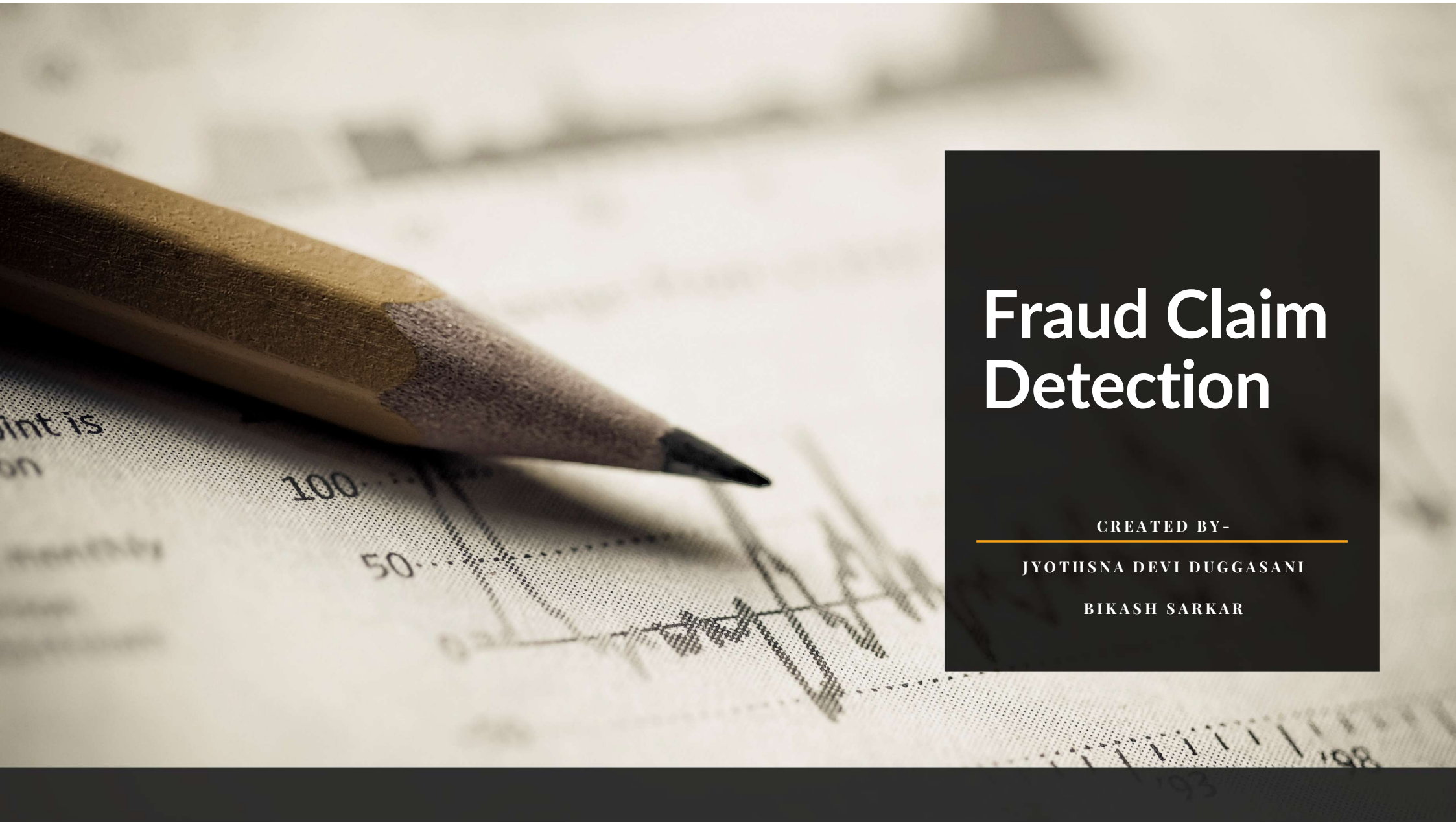


Fraud Claim Detection

CREATED BY-

JYOTHSNA DEVI DUGGASANI

BIKASH SARKAR

Table of 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 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 process.

