

# Fraudulent Claim Detection Case Study

Submitted By

Pradeep Harry Michael

Pavan Kumar BN

# Contents



## Problem Statement



## Analysis and Model Building

- 1> Data Preparation and Cleaning
- 2> EDA
- 3> Feature Engineering
- 4> Model Building (LR / Random Forest)
- 5> Predicting and Evaluation



## Business Summary



# Problem Statement

- **Challenge:** Global Insure, a major player in the insurance industry, is experiencing substantial financial losses due to a high volume of fraudulent claims. The existing fraud detection system relies heavily on manual inspections, which are not only time-consuming and labour-intensive but also inefficient. As a result, many fraudulent claims are detected only after payouts have been made, limiting the company's ability to prevent losses and straining operational resources.
- **Objective:** To address this issue, Global Insure seeks to enhance its fraud detection capabilities by leveraging data-driven insights and advanced analytics. The goal is to implement an intelligent system that can accurately classify claims as fraudulent or legitimate at an early stage in the approval process. This proactive approach would help the company significantly reduce financial losses, improve the speed and accuracy of claims processing, and optimize overall efficiency in claims management.



# Data Preparation and Cleaning

## Data Understanding and Cleaning activities performed

- Drop Columns which do not have any values across all Rows.
- Drop Rows which have all column values as NA or Null
- Data type Changes
- Drop Columns which have large portion of values as unique
- Drop rows where features have invalid negative values
- Handle missing data like '?' as 'Unknown'

## Outcome

- The initial Data Frame of shape(rows , columns) (1000,40) was reduced to (908, 36)
- The Key Columns for analysis were identified

### Numerical Fields

- 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, auto\_year

### Categorical Field

- 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, incident\_location, property\_damage
- police\_report\_available, auto\_make, auto\_model

