# Fraudulent Claim Detection CaseStudy-Report

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.

(38577, 54)

# 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

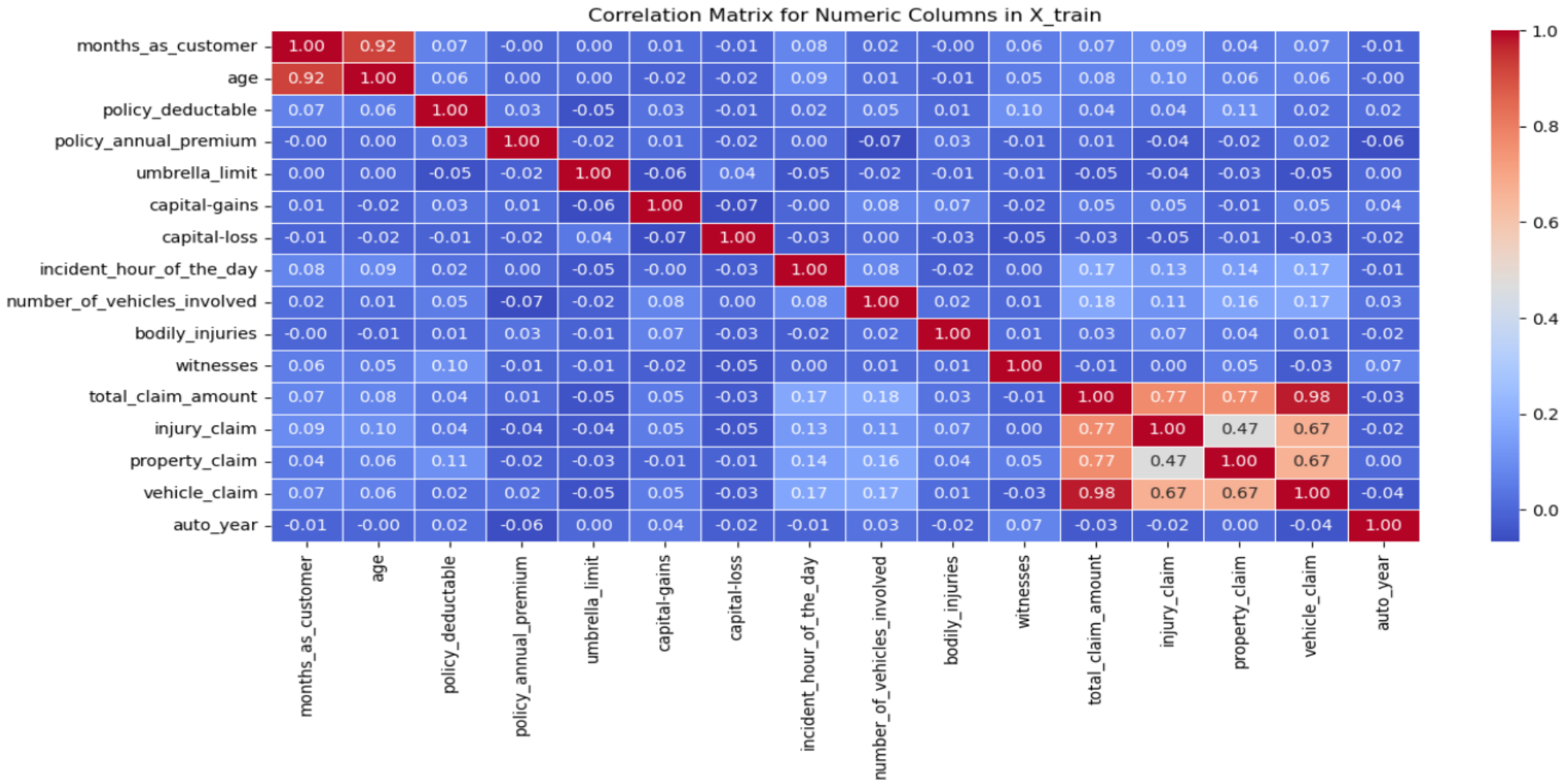
Target Field
fraud_reported

# EDA : Numeric Correlation

High Correlation Identified between the following fields

