# Insurance Fraud Claim Detection Case Study Report

## Participants

Syeda Farhana Yasmin

## Introduction

Global Insure aims to develop a predictive model that classifies insurance claims as fraudulent or legitimate using historical claim data and customer profiles. The goal is to proactively identify potentially fraudulent claims before they are approved by analyzing past patterns and key indicators such as claim amounts, claim types, and customer behavior. This involves detecting fraud-related trends in historical data, identifying the most predictive features of fraudulent activity, and building machine learning models to estimate the likelihood of fraud in incoming claims. Additionally, the insights generated from the model will be used to enhance the overall fraud detection process by improving decision-making, optimizing investigation efforts, and enabling early intervention

## Fraudulent Claim Detection Approach

1. Data Understanding

2. Data Pre-processing

3. Train Test Split

4. EDA on Training Data

5. Feature Engineering

6. Model Building

7. Predicting and Model Evaluation

## Data Understanding & Pre-Processing

The dataset provided in CSV format is the historical claim details with customer profile. It is a dataset of 1000 rows and 40 columns. Following actions were taken in data pre-processing:

- Examine the columns to determine if any value or column needs to be treated
- Identify and handle redundant values and columns
- Fix The Data Types

## Train - Test Split

Breakout and define the predictive features and target variable. Split the data into train and test datasets using stratification. The train-test split is the ratio of 70-30.

**Exploratory Data Analysis**

Descriptive Statistics

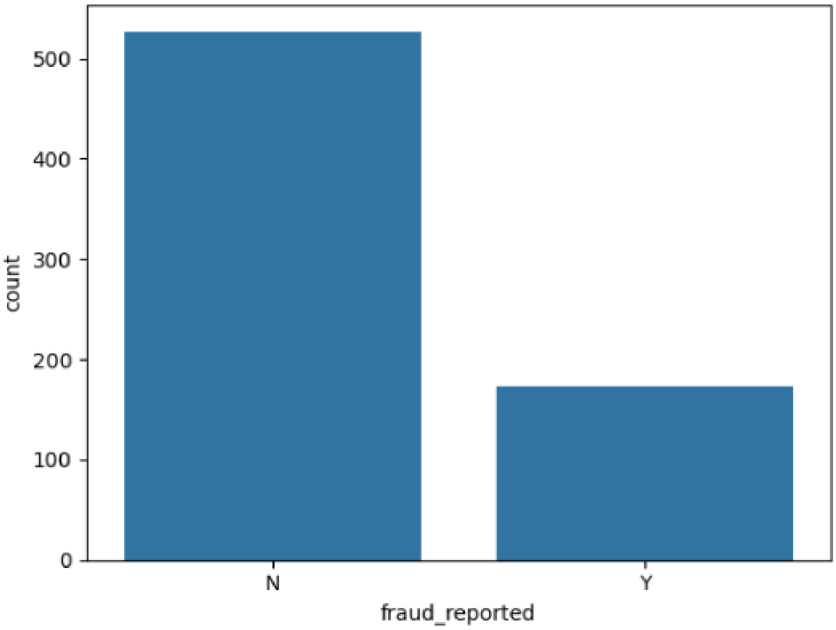
Summarizing the main features of the data:

- Majority of customer with peak concentration is with Global Insure for 100 to 300 months.
- Majority of customers are between 30 to 45 years of age.
- Most customers have annual premium between 800 and 1200.
- The vast majority (672 cases) involved only 1 vehicle, indicating mostly single-vehicle incidents or claims.
- Most claims are for claim amount of 50K to 65K.
- Out of the total claims – ~500 claims are genuine and ~180 are fraudulent claims.
- Using Likelihood analysis and p_value analysis, most of the categorical features do not significantly affect the target variable. Only significant features are 'incident_state' and 'authorities_contacted'.

**Visualization**

Used count plots, scatter plots and heatmap to understand data distributions and relationships.