### DateField

incident\_date  
policy\_bind\_date

Target Field  
fraud\_reported

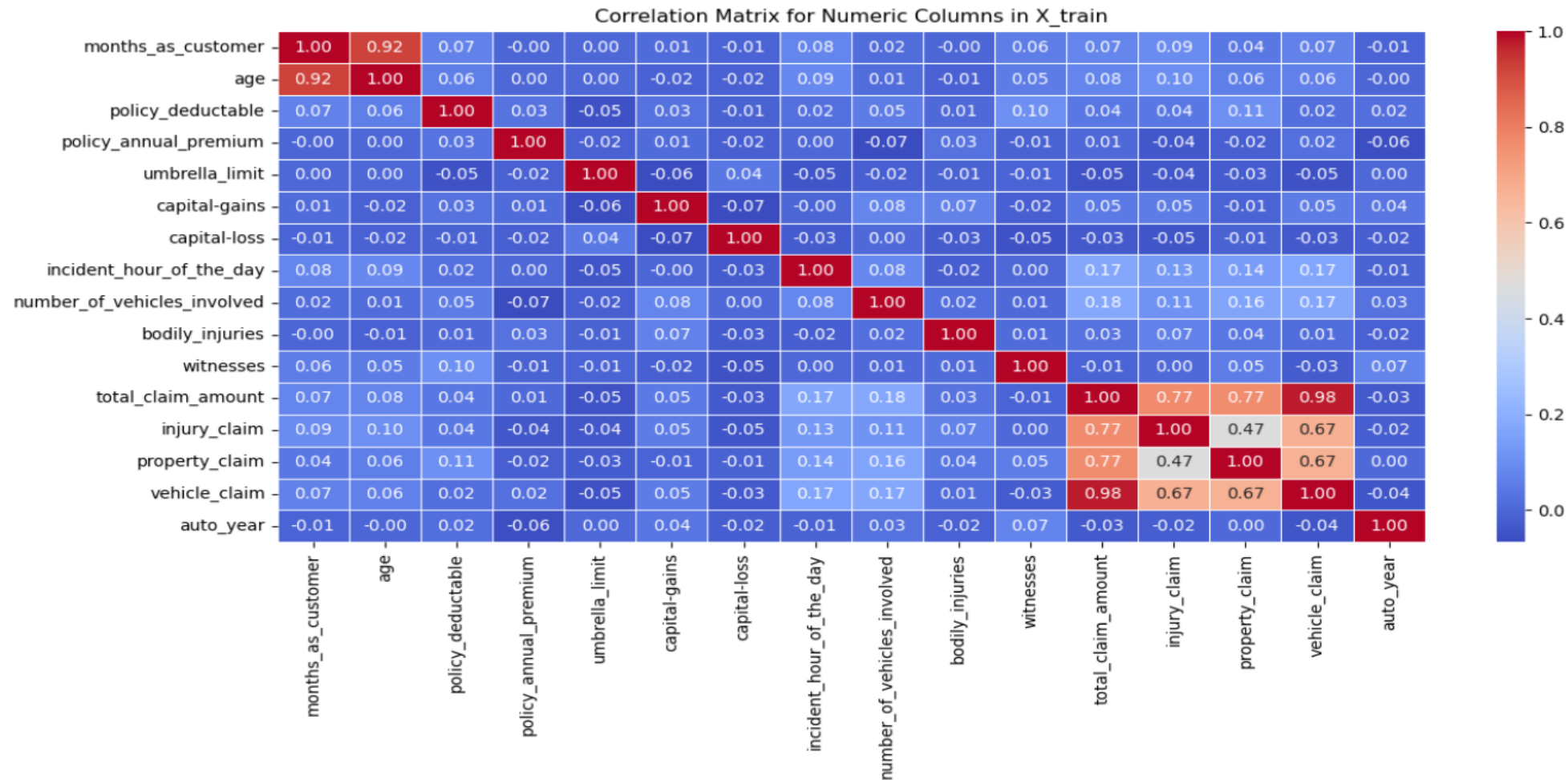


# EDA : Numeric Correlation

High Correlation Identified between the following fields

