Fraudulent Claim Detection Project Overview:

- **Objective:** To improve fraud detection at Global Insure by developing a data-driven model that classifies insurance claims as fraudulent or legitimate early in the approval process.
- **Problem:** Global Insure faces financial losses due to a significant percentage of fraudulent claims, often detected late. The current manual inspection is time-consuming and inefficient.
- **Business Goal:** Minimize financial losses and optimize the claims handling process.

Agenda:

1. Defining the Problem
2. Data Preparation
3. Data Cleaning
4. EDA (Exploratory Data Analysis)
5. Feature Engineering
6. Model Building
7. Predicting and Evaluation

1. Defining the Problem:

- **Company:** Global Insure
- **Task:** Build a model to classify insurance claims (fraudulent/legitimate) using historical data.
- **Data:** Claim amounts, customer profiles, claim types.

2. Data Preparation:

- **Data Dictionary:** Used to understand the dataset.
- **Dataset Size:** 1000 rows, 40 columns.
- **Data Types:** 21 categorical features, 19 numerical features.

Categorical Features (List):

- policy_bind_date

- policy_state
- policy_csl
- insured_sex
- insured_education_level
- insured_occupation
- insured_hobbies
- insured_relationship
- incident_date
- incident_type
- collision_type
- incident_severity
- authorities_contacted
- incident_state
- incident_city
- incident_location
- property_damage
- police_report_available
- auto_make
- auto_model
- fraud_reported

Numerical Features (List):

- months_as_customer
- age
- policy_number
- policy_deductable
- umbrella_limit
- insured_zip
- 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
- policy_annual_premium
- _c39

3. Data Cleaning:

- **Missing Data:**
    - 38/40 features: 0% missing data.
    - authorities_contacted: 9.1% missing values.
    - _c39: 100% missing values.
- **Data Quality:**
    - Ambiguous values ("?") in property_damage and police_report_available.
    - Low variance/_c39 column.
    - High cardinality in policy_number and incident_location.
- **Cleaning Steps:**
    - Remove rows with null values in authorities_contacted.
    - Drop column _c39.
    - Remove high-cardinality columns (policy_number, incident_location, policy_annual_premium, insured_zip).
    - Delete rows with negative values in umbrella_limit.
    - Replace "?" with 'UNKNOWN' in police_report_available, property_damage, and collision_type.
    - Convert policy_bind_date and incident_date to datetime.

4. EDA (Exploratory Data Analysis):

- **Analysis:** Distribution of columns, correlation matrices, feature imbalances, relationships with the target feature.
- **Key Numeric Insights:**
    - High Cardinality: policy_number, incident_location (100% unique values).
    - Skewed Claims Data: total_claim_amount, injury_claim, property_claim, vehicle_claim.
    - Negative values in umbrella_limit.
- **Categorical Features:** 21 features related to customer demographics, incidents, and vehicle/policy info.
    - Collision Type: Rear (292), Side (275), Front (254), Unknown (8).
    - Property Damage: Unknown (36.23%), No (33.48%), Yes (30.29%).
    - Police Report Availability: Unknown (34.69%), No (33.81%), Yes (31.50%).
- **Handling Data (EDA Reiterated):** Same steps as in "Data Cleaning" section.

5. Feature Engineering:

- **New Features:**
    - incident_month (from incident_date)
    - incident_day_of_week (from incident_date)
    - policy_age (difference between policy_bind_date and incident_date)
    - injury_claim_ratio, property_claim_ratio, vehicle_claim_ratio (ratio to total_claim_amount).
- **Handling Categorical Features:**
    - insured_occupation: Categories < 6% replaced with 'Other'.
    - insured_hobbies: Categories < 5% grouped into 'Other'.

