FRAUDULENT CLAIM DETECTION

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.

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.

HOW CAN WE ANALYZE HISTORICAL CLAIM DATA TO DETECT PATTERNS THAT INDICATE FRAUDULENT CLAIMS?

- The steps include to analyze above details:

Data Preparation and Data Cleaning

Train Test Split

EDA

Feature Creation

Model Budling

WHICH FEATURES ARE MOST PREDICTIVE OF FRAUDULENT BEHAVIOUR?

auto_model

insured_hobbies

collison_type

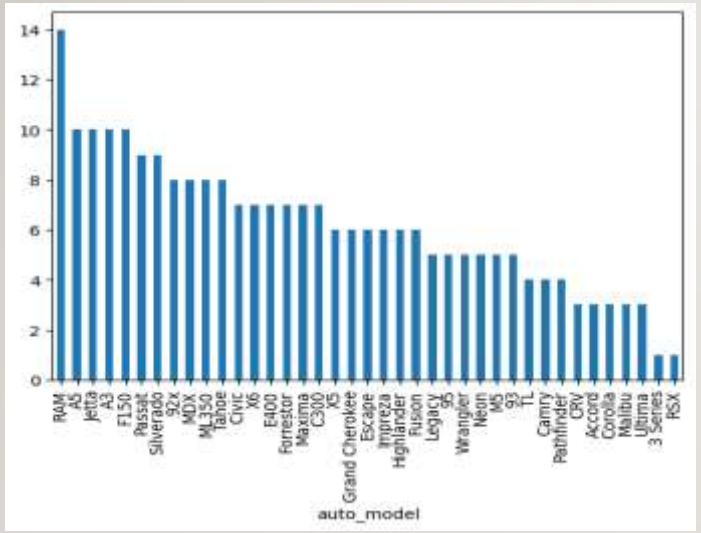
Incident_severity

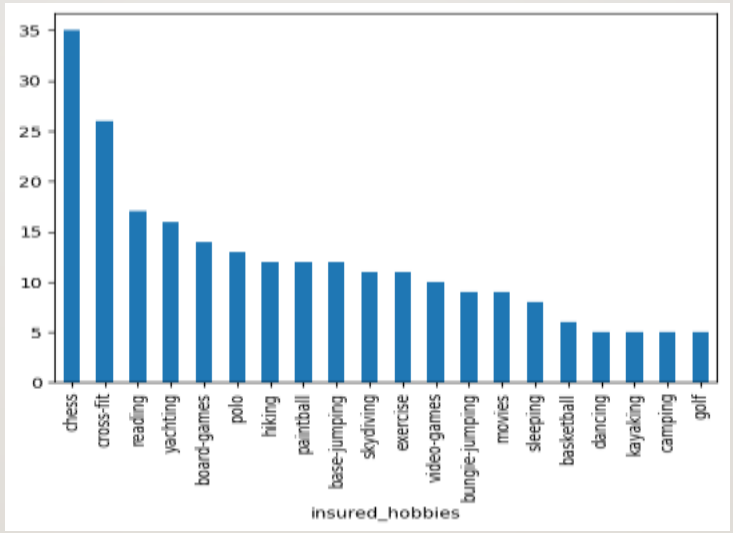
Incident_state

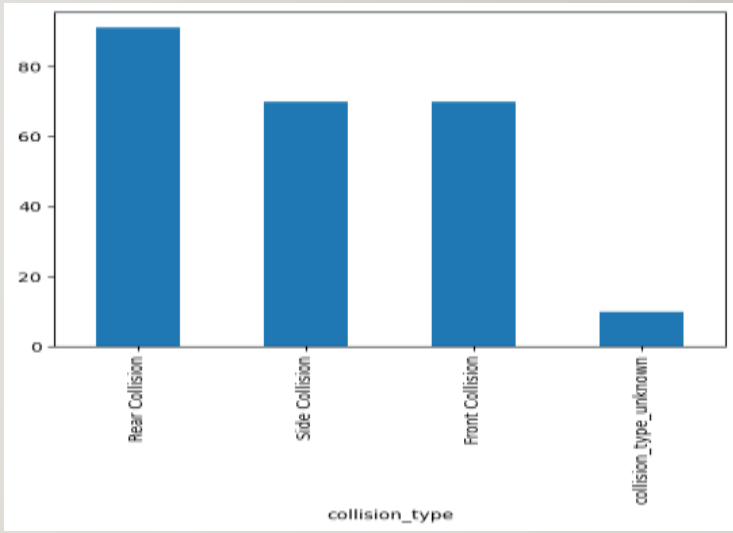
Insured_occupation

unmrella_limit

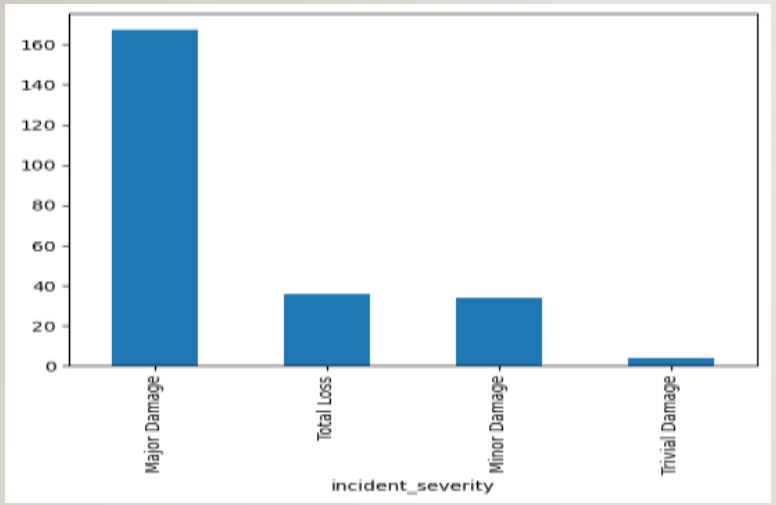
UNIVARIATE ANALYSIS OF FEW CATEGORICAL VARIABLES IN DATA

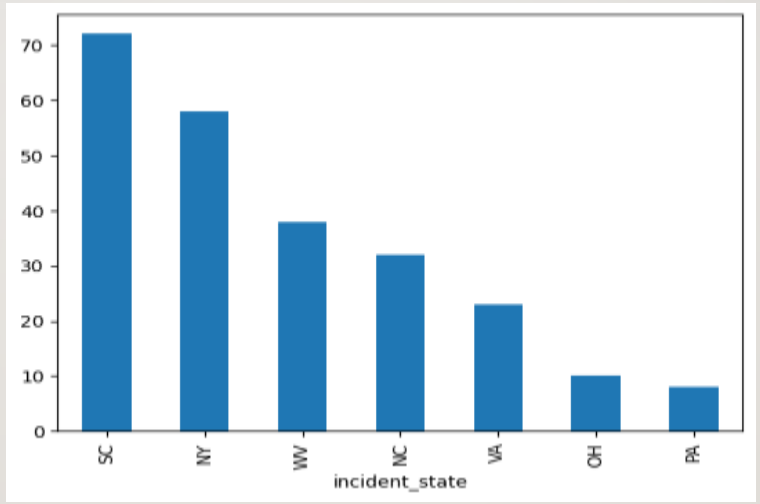


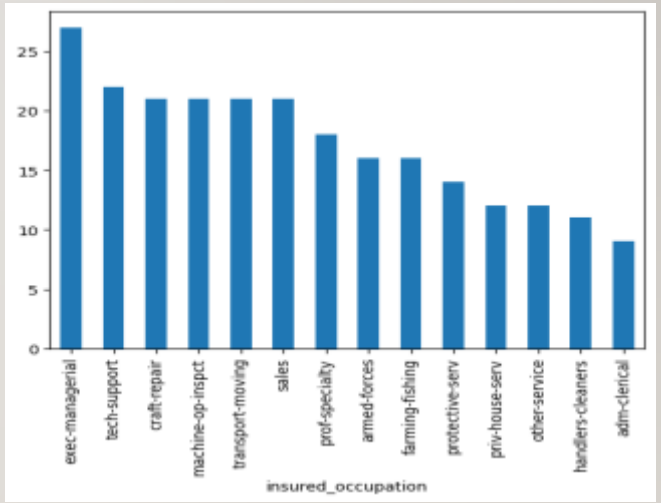




UNIVARIATE ANALYSIS OF FEW CATEGORICAL VARIABLES IN DATA







CAN WE PREDICT THE LIKELIHOOD OF FRAUD FOR AN INCOMING CLAIM, BASED ON PAST DATA?

- Yes. We have built the model using the below two model building process:
  - Logistic Regression
  - Random Forest Model

We have found that in our case logistic Regression model is showing better performance. For a new incoming claim we can follow the below process too predict likelihood of fraud:

Predicting Fraud Likelihood for Incoming Claims:

- **Feature Extraction:** When a new claim comes in, we extract the same set of features that were used to train the model.
- **Model Scoring:** The trained machine learning model then processes these features and outputs a **fraud risk score** or a **probability of fraud** for the new claim.