1> vehicle\_claim and total\_claim\_amount

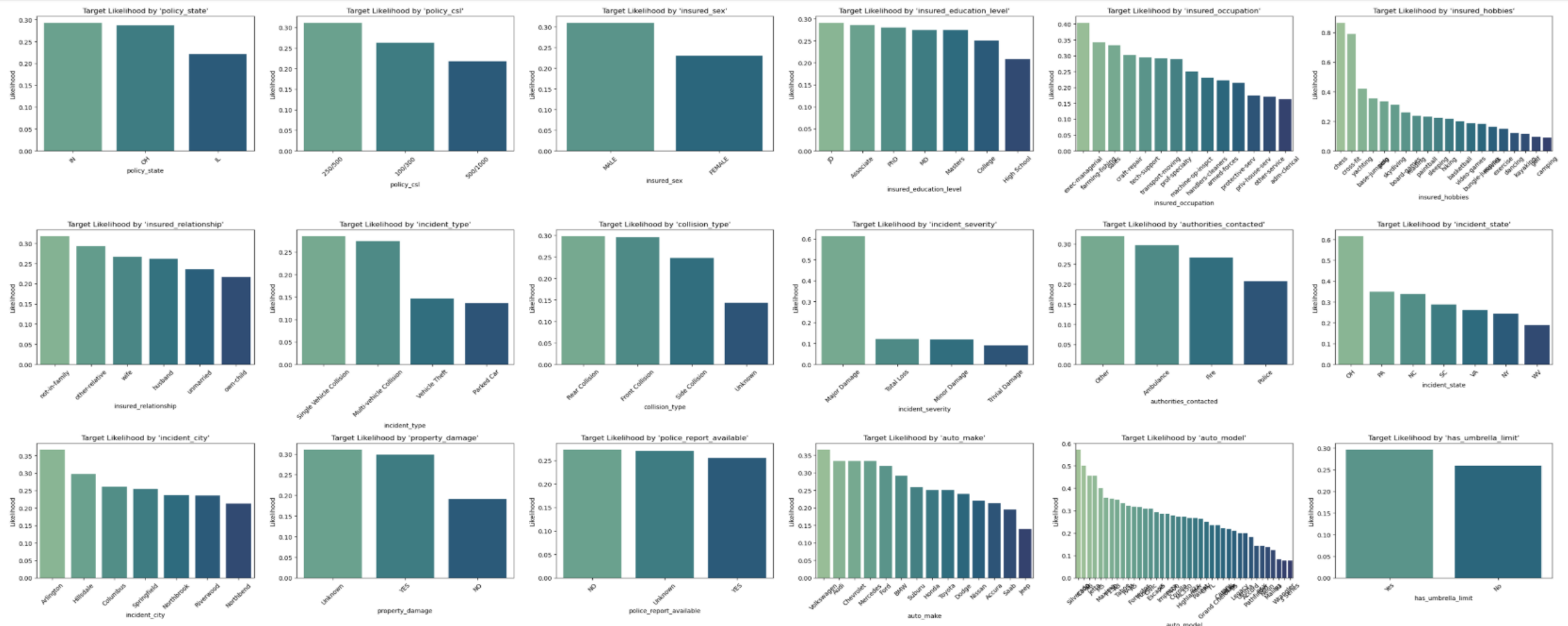
2> months\_as\_customer and age





# EDA : Target Likelihood for Categorical Variables

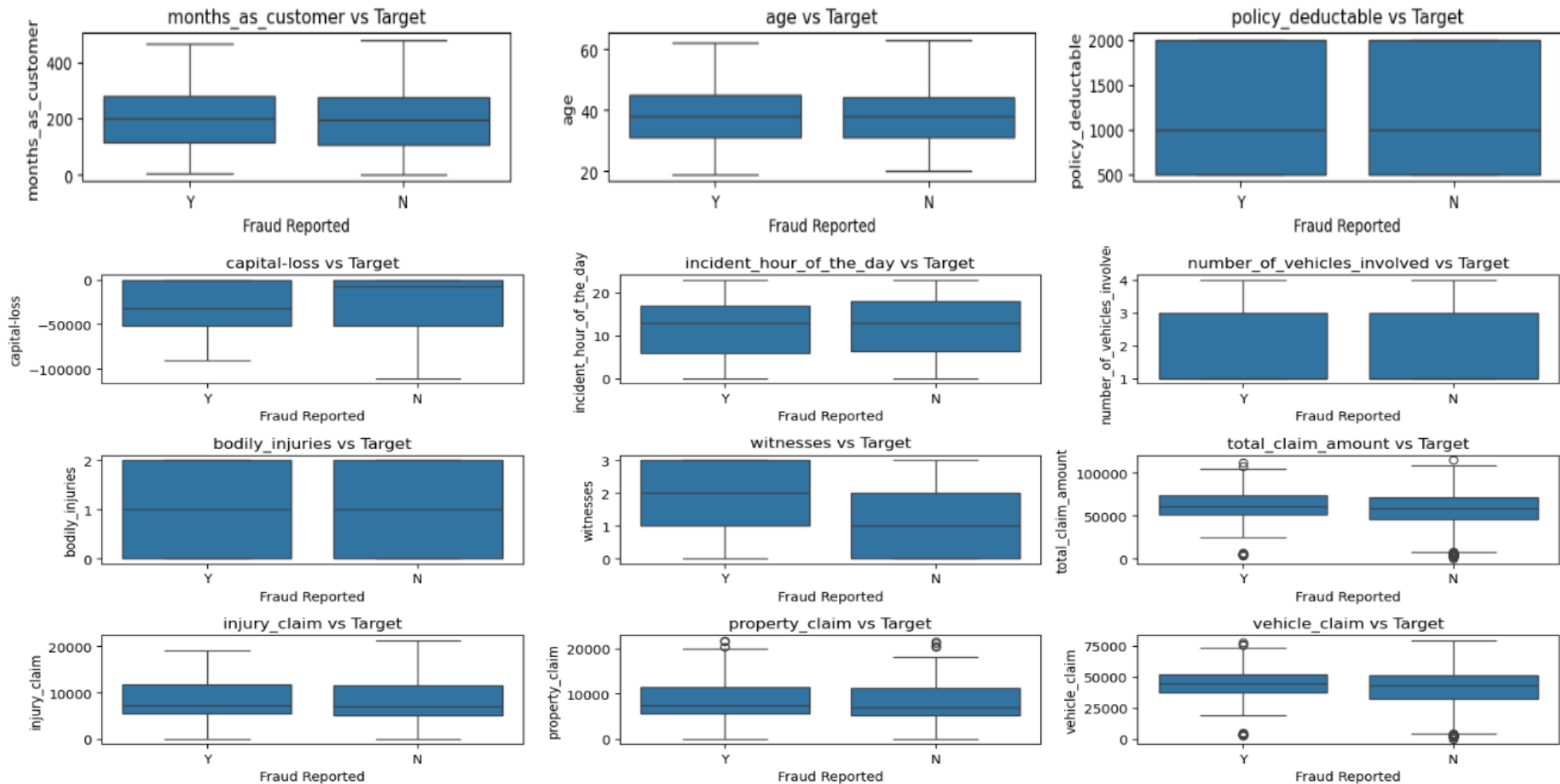
Categorical Variables like **insured\_hobbies**, **insured\_occupation**, **incident\_state** , **incident\_severity** and **auto make/model** have more impact of the target likelihood





