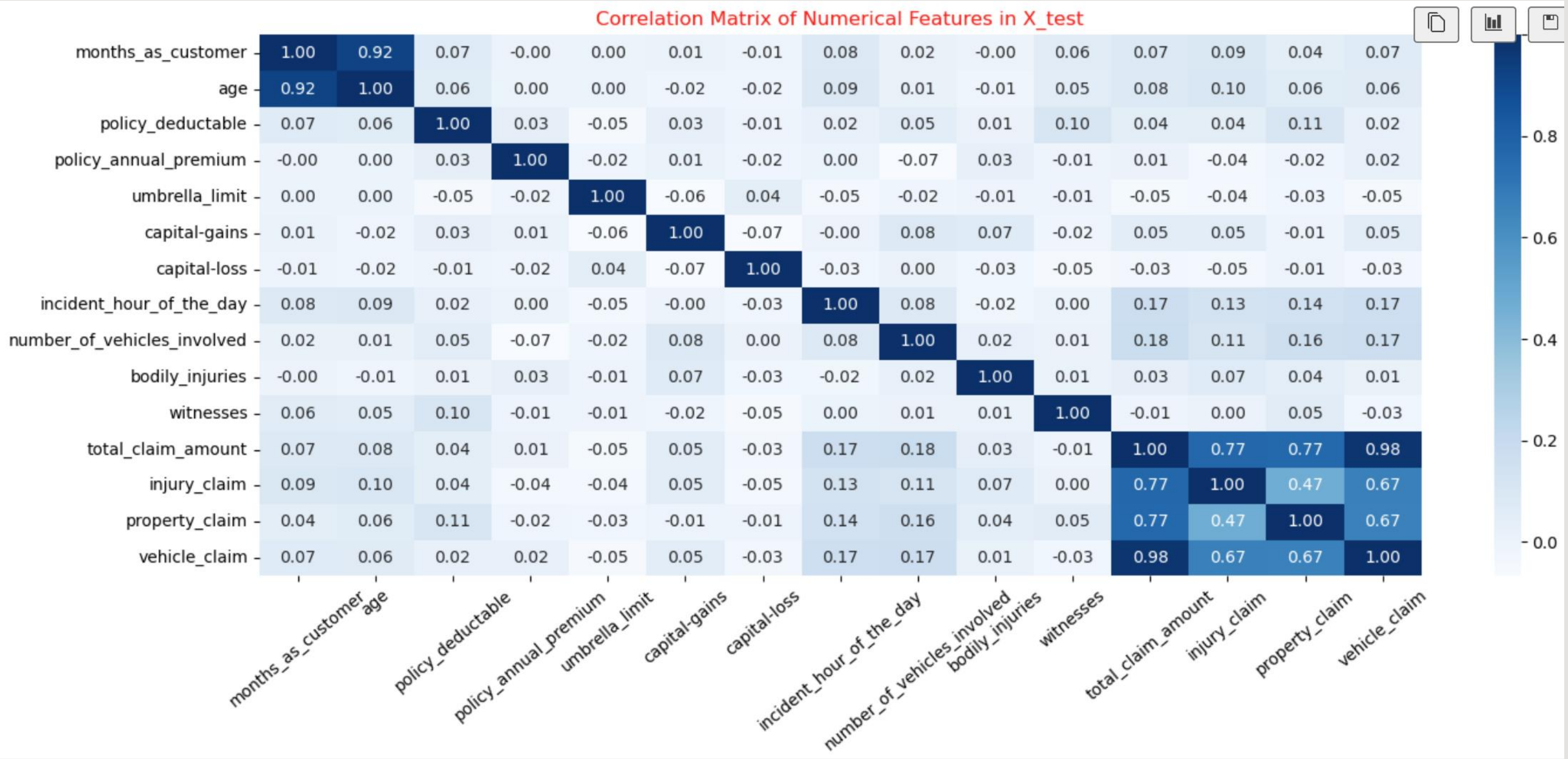
'fraud_reported'

Data Analysis

EDA: Numerical Feature Correlation

Feature pairs with high correlation:

- ‘injury_claim’, ‘property_claim’, ‘vehicle_claim’ **vs.** ‘ total_claim_amount’
- ‘months_as_customer’ **vs**. ‘age’