1> vehicle_claim and total_claim_amount

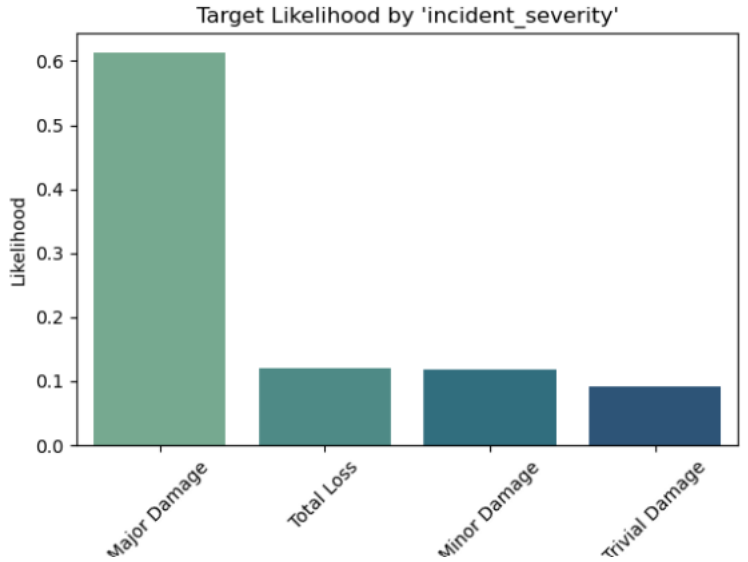
2> months_as_customer and age

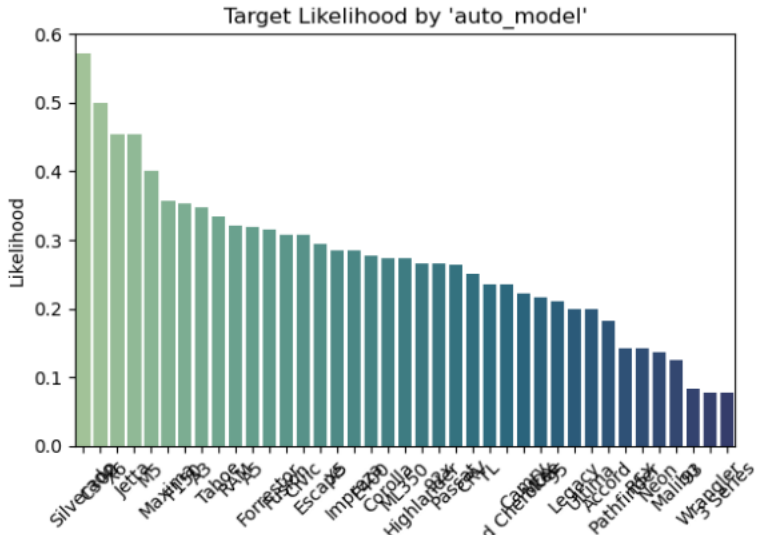


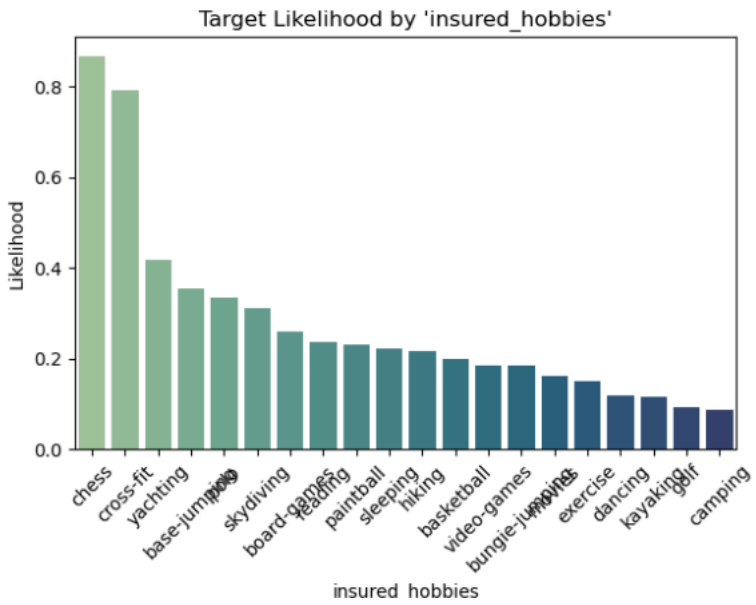
# EDA : Target Likelihood for Categorical Variables

