REPORT

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.

Understanding the Rise of Insurance Fraud in Rural India

When we think about money fraud, we usually imagine smart people in big cities using computers and the internet to trick others.But when it comes to insurance fraud, the reality is quite different. Surprisingly, rural areas in India have seen a major spike in fraudulent claims over the past decade and it's becoming a serious problem.

Insurers have discovered well-organized groups operating in these regions almost like companies with trained people collecting information, forging documents, and filing fake claims. These fraud rings often target families of people who are critically ill or near death, making it easier to fake life or health claims. On top of that, the lack of strict checks when policies are issued makes it easier for such fraud to slip through the cracks.

To fight this, we need more than just basic checks. Insurance companies must start using **technology like AI and machine learning** to catch suspicious patterns early. At the same time, there's a need for **stricter punishments under the law**, and a **central database** of known fraudsters to prevent repeat offenses. Without strong action and better systems in place, rural insurance fraud may continue to grow causing huge losses and eroding trust in the system.

Objective:

- Move beyond traditional, rule-based fraud detection methods used in insurance.
- Build a **data-driven model** using historical claims data to detect fraud automatically.
- Use **machine learning algorithms** (like Random Forest and Logistic Regression) to classify claims as fraudulent or legitimate.
- Identify complex fraud patterns that manual checklists or fixed rules often miss.
- Enable **early fraud detection** to prevent financial losses and speed up decision-making.
- Support insurers in making smarter, more accurate claim approvals with less manual effort.

Libraries used:

1) warnings

2) numpy

3) pandas

4) seaborn

5) matplotlib.pyplot

6) sklearn (specifically sklearn.model_selection, sklearn.metrics, sklearn.ensemble)

7) imblearn.over_sampling

8) statsmodels.api

9) statsmodels.stats.outliers_influence

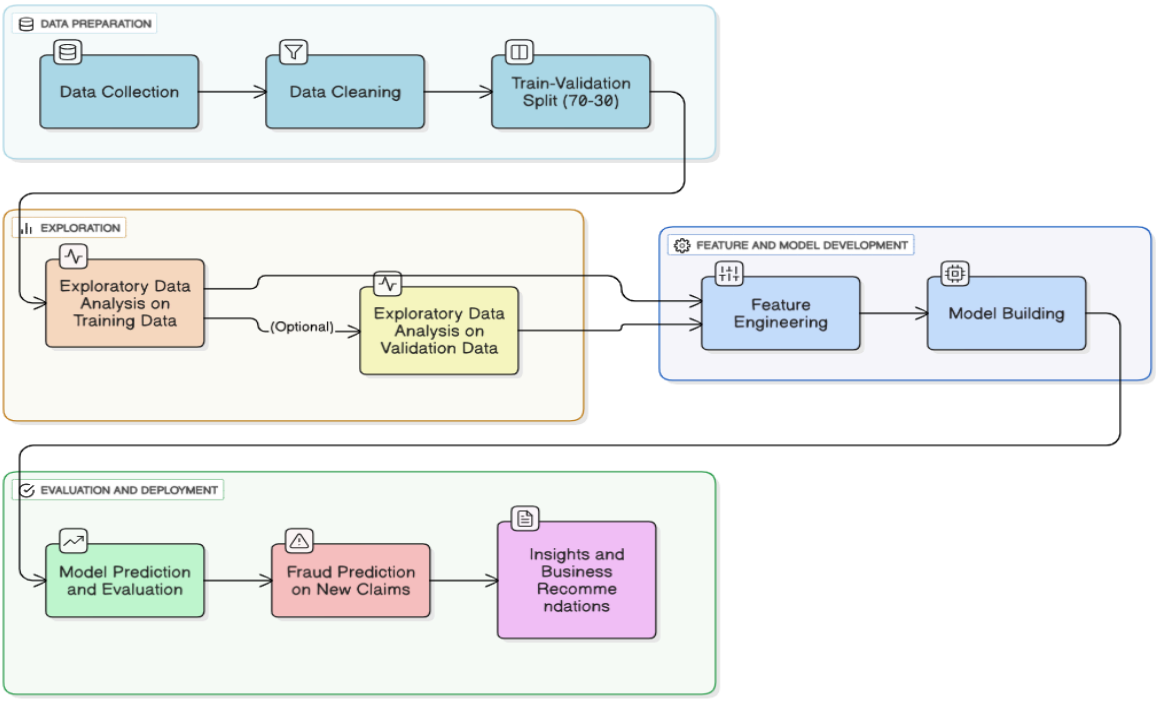
10) sklearn.preprocessing

Algorithms used:

Algorithms that are used:

1) **random forest**
2) **logistic regression**

DESIGN AND ROADMAP:



Data Description:

The data contains the following attributes:

Features:

Customer & Policy Details

- months_as_customer: Number of months the customer has been with the insurance company.
- age: Age of the insured person.
- policy_number: Unique identifier for each insurance policy.

- policy_bind_date: The date when the policy started.
- policy_state: State where the policy was issued.
- policy_csl: Combined Single Limit – coverage for bodily injury within total damages.
- policy_deductable: Amount paid by the policyholder before the insurer starts covering costs.
- policy_annual_premium: Yearly premium paid for the policy.
- umbrella_limit: Extra liability protection beyond basic policy limits.

Customer Profile

- insured_zip: Zip code of the insured person.
- insured_sex: Gender of the insured person.
- insured_education_level: Highest qualification of the policyholder.
- insured_occupation: Occupation/profession of the policyholder.
- insured_hobbies: Hobbies and activities of the insured person.
- insured_relationship: The relationship status or dependents listed under the policy.
- capital-gains: Profit earned by the customer from investments.
- capital-loss: Loss incurred by the customer from investments.

Incident & Claim Information

- incident_date: Date when the accident or event occurred.
- incident_type: Type of incident (e.g., theft, collision).
- collision_type: Nature of the vehicle collision.
- incident_severity: Level of damage or injury due to the incident.
- authorities_contacted: Authorities informed after the incident.
- incident_state, incident_city, incident_location: Where the incident occurred.
- incident_hour_of_the_day: Time of day when the incident took place.

- number_of_vehicles_involved: Number of vehicles in the incident.
- property_damage: Whether property was damaged.
- bodily_injuries: Number of bodily injuries.
- witnesses: Count of people who witnessed the event.
- police_report_available: Whether a police report was filed.

Claim Amounts

- total_claim_amount: Total amount claimed.
- injury_claim: Amount claimed for injuries.
- property_claim: Amount claimed for property damage.
- vehicle_claim: Amount claimed for vehicle repair/replacement.

Vehicle Information

- auto_make: Car manufacturer (e.g., Honda, Ford).
- auto_model: Specific model of the car.
- auto_year: Year of car manufacture.

Target Variable

- fraud_reported: Indicates if the claim is fraudulent (Y) or not (N).