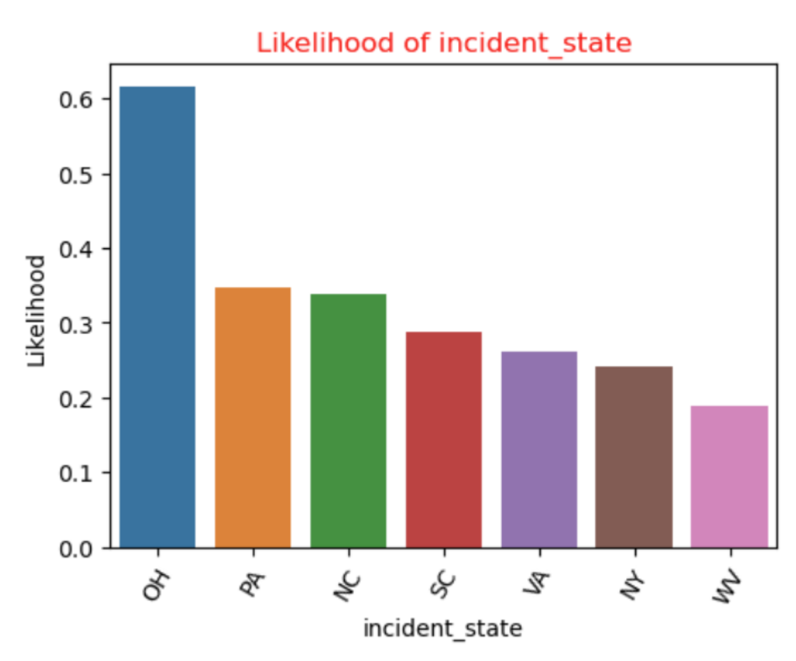


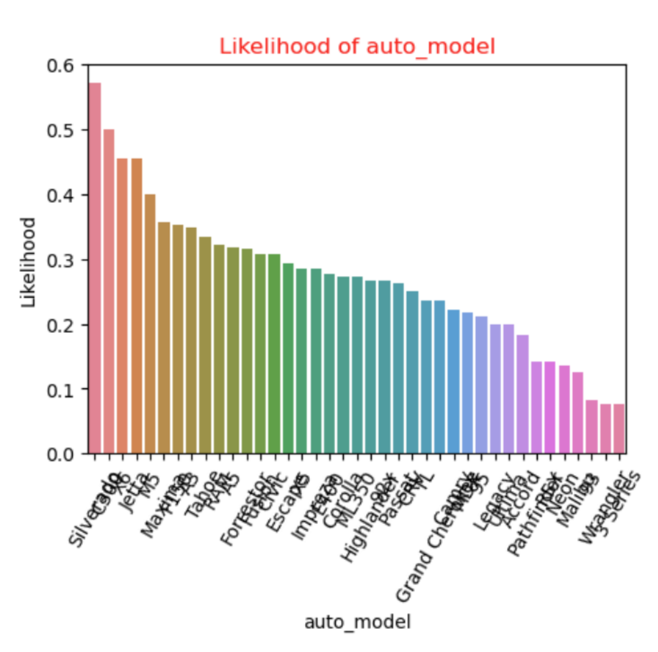
Data Analysis

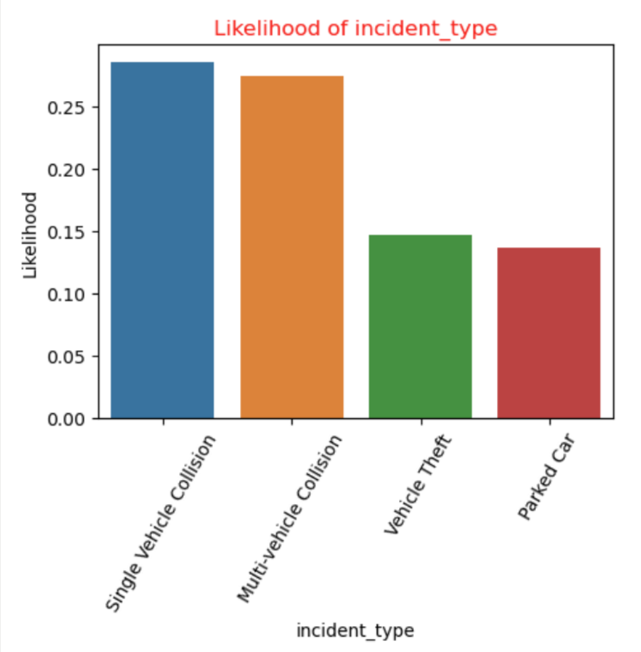
Target likelihood analysis for categorical variables

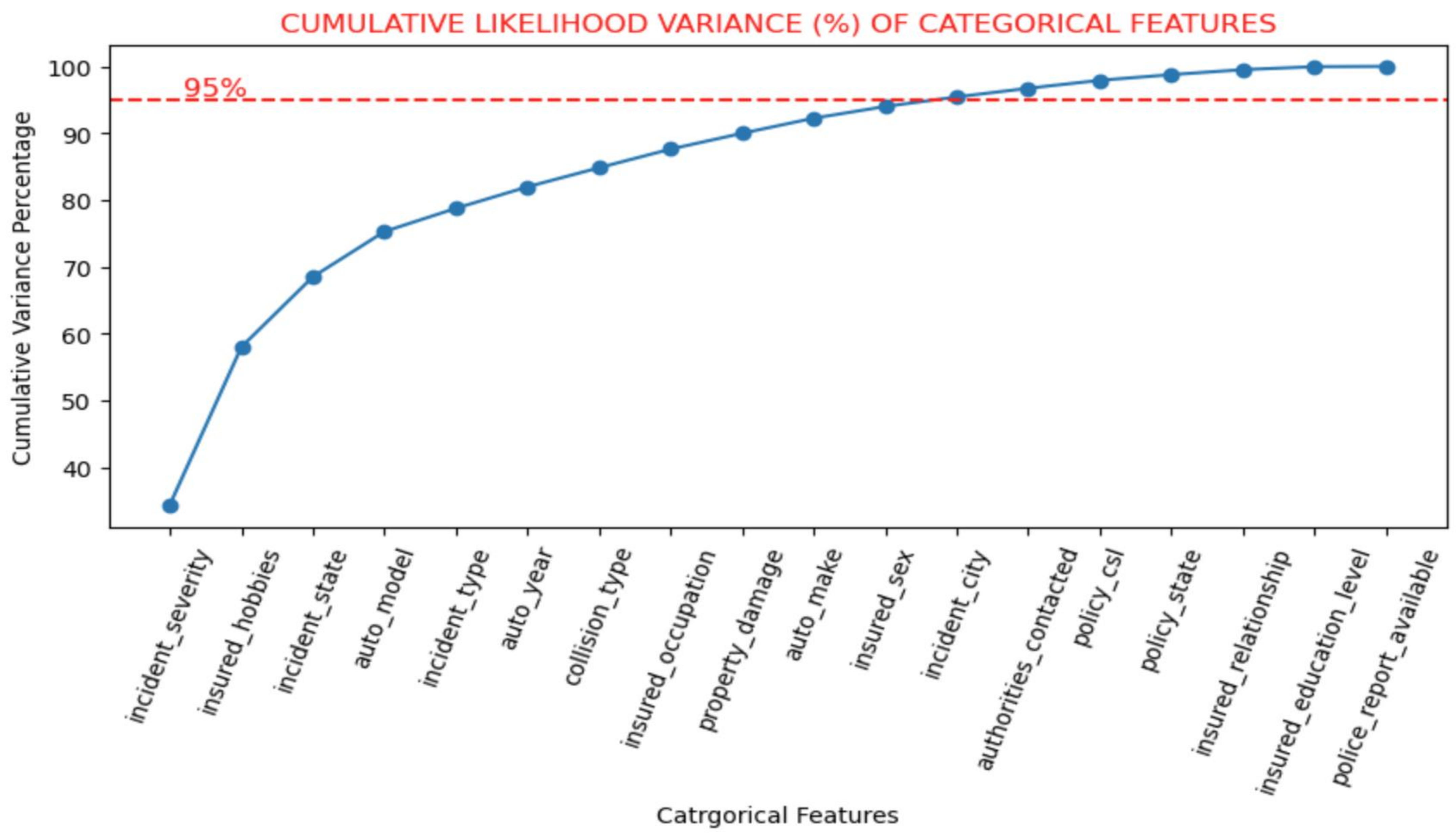
5 features with highest likelihood variances:

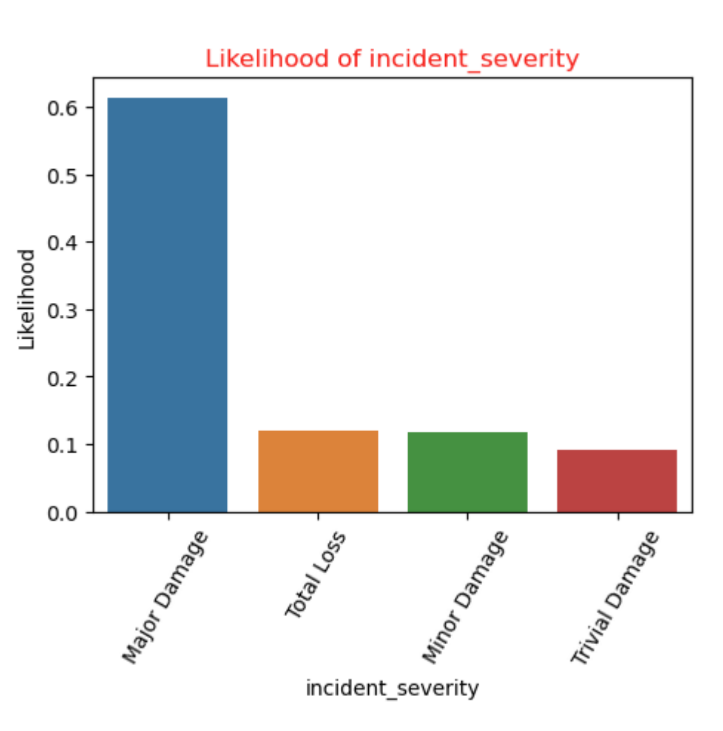
- **incident_severity**: 0.0634
- **insured_hobbies**: 0.0435
- **incident_state**: 0.0194
- **auto_model**: 0.0125
- **incident_type**: 0.0064