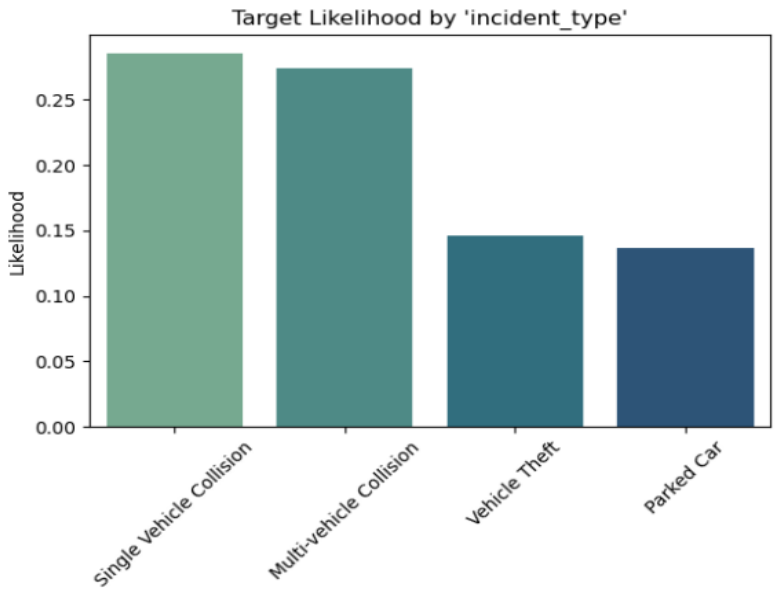
Top 5 Categorical Features with Highest Variation in 'Y' Likelihood:

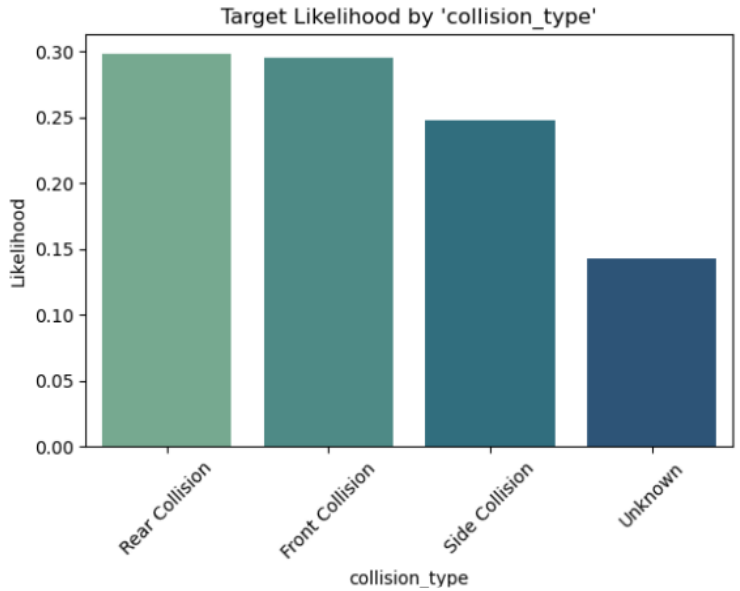
- incident_severity
- auto_model
- insured_hobbies
- incident_type
- collision_type