# EDA : Target Likelihood for Numerical Variables

Numerical Variables like **capital-loss**, **wtnesses** have more impact of the target likelihood

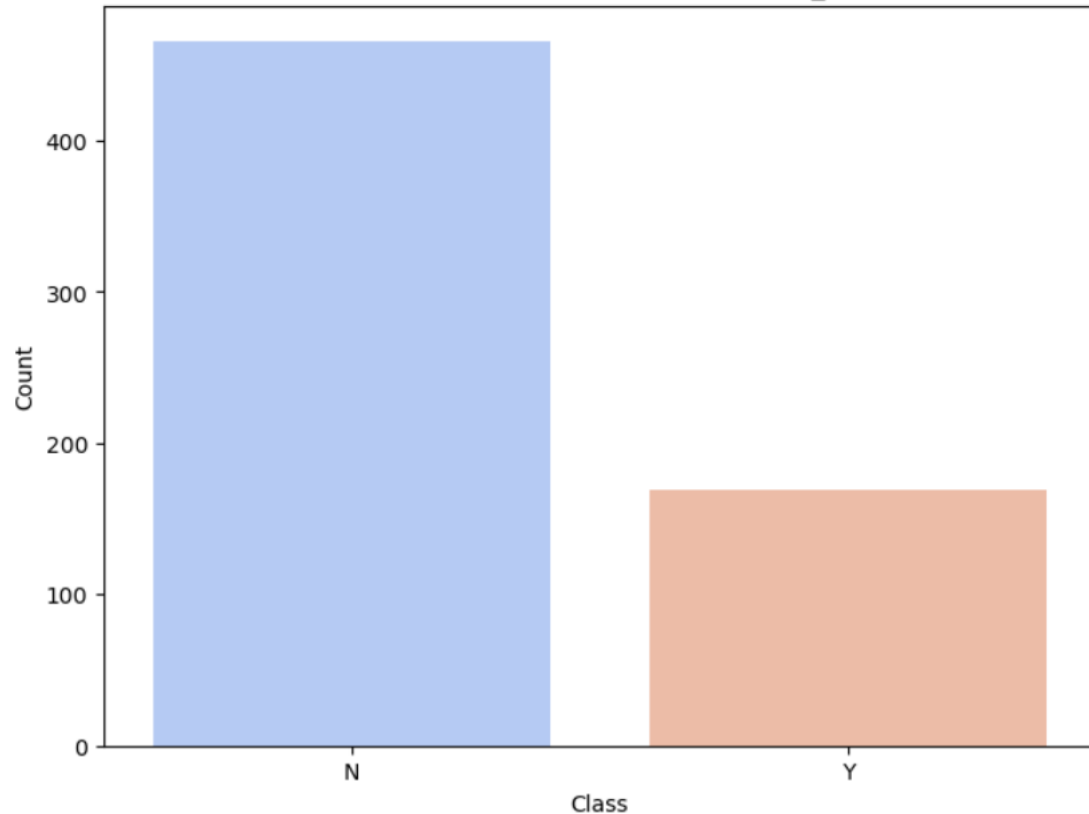




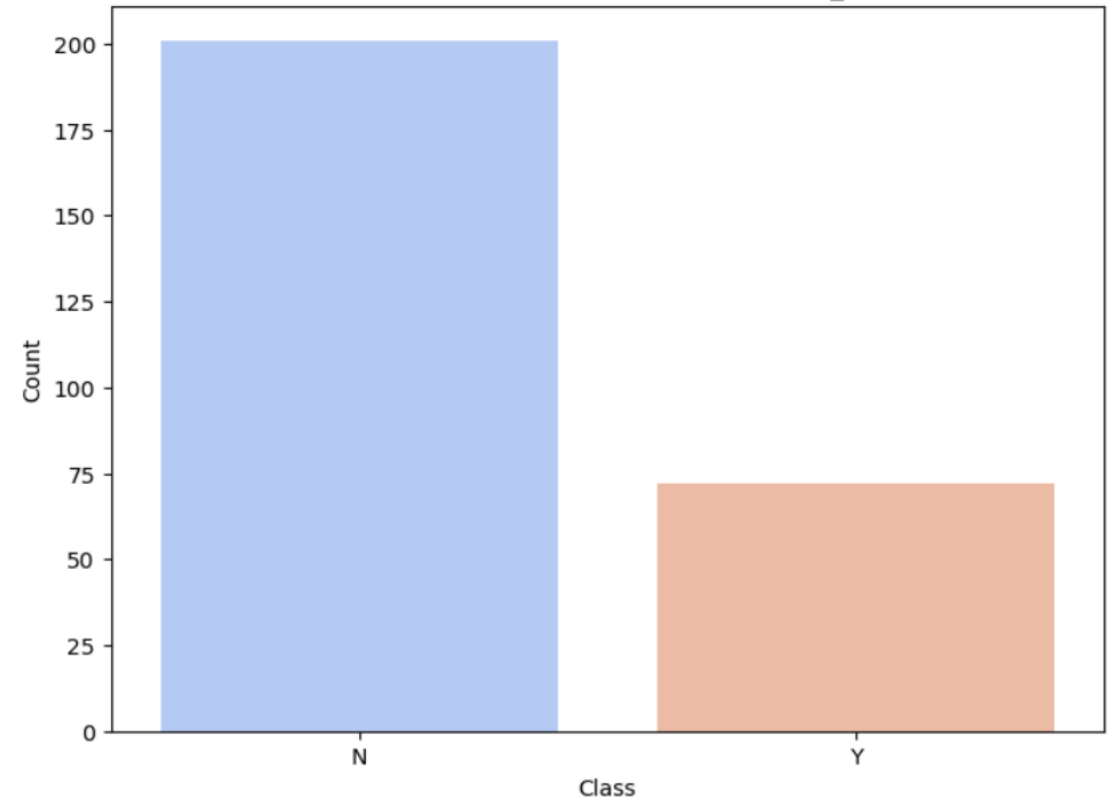
# Feature Engineering :

- Resampling using RandomOverSampler to handle Class Imbalance
- Deriving new Features from Date Columns
- Combining Features like auto model and auto make
- Handle Redundant Columns
- Dummy Variables for Categorical Columns
- Feature Scaling

Class Distribution in Target Variable (y\_train)



Class Distribution in Target Variable (y\_val)