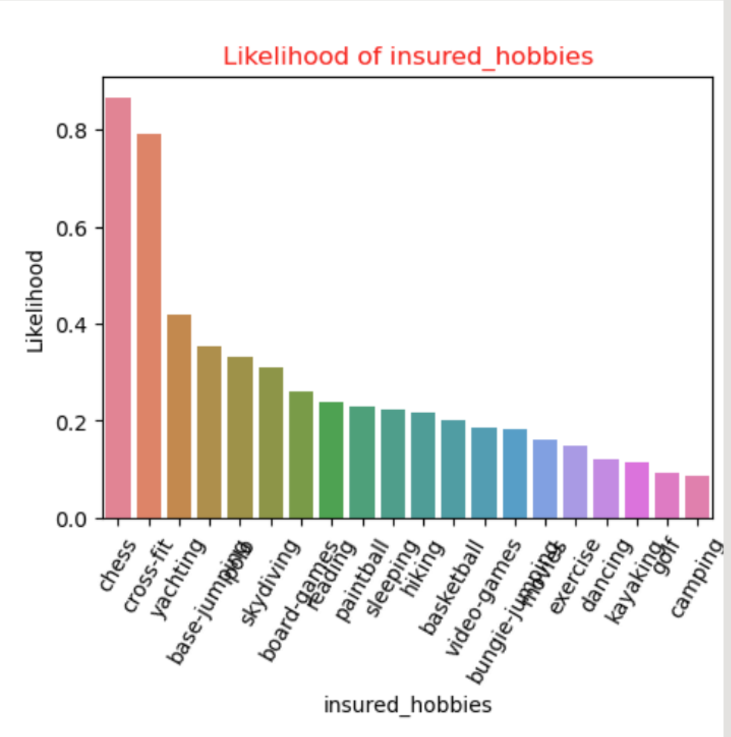










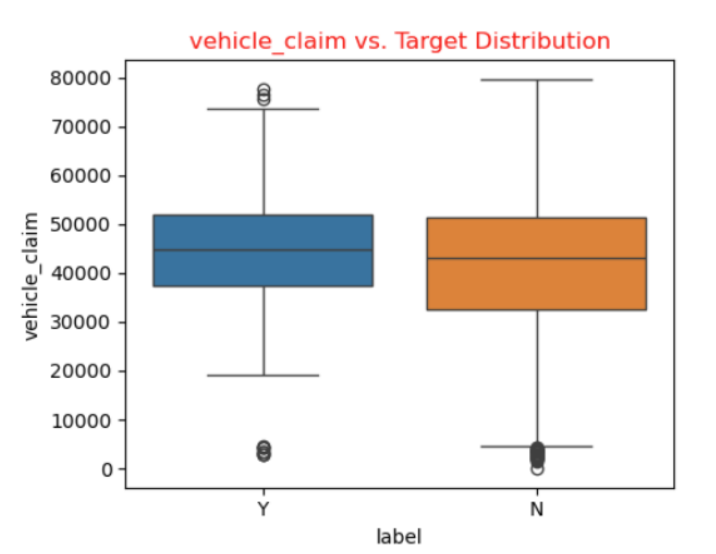


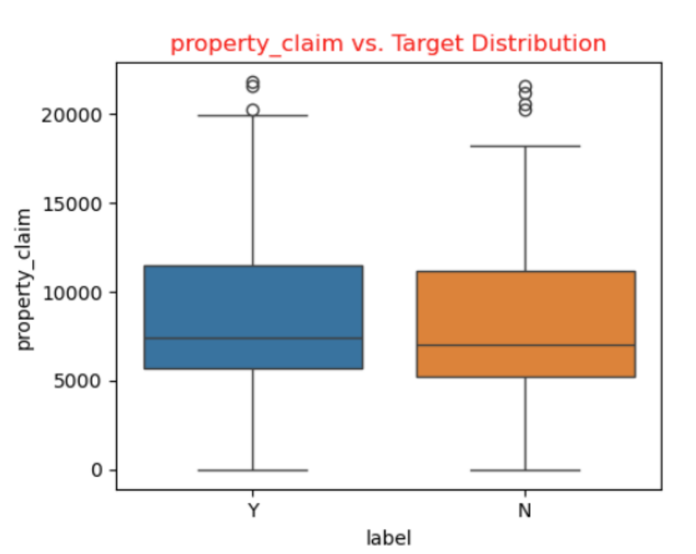
Data Analysis

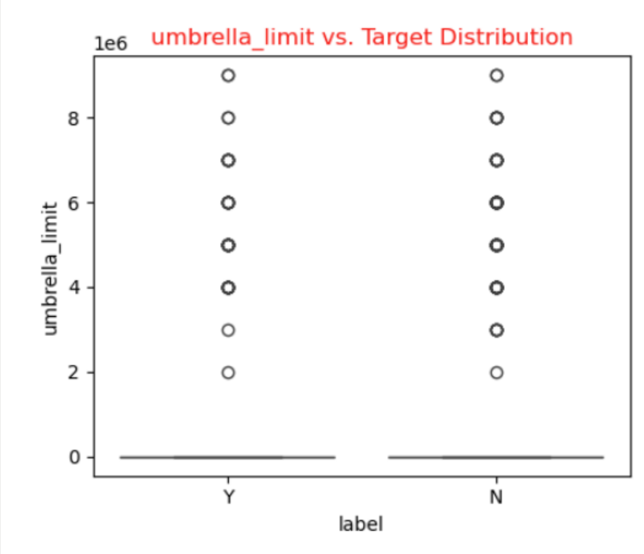
Target likelihood analysis for numerical variables

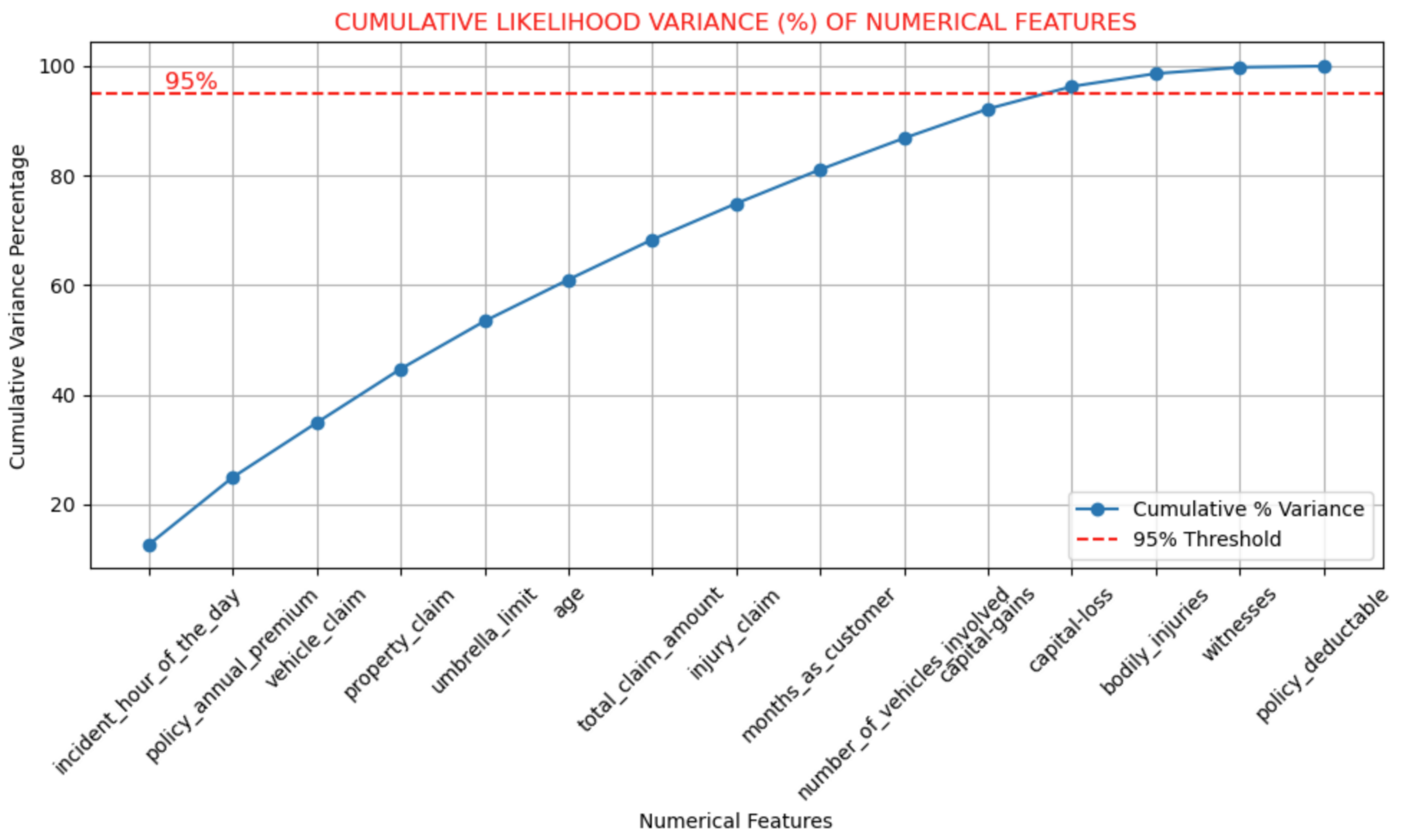
5 features with highest likelihood variances:

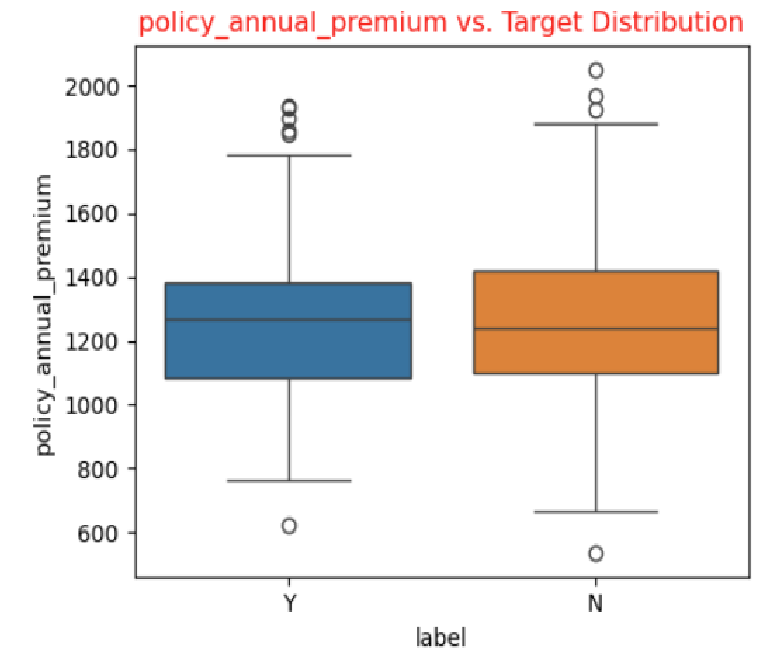
- **incident_hour_of_the_day:** 0.0113
- **policy_annual_premium:** 0.0109
- **vehicle_claim:** 0.0089
- **property_claim:** 0.0087
- **umbrella_limit:** 0.0078