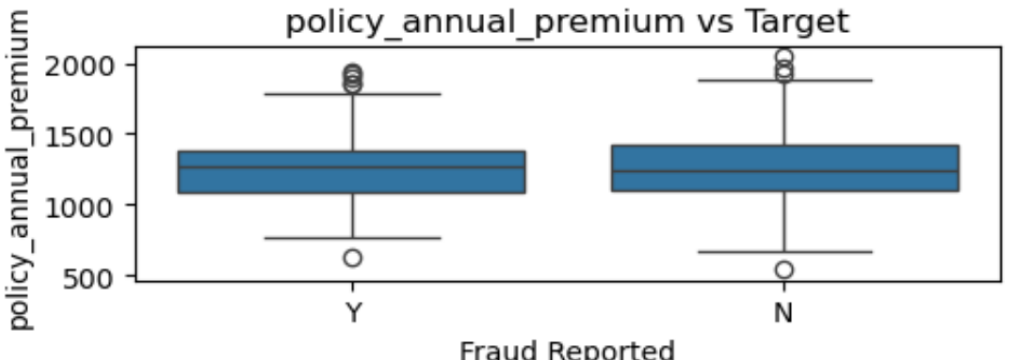



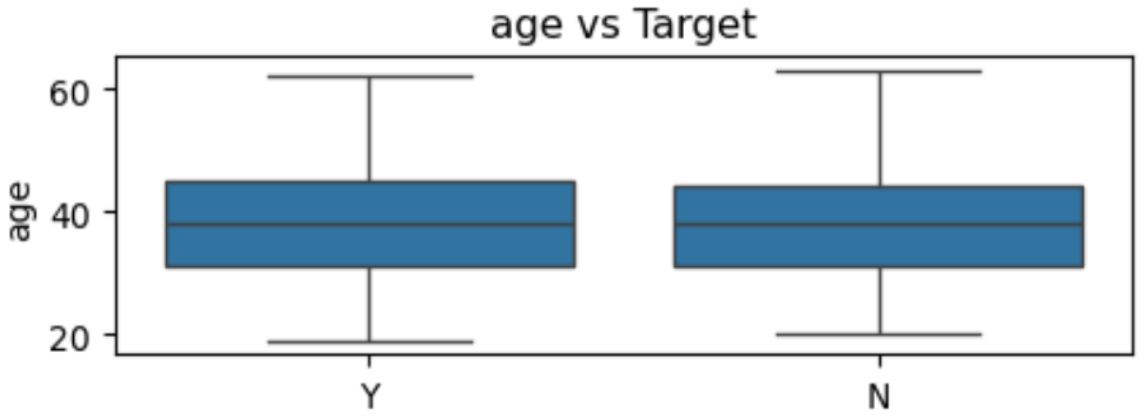



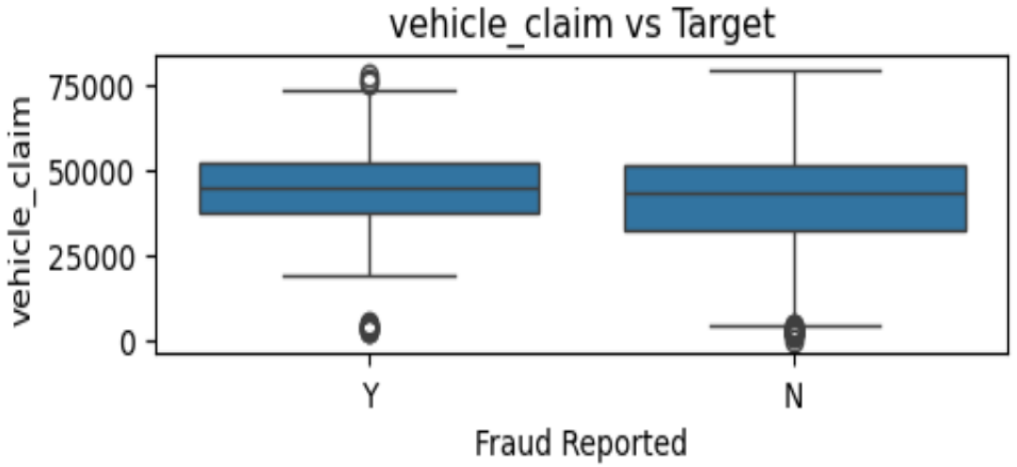
# EDA : Target Likelihood for Numerical Variables