Python Code

GITHUB LINK-
HTTPS://GITHUB.COM/BIKASH3/INSURANCECLAIMFRAUDDETECTION

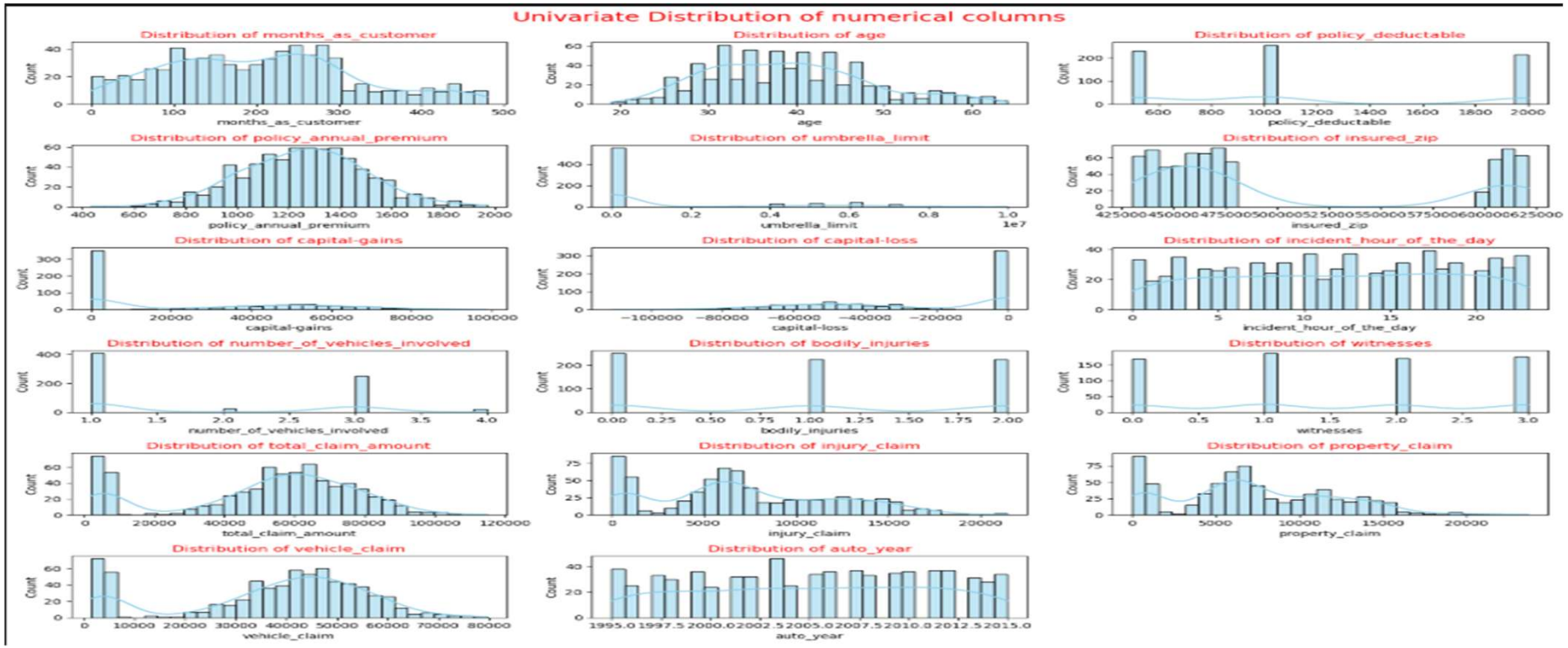
Project Summary

This project aims to detect fraudulent insurance claims using machine learning models, helping insurers reduce financial losses and improve operational efficiency. The process began with thorough data preprocessing and feature engineering to clean the dataset, handle missing values, and create meaningful features that capture fraud patterns. Two models—Logistic Regression and Random Forest—were developed to classify claims as either fraudulent or legitimate. To enhance model performance, hyperparameter tuning and threshold optimization were performed using cross-validation and grid search techniques. The final models were evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to ensure robustness and reliability. This end-to-end approach demonstrates how machine learning can be effectively leveraged for fraud detection in the insurance sector.