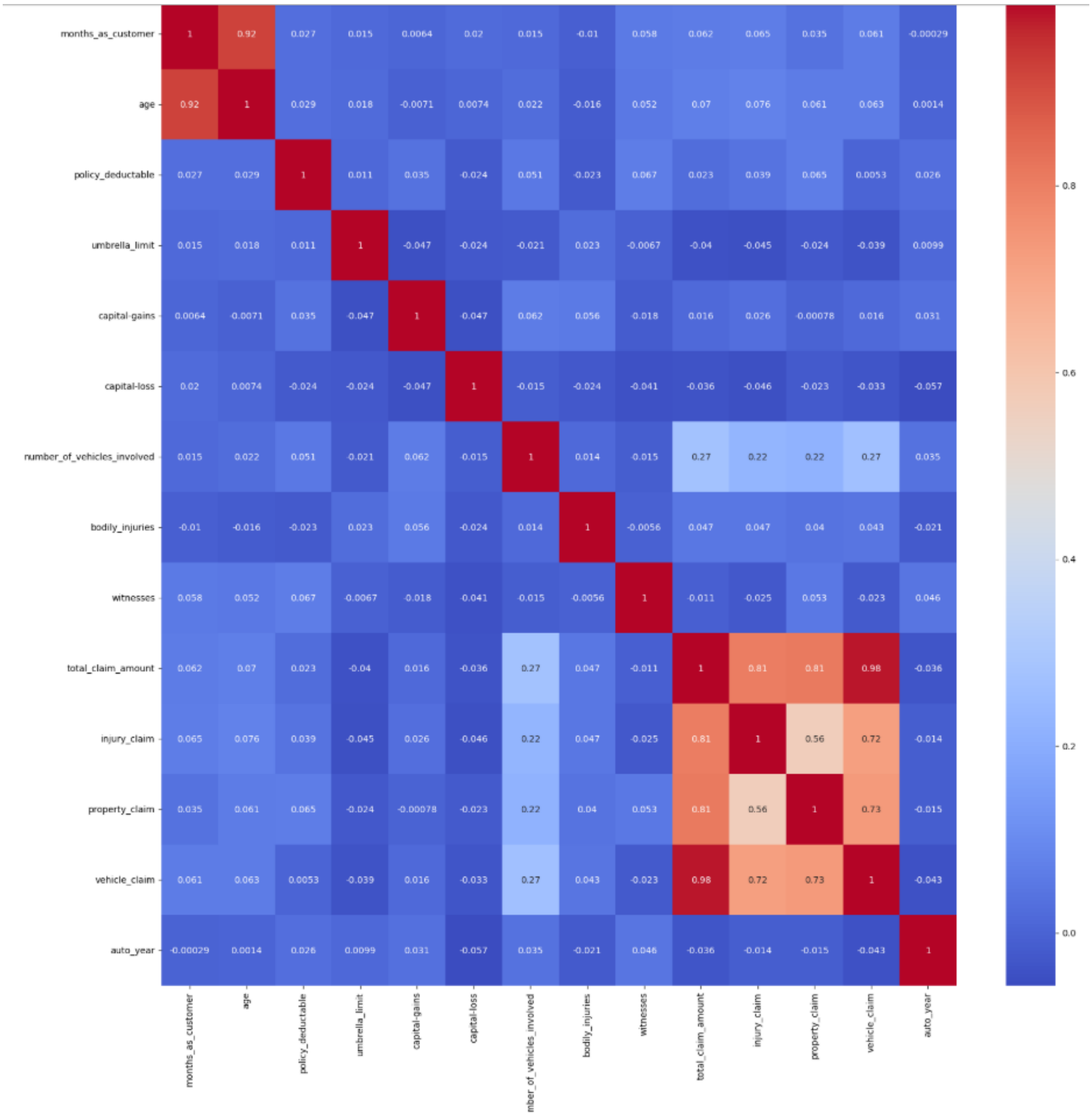
1. Distribution of Target Variable to highlight the data imbalance.