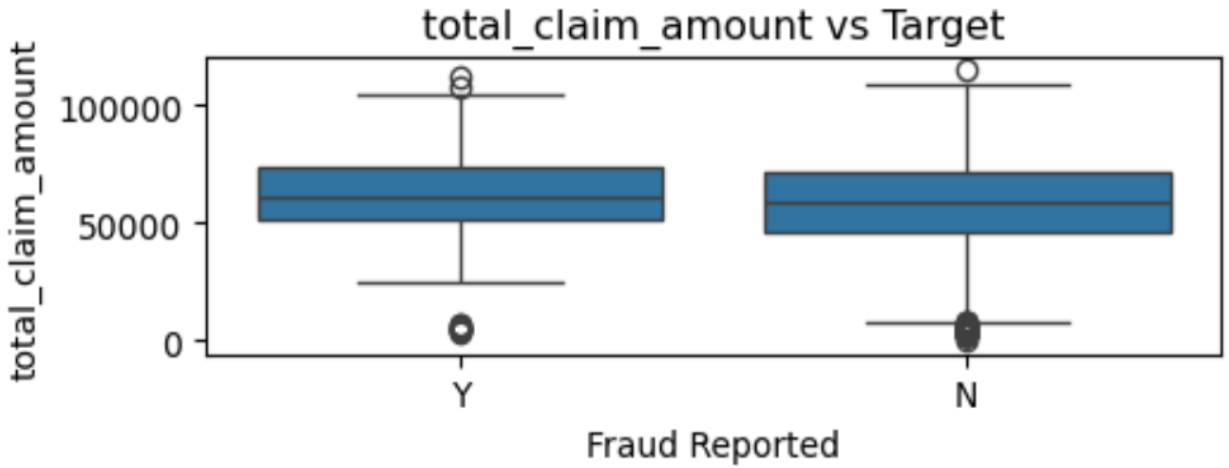
Top 5 Numerical Features with Highest Variation in 'Y' Likelihood:

- policy_annual_premium
- number_of_vehicles_involved
- total_claim_amount
- vehicle_claim
- age



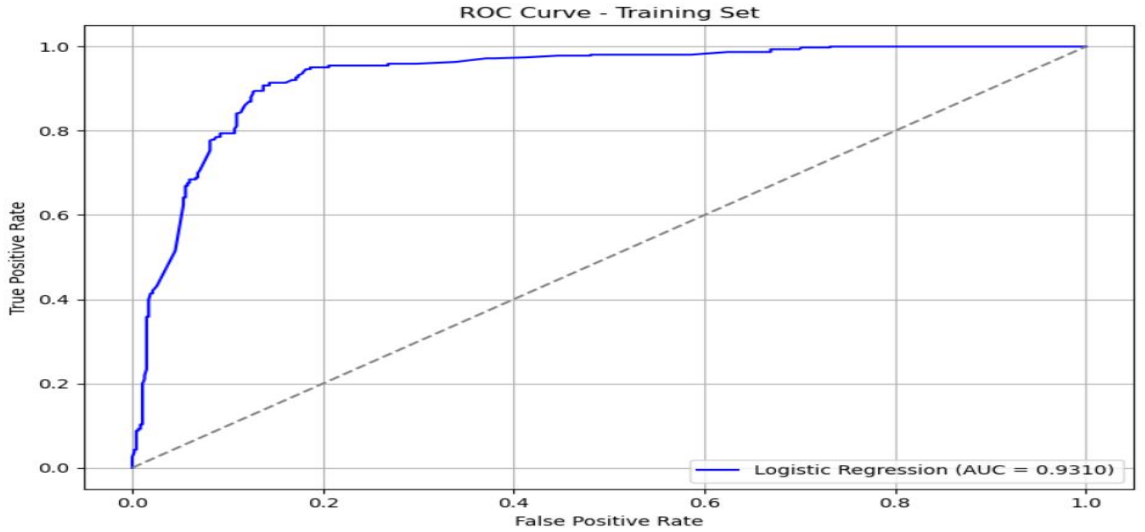
# Feature Engineering :