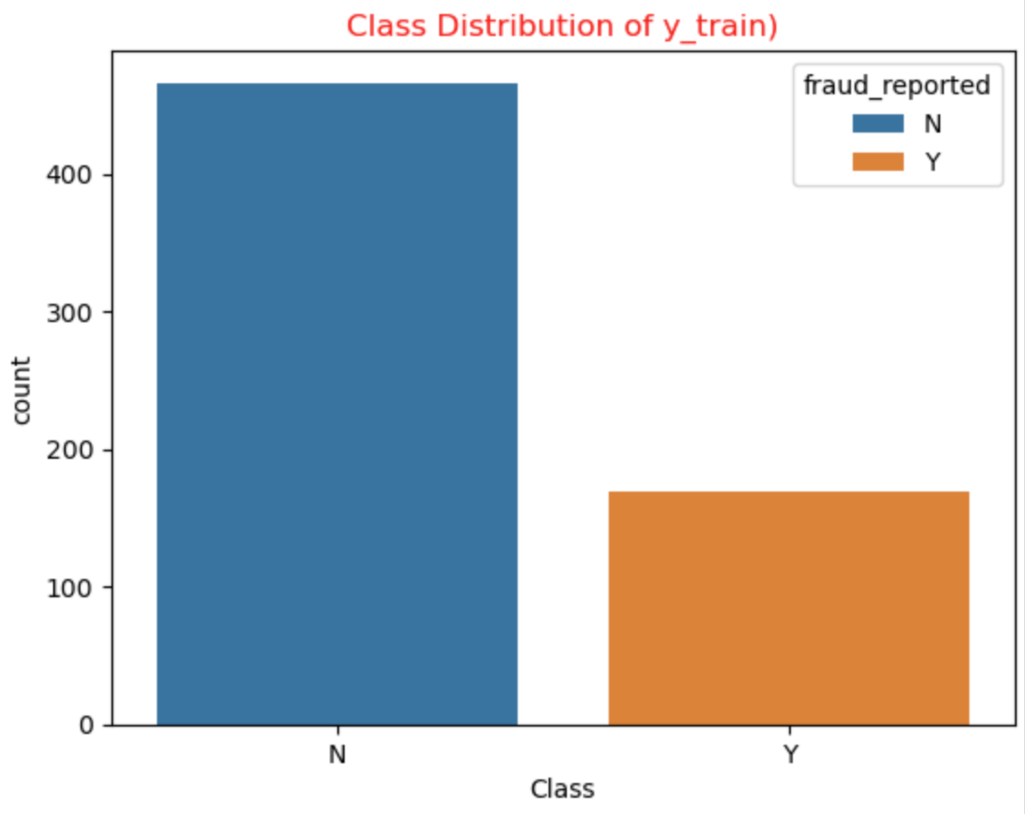


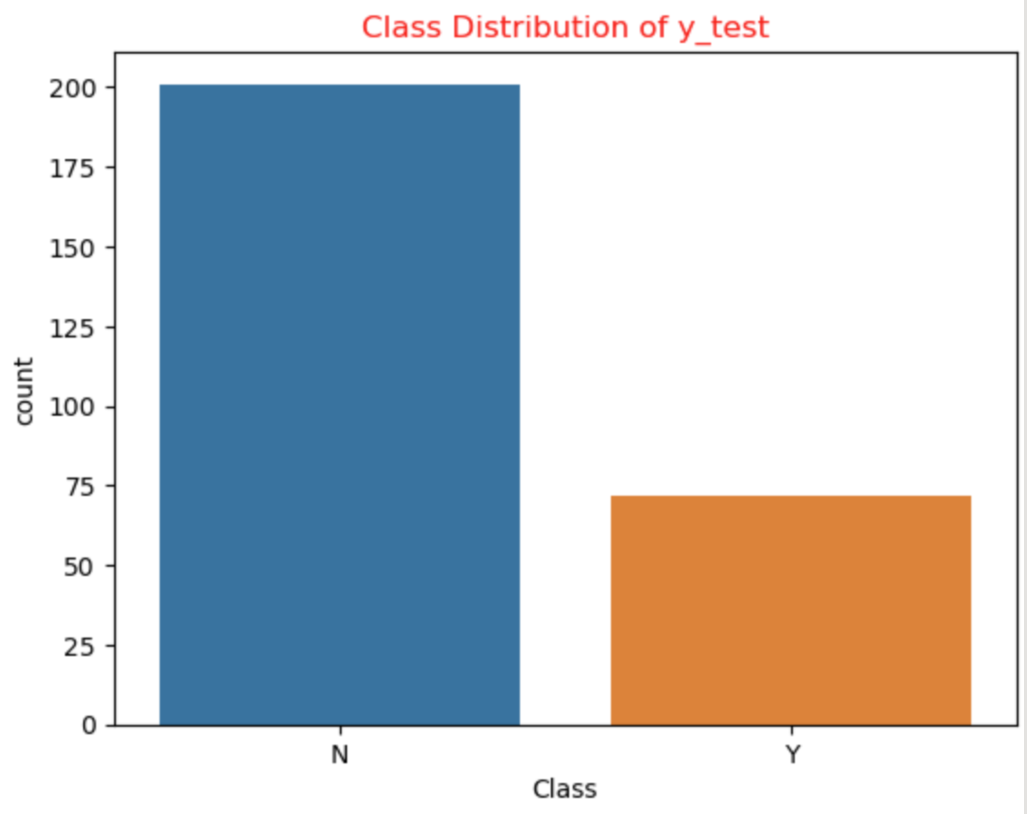


Data Preparation

Feature Engineering:

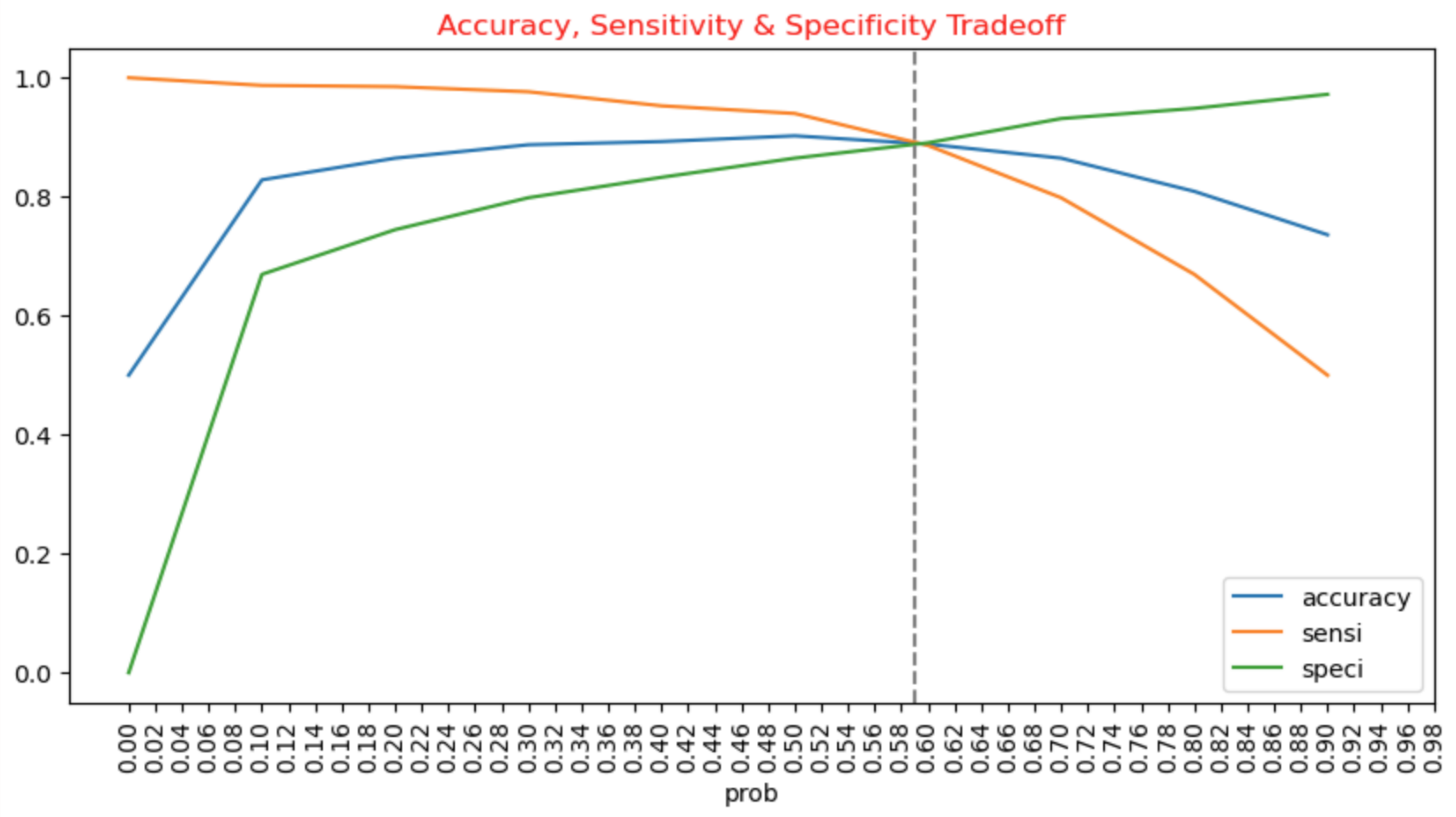
- Handle **class imbalance** by resampling, using **RandomOverSampler()**
- Derive the features ‘**policy_bind_month**’, ‘**is_weekend_incident**’, ‘**incident_month**’ from ‘**policy_bind_date**’ and ‘**incident_date**’
- Derive the features ‘**policy_csl_max’, ‘capital_gain_net’**, ‘**high_premium**’, ‘**high_claim’, ‘claim_to_premium_ratio’, ‘claim_to_deductible_ratio’, ‘vehicle_claim_ratio’, ‘injury_claim_ratio’, ‘property_claim_ratio’** and **‘age_group’, ‘auto_type’**
- Handle low-requency values (covert them to ‘other’)
- Create dummy features using **get_dummies()**
- Feature Scaling using **StandardScaler()**