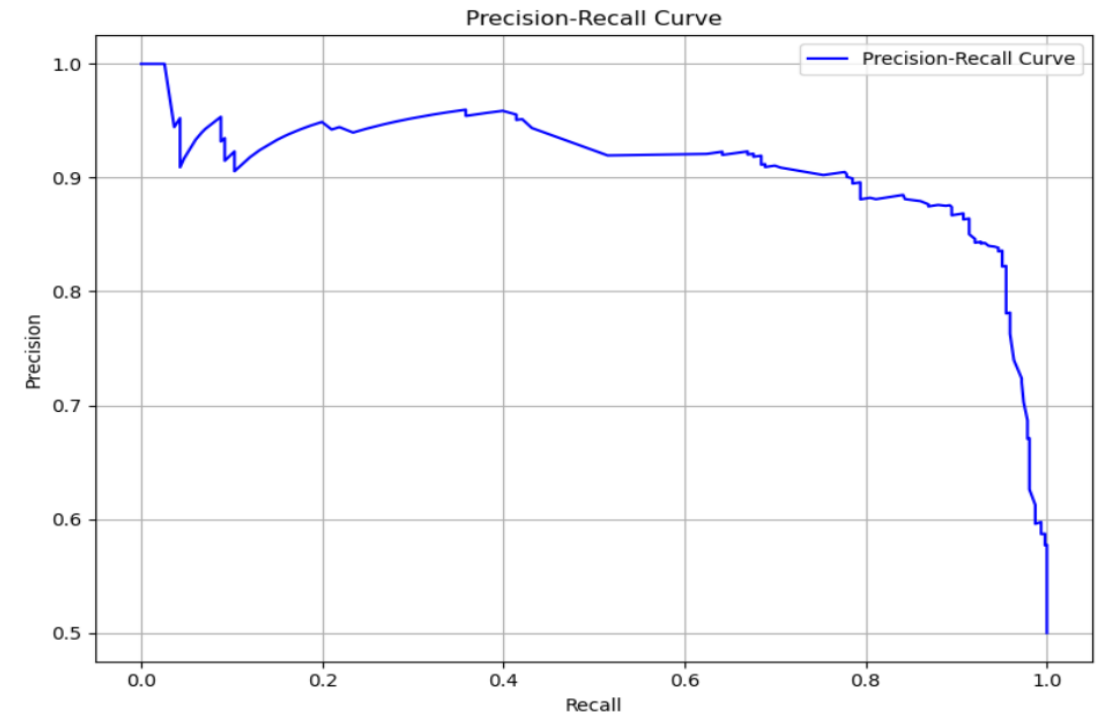
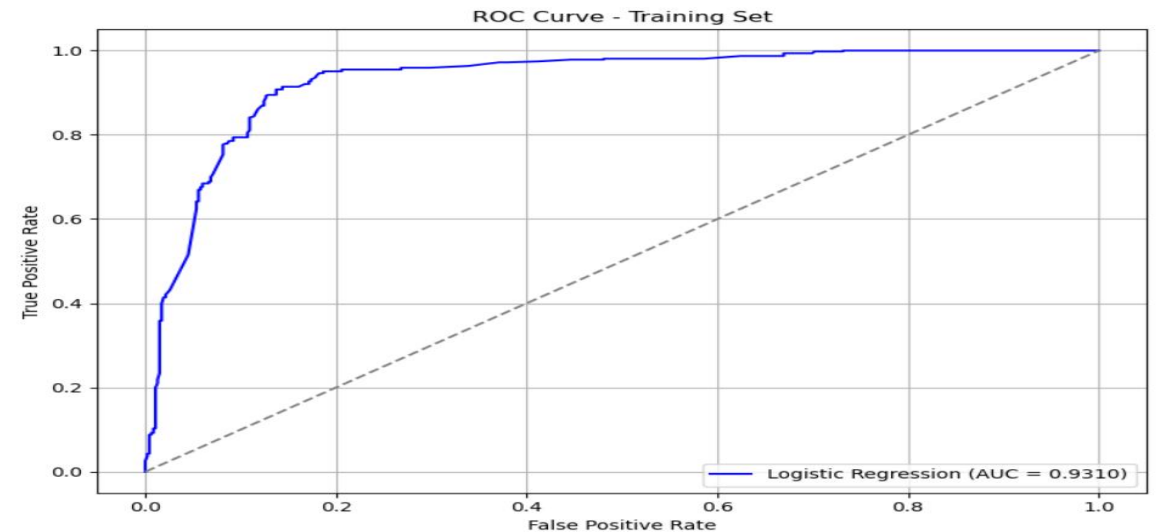
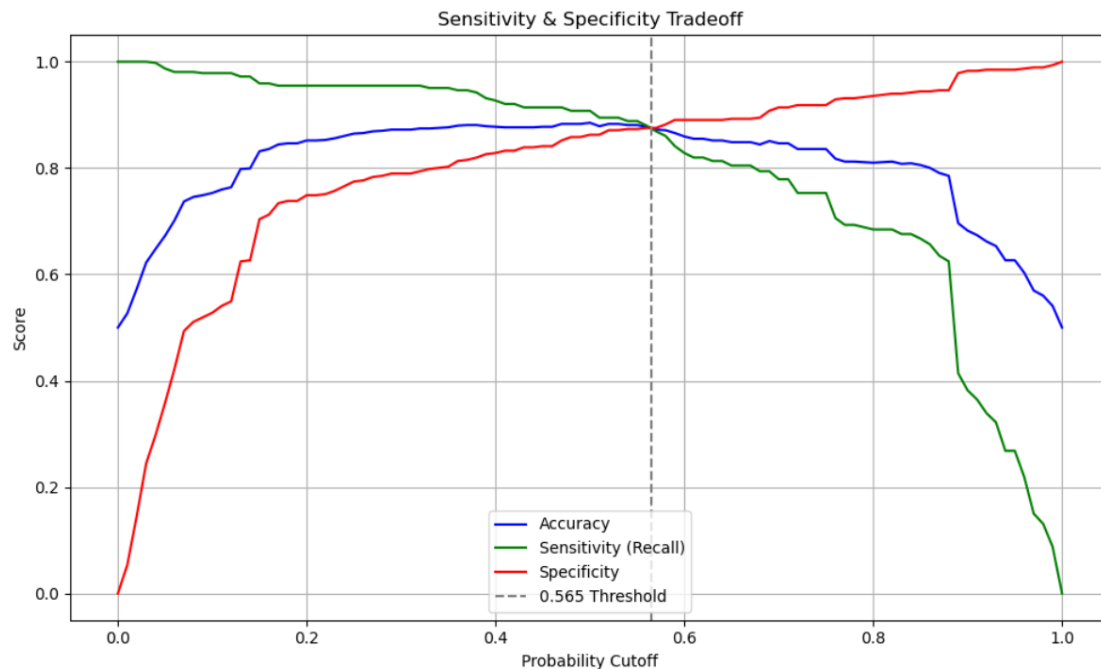