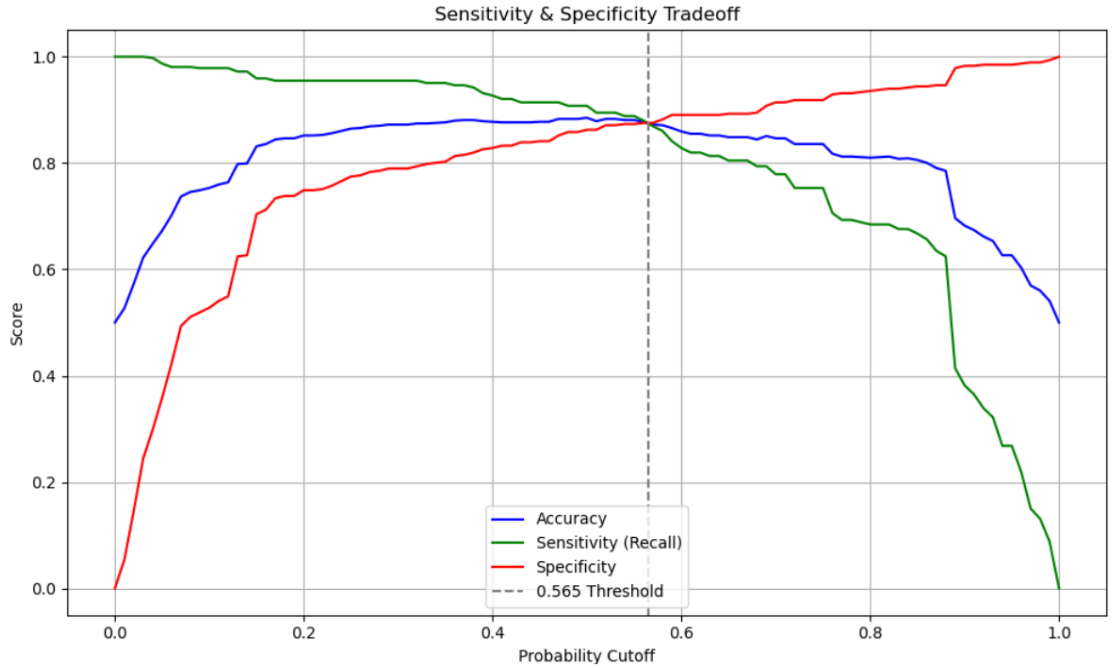
- Resampling using RandomOverSampler to handle Class Imbalance
- Deriving new Features from Date Columns
- Combining Features like auto model and auto make
- Handle Redundant Columns
- Dummy Variables for Categorical Columns
- Feature Scaling