- **Thresholding:** A predefined threshold is used to classify the claim as potentially fraudulent or non-fraudulent based on its risk score. Claims exceeding the threshold are flagged for further investigation.

LOGISTIC REGRESSION VS RANDOM FOREST

- Note- For Random Forest sensitivity,specificity,precision.recall,fi-score can vary slightly for each run.

Logistic Regression	Random Forest
Accuracy_sore- 0.8278388278388278	Accuracy_sore- 0.8095238095238095
seinsitivity- 0.7638888888888888	seinsitivity- 0.6388888888888888
specificity- 0.8507462686567164	specificity- 0.8706467661691543
Precision- 0.6470588235294118	Precision- 0.6388888888888888
Recall- 0.7638888888888888	Recall- 0.6388888888888888
F1_Score- 0.7006369426751593	F1_Score- 0.6388888888888888

WHAT INSIGHTS CAN BE DRAWN FROM THE MODEL THAT CAN HELP IN IMPROVING THE FRAUD DETECTION PROCESS?

- Determining the features that are strongly corelated to determine fraudulent claims.
- Determining the features that are most important to determine fraudulent claims.
- Grouping the low frequency/low important categorical columns to improve the determination of fraudulent claims.

- In summary, the insights drawn from a fraud prediction model go beyond simply flagging suspicious claims. They provide a deeper understanding of the underlying patterns of fraud, enabling insurance companies to refine their detection processes, optimize resource allocation, adapt to evolving threats, and ultimately reduce fraudulent payouts more effectively.