2. Heatmap indicating the corelation between features.

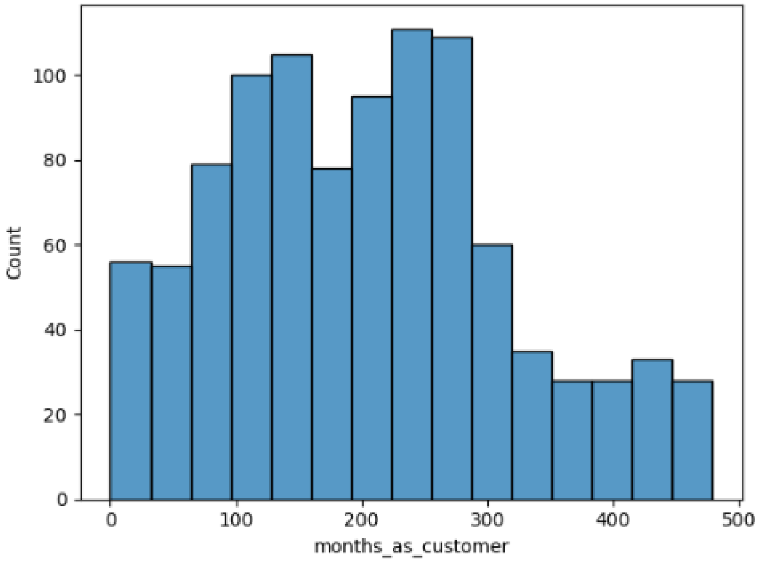
a. It could be seen that total_claim_amount is highly corelated with injury_claim, property_claim and vehicle claim.

b. Age is highly corelated with months_as_customer.