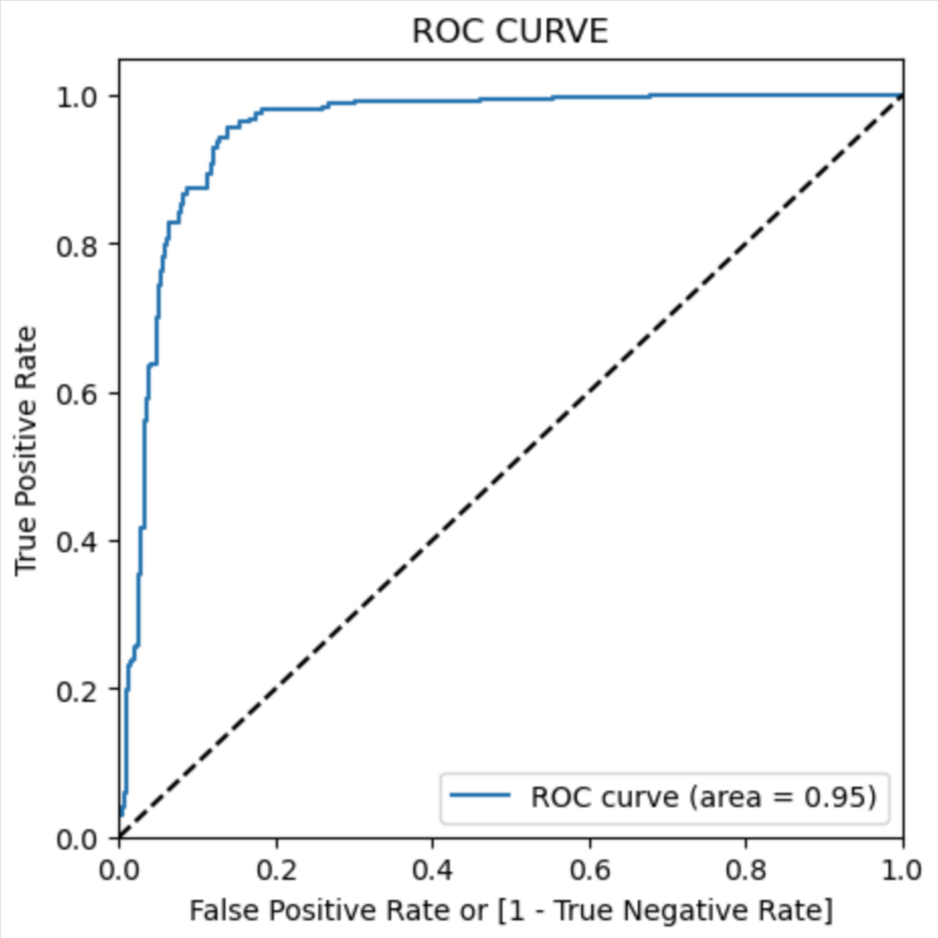


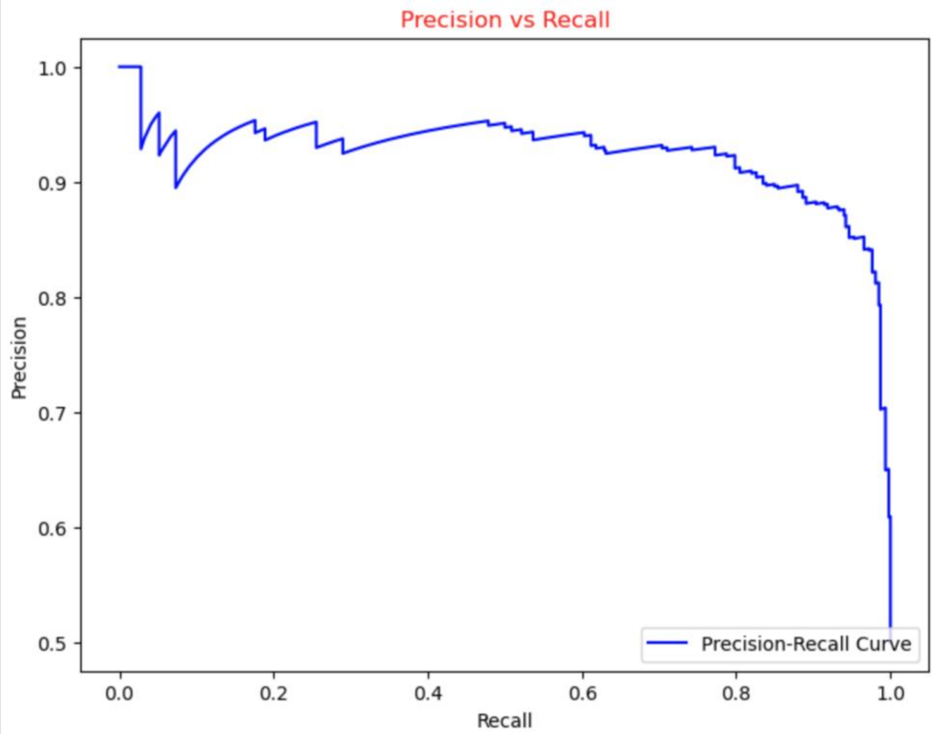


Model Building: Logistic Regression

Optimal Cuttoff	0.59
Accuracy	0.89
Sensitivity (Recall)	0.89
Specificity	0.89
Precision	0.89
F1 score:	0.91