Assignment Tasks

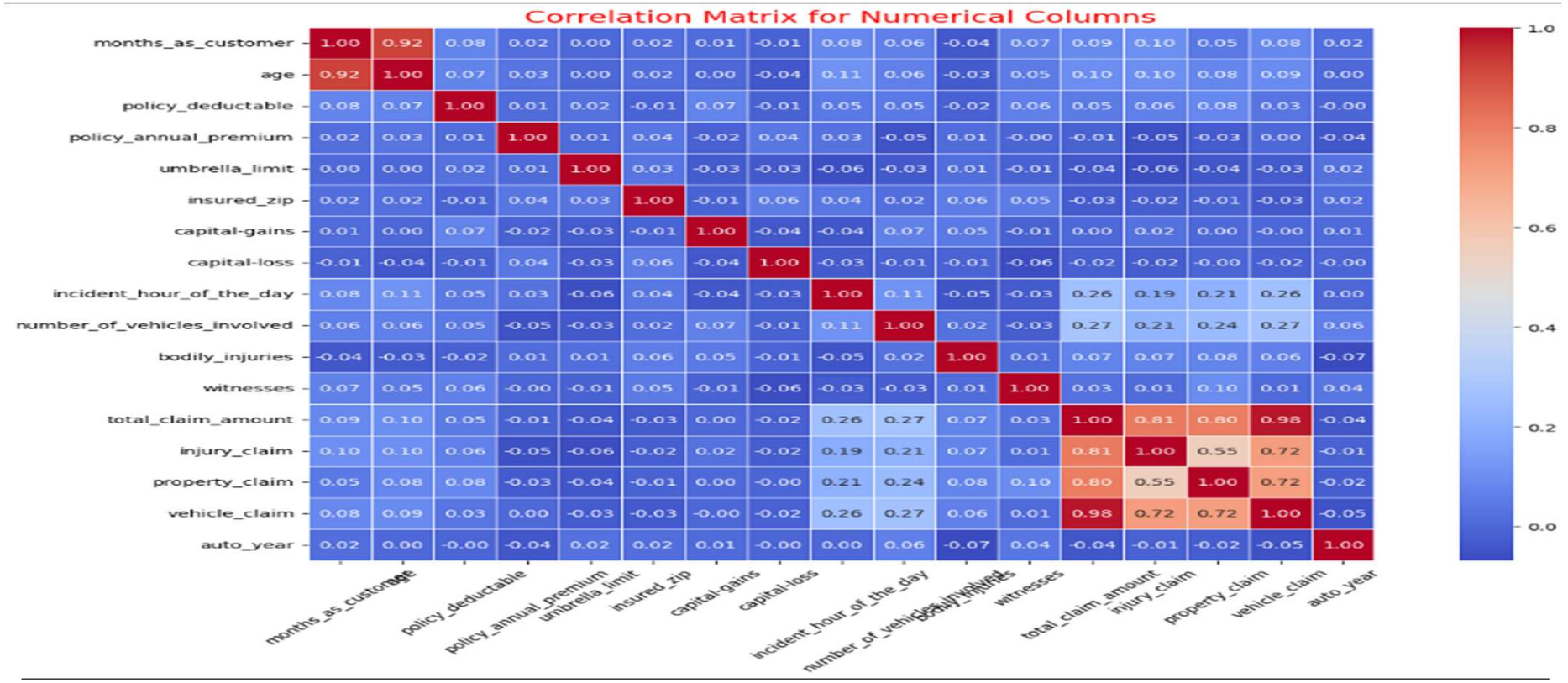
- **Data Preparation:**
  - Initial data loading and understanding of structure
- **Data Cleaning:**
  - Handling missing values and correcting inconsistencies
- **Train-Validation Split (70-30):**
  - Dividing the dataset for training and evaluation
- **Exploratory Data Analysis (EDA):**
  - Analyzing patterns, distributions, and outliers in the training data
  - EDA on validation data for additional insights
- **Feature Engineering**
  - Creating and transforming features to improve model performance
- **Model Building**
  - Training machine learning models to detect fraudulent claims
- **Prediction & Model Evaluation**
  - Generating predictions on the validation set
  - Evaluating model performance using key metrics