# Model Building : Logistic Regression

## Summary of Model (Training)

- **Probability Cutoff Chosen: 0.565**
- **Model Accuracy at Optimal Cutoff (0.565): 0.8766**
- **Sensitivity (Recall): 0.8777**
- **Specificity: 0.8755**
- **Precision: 0.8758**
- **Recall: 0.8777**
- **F1 Score: 0.8767**



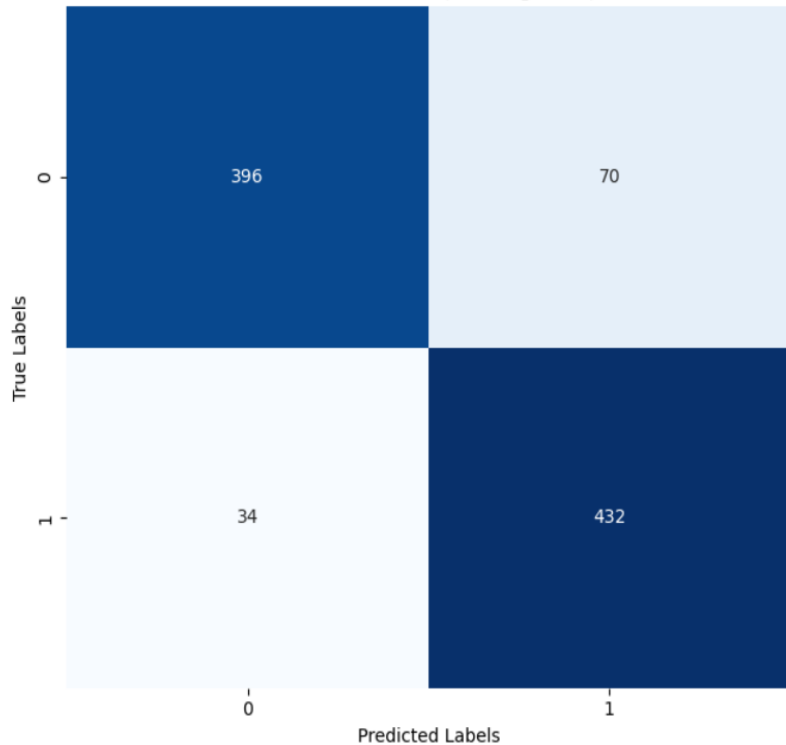


# Model Building : Random Regression

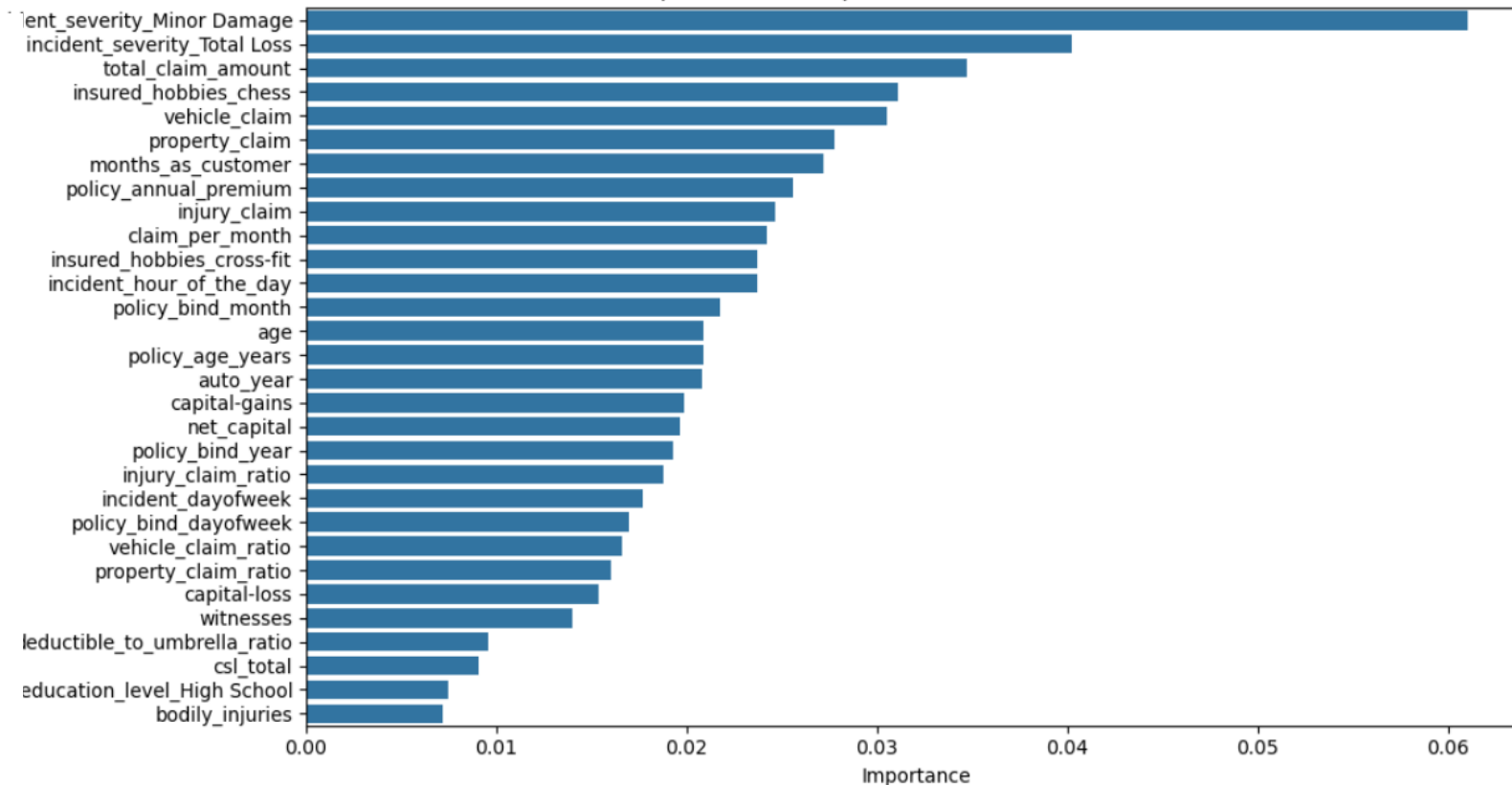
## Summary of Model (Training)

- **Best Hyperparameters Found:**  
{ 'max\_depth': 15, 'max\_features': 8, 'min\_samples\_leaf': 10, 'min\_samples\_split': 20, 'n\_estimators': 15 }
- **Training Accuracy:** 0.8884
- **Sensitivity (Recall):** 0.9270
- **Specificity:** 0.8498
- **Precision:** 0.8606
- **F1 Score:** 0.8926

Confusion Matrix (Training Data)



Top 30 Feature Importances from Random Forest





# Model Building : Prediction and Evaluation

Model	Tuning	Training Set Performance	Validation Set Performance