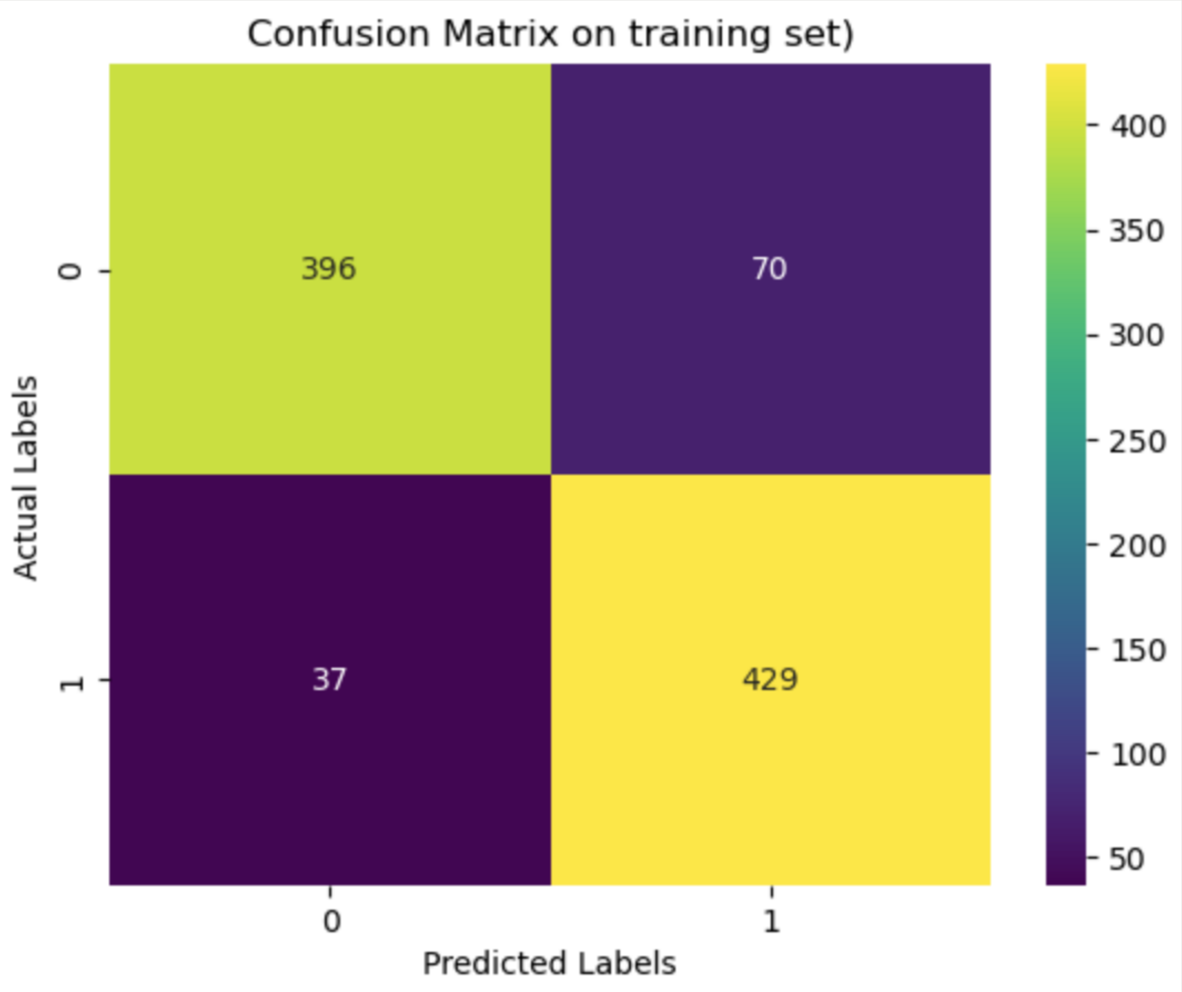


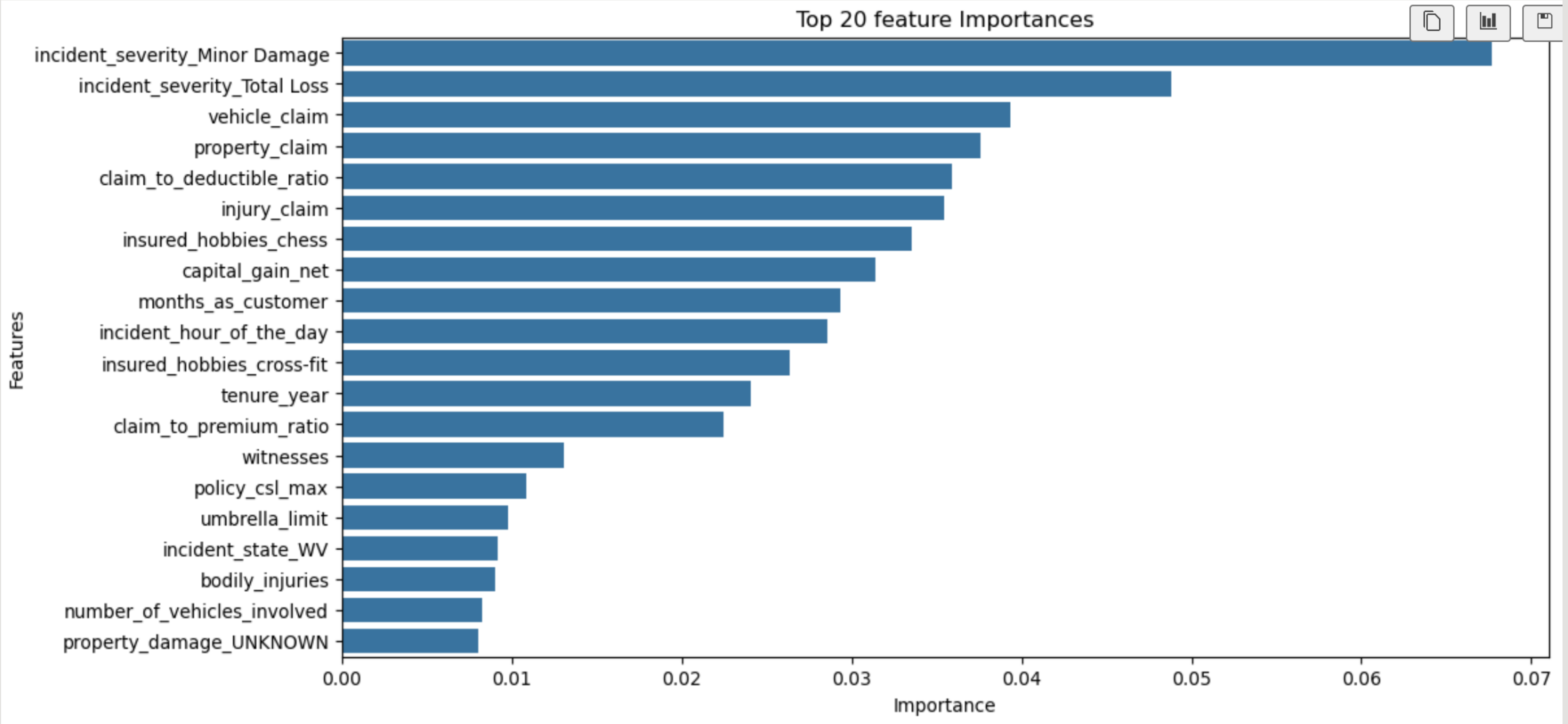




Model Building: Random Forest

Best estimator	'max_depth': 12, 'max_features': 8, 'min_samples_leaf': 10, 'min_samples_split': 10, 'n_estimators': 20
Accuracy	0.89
Sensitivity (Recall)	0.92
Specificity	0.85
Precision	0.86
F1 score:	0.91





Model Prediction & Evalualtion

Model	Best parameters	Training Set	Test Set
Logistic Regression	**Optimal Cuttoff:** 0.59	• Accuracy score: 0.89 • Sensitivity (Recall): 0.89 • Specificity: 0.89 • Precision 0.89 • F1 score: 0.91	• Accuracy score: 0.77 • Sensitivity (Recall): 0.61 • Specificity: 0.83 • Precision: 0.56 • F1 score: 0.58
Random forest	**Best estimator:** {'max_depth': 12, 'max_features': 8, 'min_samples_leaf': 10, 'min_samples_split': 10, 'n_estimators’: 20}	• Accuracy score: 0.89 • Sensitivity (Recall): 0.92 • Specificity: 0.85 • Precision: 0.86 • F1 score: 0.91	• Accuracy score: 0.89 • Sensitivity (Recall): 0.92 • Specificity: 0.85 • Precision: 0.86 • F1 score: 0.58

Summary

Data-driven analysis of past claims revealed patterns of fraudulent behavior.

Logistic Regression and **Random Forest** models predicted fraud probability, with **Random Forest showing slightly better performance**.

Feature importance analysis of categorical and numerical features identified key fraud indicators.

High-variance features like **incident_severity, insured_hobbies**, **policy_annual_premium, and policy_annual_premium** components were strong fraud predictors.

Low-impact features (e.g., **insured_relationship**, **policy_state**, …) had minimal predictive power and could be deprioritized for better model efficiency