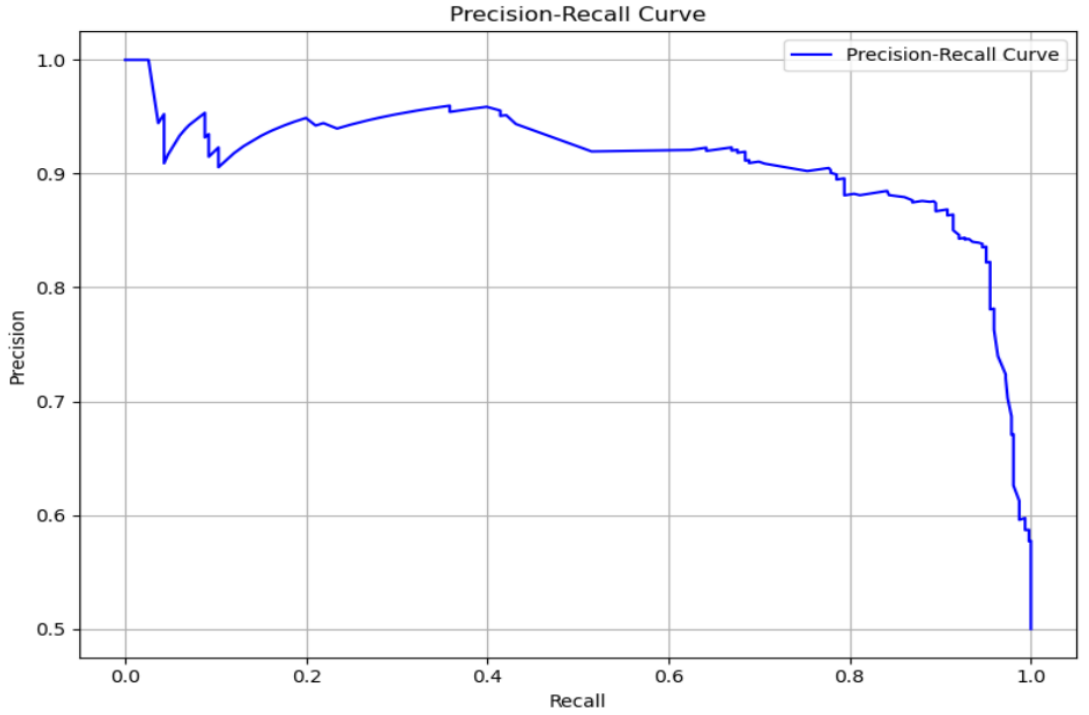

# Model Building : Logistic Regression

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



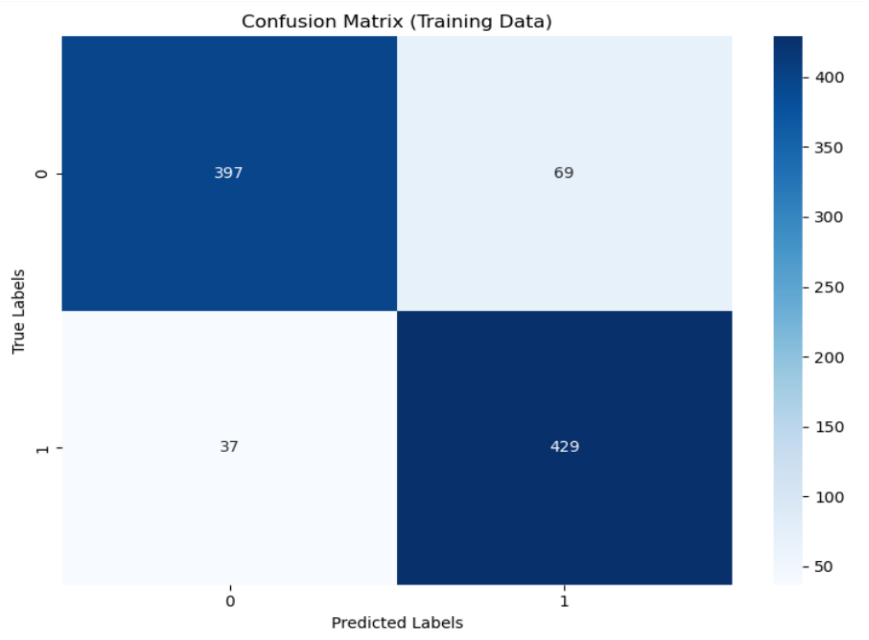



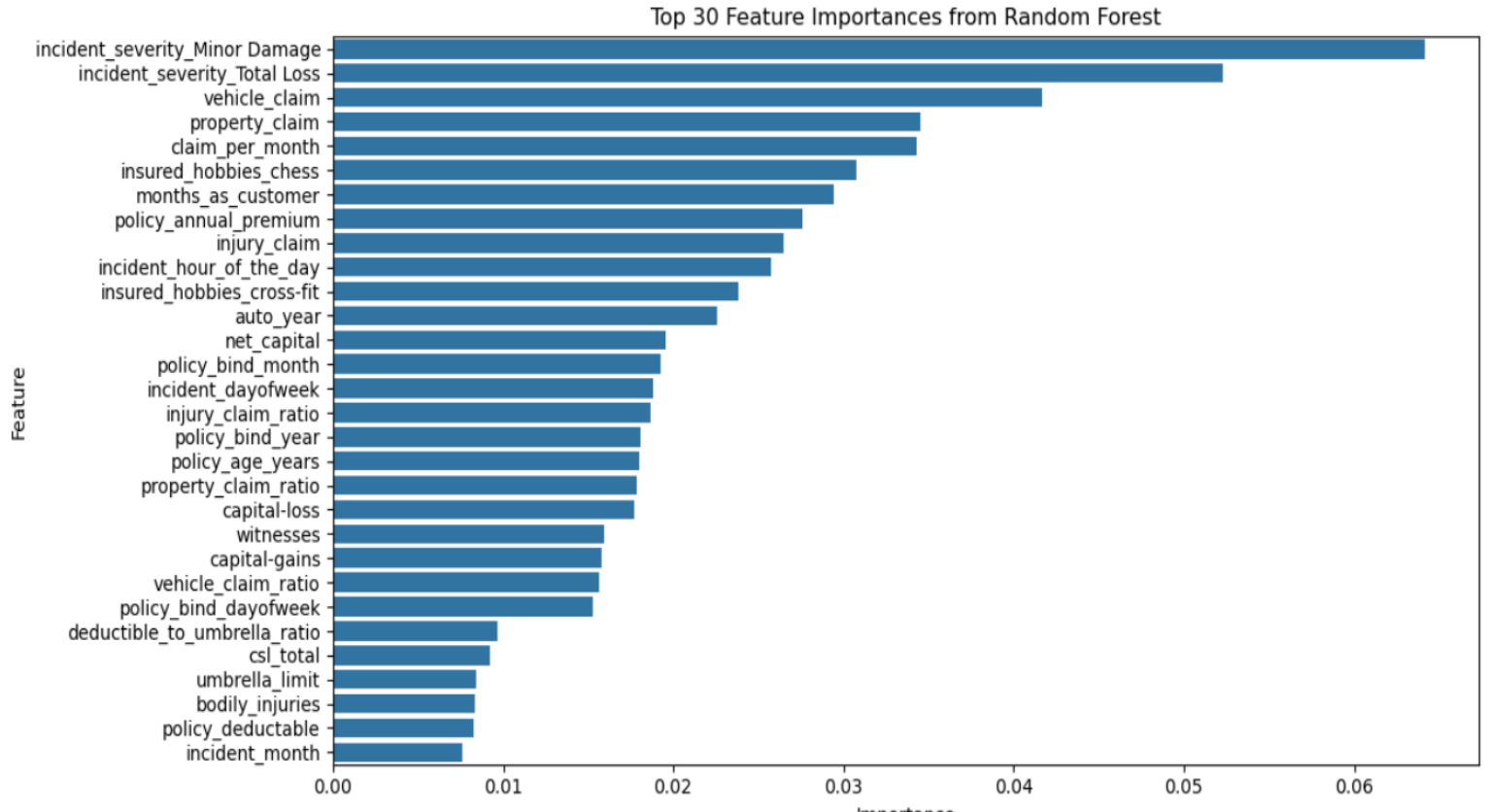
# Model Building : Random Regression

Summary of Model (Training)

- **Best Hyperparameters Found:**
  {'max_depth': 10, 'max_features': 12, 'min_samples_leaf': 10, 'min_samples_split': 20, 'n_estimators': 20}

- **Training Accuracy:** 0.8863
- **Sensitivity (Recall):** 0.9206
- **Specificity:** 0.8519
- **Precision:** 0.8614
- **F1 Score:** 0.8900



# Model Building : Prediction and Evaluation

| Model | Tuning | Training Set Performance | Validation Set Performance |
|---|---|---|---|
| Logistic Regression | **Probability Cutoff Chosen:** 0.565 | **Model Accuracy at Optimal Cutoff (0.565):** 0.8766<br>**Sensitivity (Recall):** 0.8777<br>**Specificity:** 0.8755<br>**Precision:** 0.8758<br>**Recall:** 0.8777<br>**F1 Score:** 0.8767 | **Model Accuracy at Optimal Cutoff (0.565):** 0.8388<br>**Sensitivity (Recall):** 0.8194<br>**Specificity:** 0.8458<br>**Precision:** 0.6556<br>**Recall:** 0.8194<br>**F1 Score:** 0.7284 |