Logistic Regression	<b>Probability Cutoff Chosen:</b> 0.565	<b>Model Accuracy at Optimal Cutoff (0.565):</b> 0.8766 <b>Sensitivity (Recall):</b> 0.8777 <b>Specificity:</b> 0.8755 <b>Precision:</b> 0.8758 <b>Recall:</b> 0.8777 <b>F1 Score:</b> 0.8767	<b>Model Accuracy at Optimal Cutoff (0.565):</b> 0.8388 <b>Sensitivity (Recall):</b> 0.8194 <b>Specificity:</b> 0.8458 <b>Precision:</b> 0.6556 <b>Recall:</b> 0.8194 <b>F1 Score:</b> 0.7284
Random Forest	<b>Best Hyperparameters Found:</b> {'max_depth': 15, 'max_features': 8, 'min_samples_leaf': 10, 'min_samples_split': 20, 'n_estimators': 15}	<b>Training Accuracy:</b> 0.8884 <b>Sensitivity (Recall):</b> 0.9270 <b>Specificity:</b> 0.8498 <b>Precision:</b> 0.8606 <b>F1 Score:</b> 0.8926	<b>Validation Accuracy:</b> 0.8352 <b>Sensitivity (Recall):</b> 0.8750 <b>Specificity:</b> 0.8209 <b>Precision:</b> 0.6364 <b>F1 Score:</b> 0.7368