Analysis

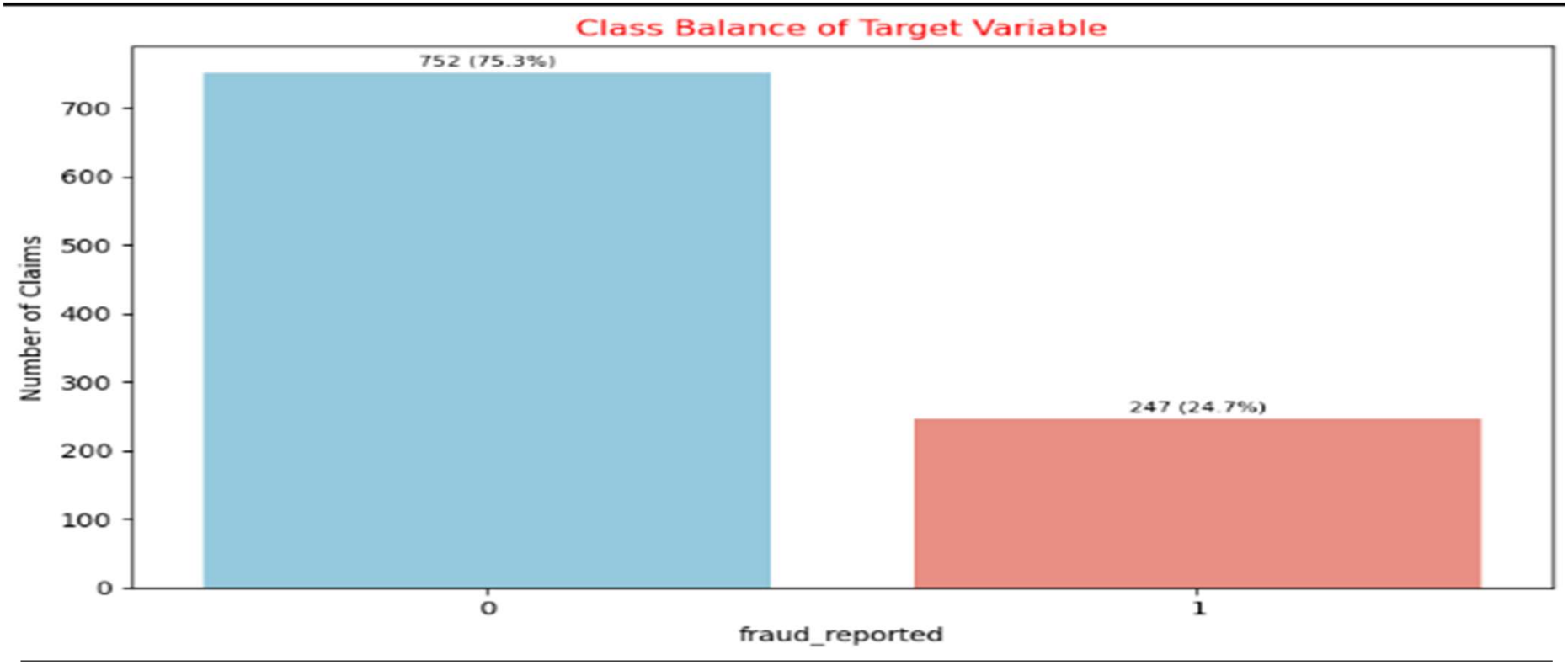


UNIVARIATE ANALYSIS

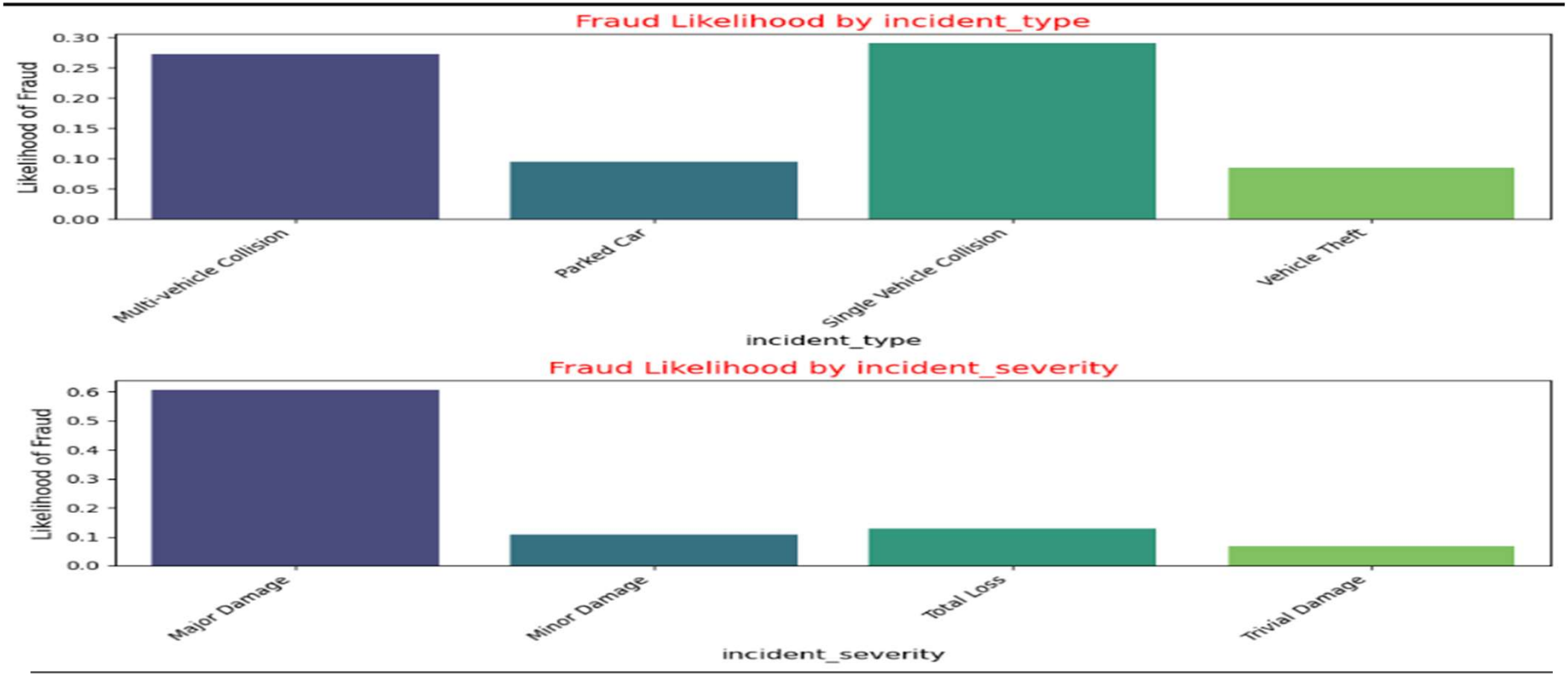
DISTRIBUTION OF NUMERICAL FEATURES



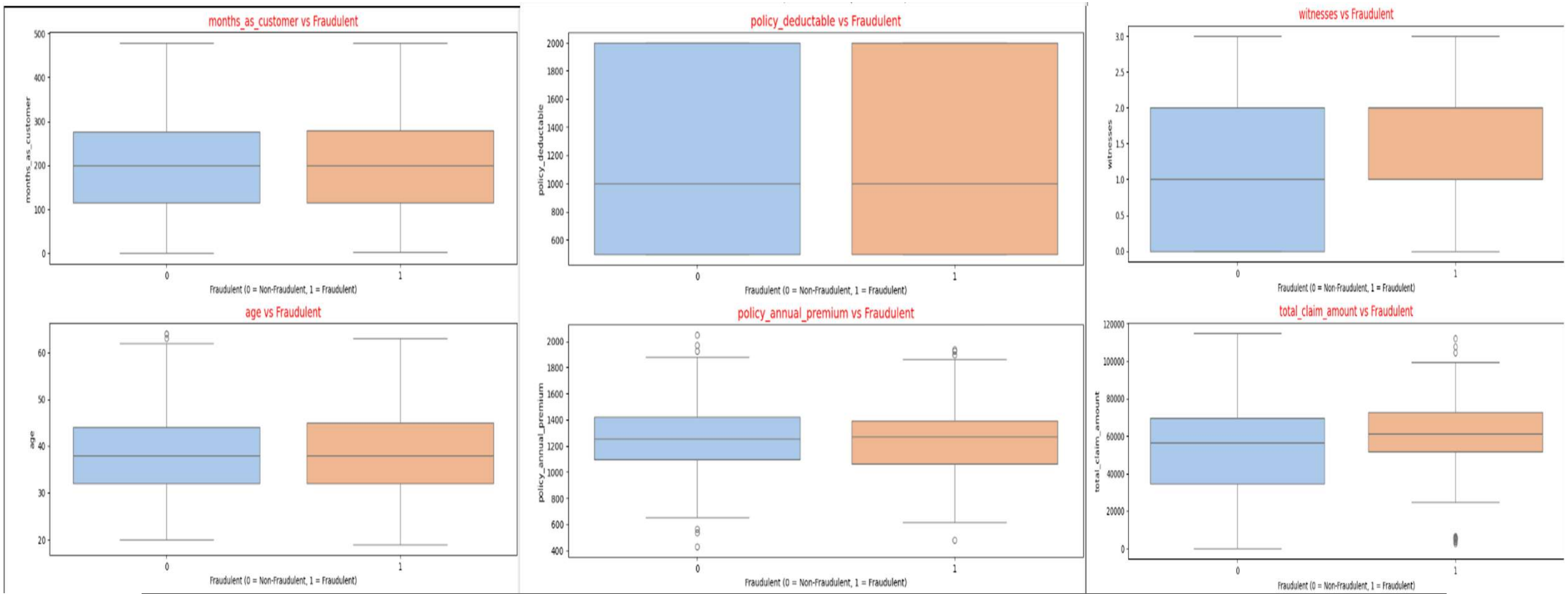
CORRELATION ANALYSIS



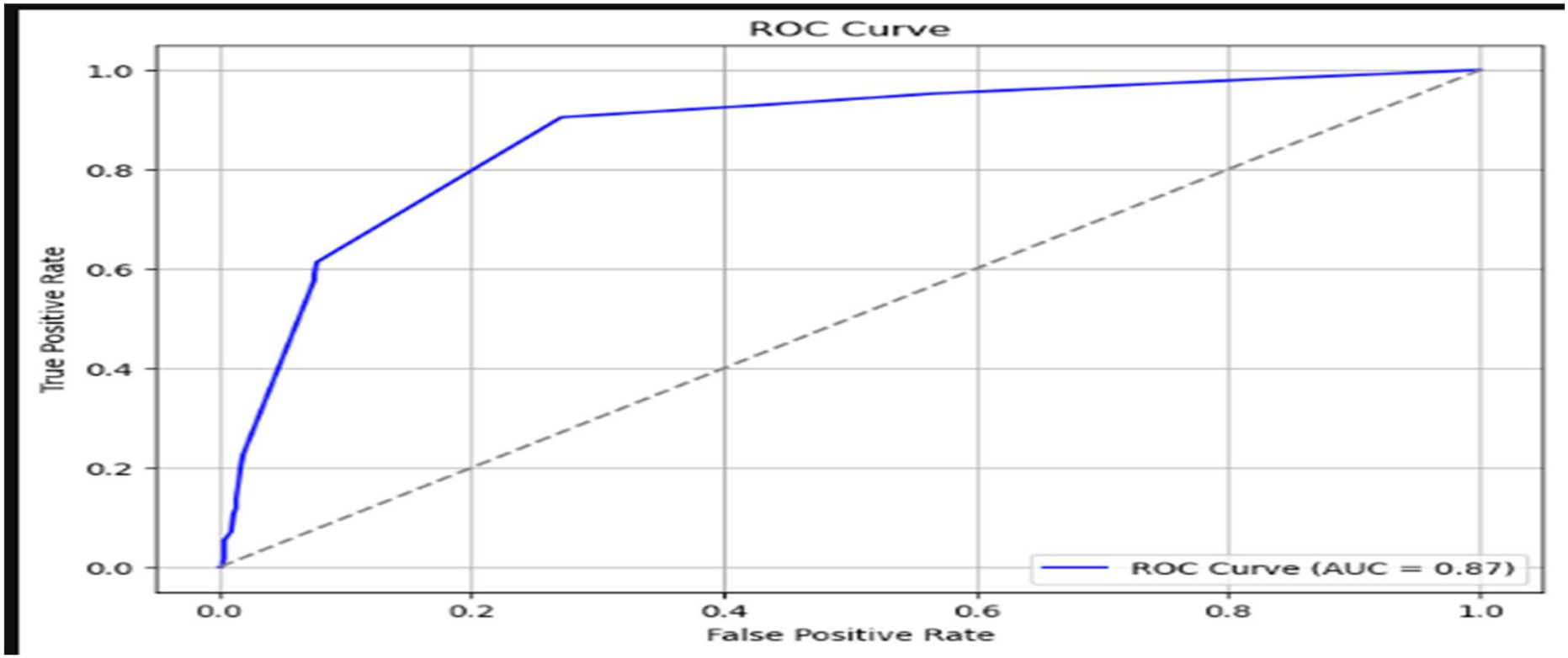
CLASS BALANCE OF TARGET VARIABLE



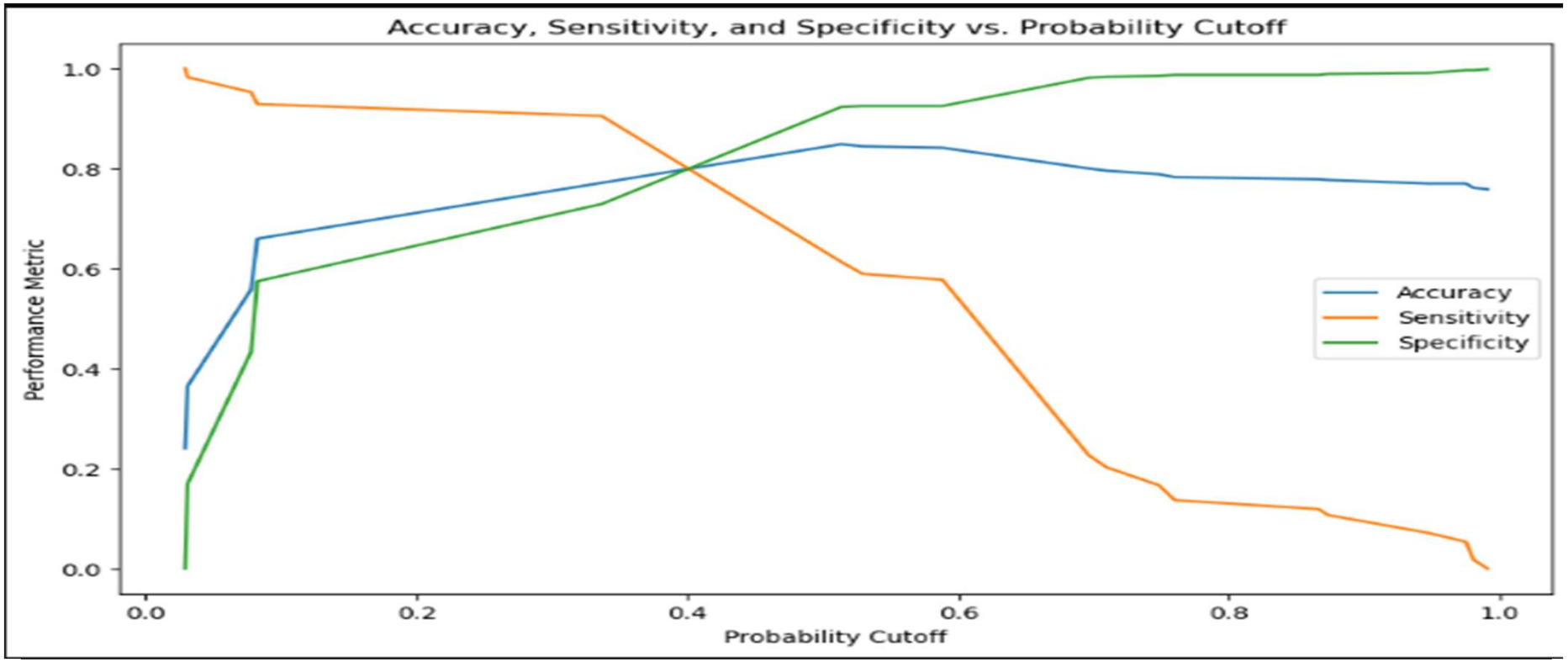
LIKELIHOOD FOR CATEGORICAL FEATURES