| Random Forest | **Best Hyperparameters Found:**<br>{'max_depth': 10,<br>'max_features': 12,<br>'min_samples_leaf': 10,<br>'min_samples_split': 20,<br>'n_estimators': 20} | •**Training Accuracy:** 0.8863<br>•**Sensitivity (Recall):** 0.9206<br>•**Specificity:** 0.8519<br>•**Precision:** 0.8614<br>•**F1 Score:** 0.8900 | **Validation Accuracy:** 0.8388<br>**Sensitivity (Recall):** 0.8889<br>**Specificity:** 0.8209<br>**Precision:** 0.6400<br>**F1 Score:** 0.7442 |

(38577, 54)

# Business Summary

**A data-driven approach to analyze historical claim records has revealed clear patterns associated with fraudulent behaviour.**

**Both Logistic Regression and Random Forest models were trained to predict fraud probability, with Random Forest achieving slightly better performance.**

**Categorical and numerical likelihood analysis identified features that significantly influence the likelihood of fraud.**

**High-variation features like incident_severity, insured_hobbies , total_claim_amount(vehicle/property/injury), policy_annual_premium, months_as_customer were strong fraud indicators.**

**Low-impact features (e.g., insured_sex, policy_state) contributed minimally and can be deprioritized or dropped to reduce noise.**