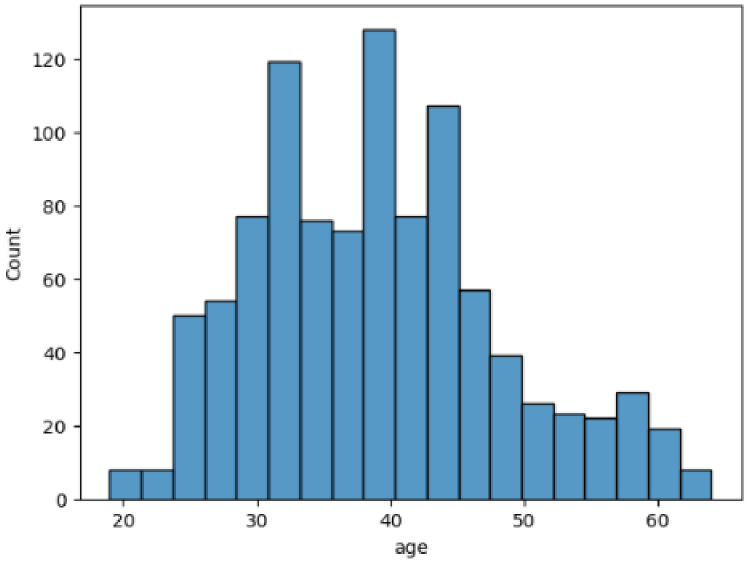


| | months_as_customer | age | policy_deductable | umbrella_limit | capital-gains | capital-loss | mber_of_vehicles_involved | bodily_injuries | witnesses | total_claim_amount | injury_claim | property_claim | vehicle_claim | auto_year |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| months_as_customer | 1 | 0.92 | 0.027 | 0.015 | 0.0064 | 0.02 | 0.015 | -0.01 | 0.058 | 0.062 | 0.065 | 0.035 | 0.061 | -0.00029 |
| age | 0.92 | 1 | 0.029 | 0.018 | -0.0071 | 0.0074 | 0.022 | -0.016 | 0.052 | 0.07 | 0.076 | 0.061 | 0.063 | 0.0014 |
| policy_deductable | 0.027 | 0.029 | 1 | 0.011 | 0.035 | -0.024 | 0.051 | -0.023 | 0.067 | 0.023 | 0.039 | 0.065 | 0.0053 | 0.026 |
| umbrella_limit | 0.015 | 0.018 | 0.011 | 1 | -0.047 | -0.024 | -0.021 | 0.023 | -0.0067 | -0.04 | -0.045 | -0.024 | -0.039 | 0.0099 |
| capital-gains | 0.0064 | -0.0071 | 0.035 | -0.047 | 1 | -0.047 | 0.062 | 0.056 | -0.018 | 0.016 | 0.026 | -0.00078 | 0.016 | 0.031 |
| capital-loss | 0.02 | 0.0074 | -0.024 | -0.024 | -0.047 | 1 | -0.015 | -0.024 | -0.041 | -0.036 | -0.046 | -0.023 | -0.033 | -0.057 |
| number_of_vehicles_involved | 0.015 | 0.022 | 0.051 | -0.021 | 0.062 | -0.015 | 1 | 0.014 | -0.015 | 0.27 | 0.22 | 0.22 | 0.27 | 0.035 |
| bodily_injuries | -0.01 | -0.016 | -0.023 | 0.023 | 0.056 | -0.024 | 0.014 | 1 | -0.0056 | 0.047 | 0.047 | 0.04 | 0.043 | -0.021 |
| witnesses | 0.058 | 0.052 | 0.067 | -0.0067 | -0.018 | -0.041 | -0.015 | -0.0056 | 1 | -0.011 | -0.025 | 0.053 | -0.023 | 0.046 |
| total_claim_amount | 0.062 | 0.07 | 0.023 | -0.04 | 0.016 | -0.036 | 0.27 | 0.047 | -0.011 | 1 | 0.81 | 0.81 | 0.98 | -0.036 |
| injury_claim | 0.065 | 0.076 | 0.039 | -0.045 | 0.026 | -0.046 | 0.22 | 0.047 | -0.025 | 0.81 | 1 | 0.56 | 0.72 | -0.014 |
| property_claim | 0.035 | 0.061 | 0.065 | -0.024 | -0.00078 | -0.023 | 0.22 | 0.04 | 0.053 | 0.81 | 0.56 | 1 | 0.73 | -0.015 |
| vehicle_claim | 0.061 | 0.063 | 0.0053 | -0.039 | 0.016 | -0.033 | 0.27 | 0.043 | -0.023 | 0.98 | 0.72 | 0.73 | 1 | -0.043 |
| auto_year | -0.00029 | 0.0014 | 0.026 | 0.0099 | 0.031 | -0.057 | 0.035 | -0.021 | 0.046 | -0.036 | -0.014 | -0.015 | -0.043 | 1 |

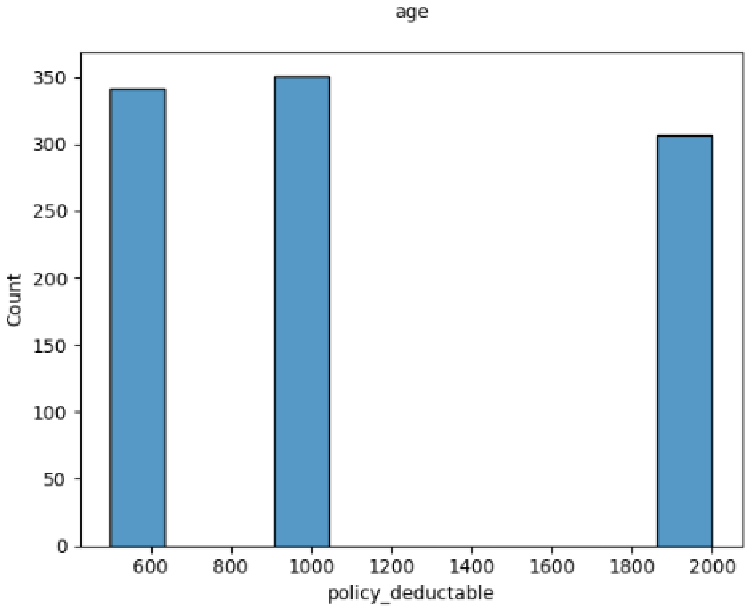
3. Majority of customer base is with Global Insure for 100 to 300 months.