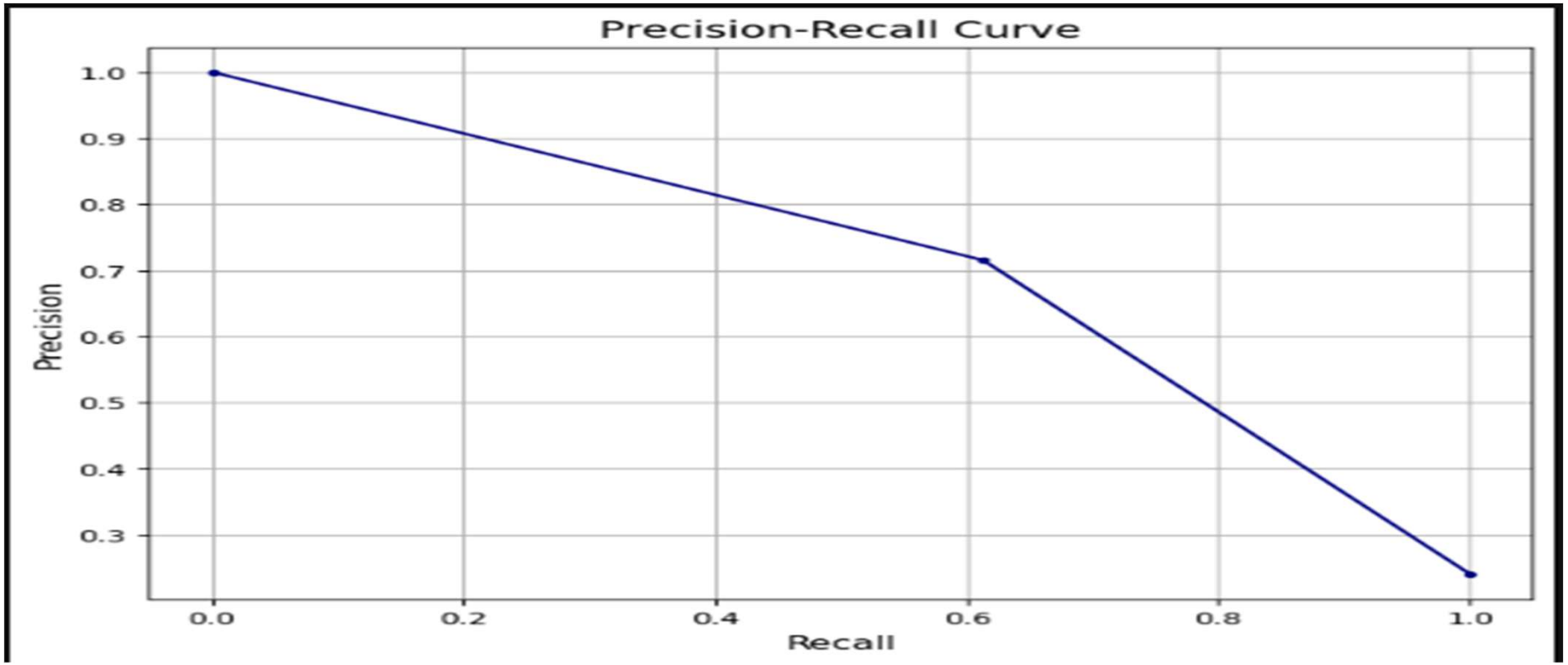
NUMERICAL FEATURES AND THE TARGET VARIABLE


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.87)

ROC CURVE



ACCURACY, SENSITIVITY, AND SPECIFICITY VS. PROBABILITY CUTOFF



PRECISION-RECALL CURVE

Key Insights

- Fraudulent claims often exhibit distinct patterns in terms of claim timing, incident severity, and policy details.
- Behavioural features (e.g., insured relationship, hobbies, claim amount) are used to identify suspicious activity.
- The most predictive features are incident_severity, insured_relationship, property_damage, incident_city and auto_make.
- Robust against overfitting and capable of handling both numerical and categorical data.
- Feature importance from the model provided interpretable insights into fraud indicators.
- Handling missing values, encoding categorical data, and transforming timestamps significantly boosted model performance.
- Multicollinearity checks and RFECV helped refine the model to include only the most informative features.

Evaluation & Conclusion

Random forest model is effective model for fraud detection. Careful preprocessing and feature selection improved model performance. This solution enables early detection of fraud, reducing financial loss and improving operational efficiency.

Recommendations

- Deploy Random Forest in real-time claims system with probability cutoff.
- Regularly retrain model with updated data.
- Focus investigations on high-risk patterns identified.
- Automate initial fraud flagging to assist human reviewers.

Thank You