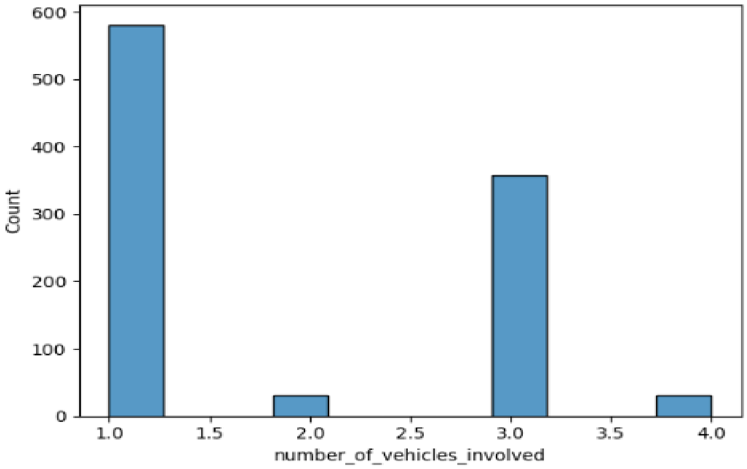
4. Majority of customers are between 30 to 45 years of age.



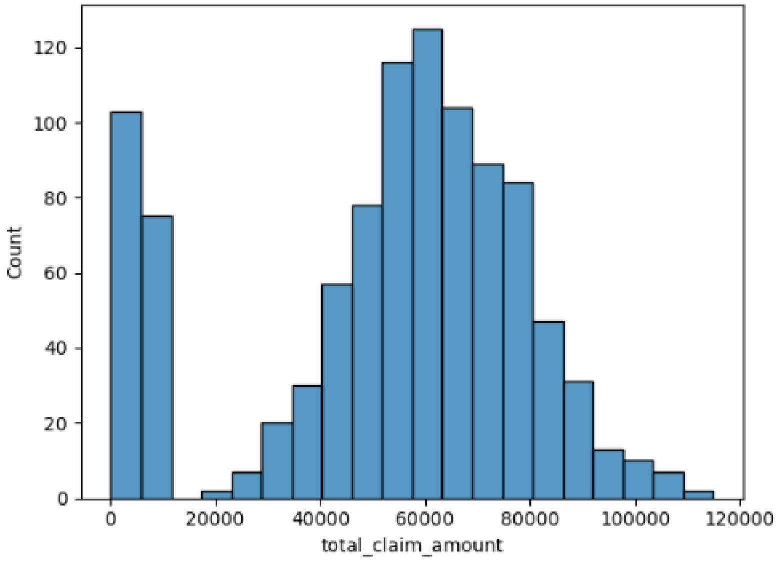
5. Most customers have annual premium between 800 and 1200.



6. The vast majority (672 cases) involved only 1 vehicle, indicating mostly single-vehicle incidents or claims.



7. Most claims are for claim amount of 50K to 65K.



**Predicting and Model Evaluation:**

- Logistic Regression Prediction: Using the cut-off of 0.3, we get a Train Recall = 84.62998102466793 % and a Test Recall = 100.0 %
- Random Forest Prediction: Using a cut-off of 0.3, we get a Train Recall = 99.05123339658444 % and Test Recall = 100.0 %.

**Conclusion:**

In this analysis, we compared the performance of two machine learning models—Logistic Regression and Random Forest—using recall as the evaluation metric, with a cut-off threshold of 0.3.

Logistic Regression demonstrated a strong performance with a train recall of 84.63%, but it achieved perfect recall of 100% on the test set. This suggests that while the model performs well on unseen data, there might be some overfitting on the training set, as it has a noticeable gap in recall performance between train and test datasets.

Random Forest, on the other hand, showed exceptional recall performance with a train recall of 99.05% and a perfect 100% test recall. The Random Forest model is able to generalize well on the test data, demonstrating both high accuracy and excellent recall performance.

Overall, both models performed well with high recall, particularly on the test set. However, the Random Forest model displayed more robust performance, with both high train and test recall, making it a potentially more reliable choice for this classification task.