6. Model Building:

- **Models:**
  - Logistic Regression
  - Random Forest
- **Logistic Regression:**
  - Feature Selection (RFECV)
  - Multicollinearity Assessment
  - Training & Evaluation
  - Optimal Cutoff
  - Final Prediction & Evaluation
- **Random Forest:**
  - Feature Importances
  - Model Evaluation
  - Cross-Validation
  - Hyperparameter Tuning (Grid Search)
  - Final Model & Evaluation

7. Model Evaluation:

- **Metrics:** Accuracy, Sensitivity, Specificity, Precision, F1-score.
- **Results Table:**

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score
Logistic Regression	0.73	0.75	0.72	0.49	0.59
Random Forest	0.83	0.79	0.84	0.64	0.71

Conclusion:

- Both Logistic Regression and Random Forest models were developed.
- Random Forest performed better (higher accuracy, F1-score).
- Random Forest Model Metrics: Accuracy (0.83), Sensitivity (0.79), Specificity (0.84), Precision (0.64), F1-score (0.71).

This is a full, detailed compilation of all the information included in the "Fraudulent Claim Detection" document.

Fraudulent Claim Detection

This document outlines the process and findings of a fraudulent claim detection project. Global Insure, a leading insurance company, processes thousands of claims annually. A significant percentage of these claims are fraudulent, resulting

in financial losses. This project aims to improve the fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process.

Agenda

1. Defining the Problem
2. Data Preparation
3. Data Cleaning
4. EDA (Exploratory Data Analysis)
5. Feature Engineering
6. Model Building
7. Predicting and Evaluation

Defining the Problem

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.

Problem Statement

Global Insure faces financial losses due to a significant percentage of fraudulent claims, which are often detected too late in the process. The current manual inspection process is time-consuming and inefficient.

Business Objective

The primary objective is to minimize financial losses and optimize the overall claims handling process by developing a data-driven model that can classify claims as fraudulent or legitimate early in the approval process.

Data Preparation

A provided data dictionary was used to understand the dataset. The dataset contains 1000 rows and 40 columns.

The dataset includes:

- 21 categorical features
- 19 numerical features

Categorical Features

List of categorical features:

- policy_bind_date
- policy_state
- policy_csl
- insured_sex
- insured_education_level
- insured_occupation
- insured_hobbies
- insured_relationship
- incident_date
- incident_type
- collision_type
- incident_severity
- authorities_contacted
- incident_state
- incident_city
- incident_location
- property_damage
- police_report_available
- auto_make
- auto_model
- fraud_reported

Numerical Features

List of numerical features:

- months_as_customer
- age
- policy_number

- policy_deductable
- umbrella_limit
- insured_zip
- 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
- policy_annual_premium
- _c39

Data Cleaning

The data cleaning process involved several steps:

- **Data Summary:** 38 out of 40 features have 0% missing data. authorities_contacted has 9.1% missing values. _c39 has 100% missing values.
- **Data Quality Observations:** Some columns contain missing or ambiguous values marked as "?", especially in property_damage and police_report_available. Several columns have low variance or limited usefulness (e.g., _c39). High cardinality in policy_number and incident_location.
- **Handling Data:**
    - Remove rows where authorities_contacted contains null values.
    - Drop column _c39.
    - Remove high-cardinality columns including policy_number, incident_location, policy_annual_premium, and insured_zip.
    - Delete rows where umbrella_limit contains negative values.
    - Replace ambiguous entries (?) with 'UNKNOWN' in police_report_available, property_damage, and collision_type.

- Convert data types of policy_bind_date and incident_date from object to datetime.

EDA (Exploratory Data Analysis)

EDA involved analyzing the distribution of numerical and categorical columns, correlation matrices, feature imbalances, and the relationship between categorical features and the target feature.

Key Numeric Insights

- High Cardinality: policy_number, incident_location have 100% unique values.
- Claims Data: total_claim_amount, injury_claim, property_claim, and vehicle_claim are highly skewed.
- Presence of negative values in umbrella_limit.

Categorical Features

There are 21 categorical features, including customer demographics, incident details, and vehicle and policy info.

Feature	Distribution Details
Collision Type	Rear: 292, Side: 275, Front: 254, Unknown: 8
Property Damage	Unknown (?): 36.23%, No: 33.48%, Yes: 30.29%
Police Report Availability	Unknown (?): 34.69%, No: 33.81%, Yes: 31.50%

Handling Data

1. Remove rows where authorities_contacted column contains null values.
2. Drop the column _c39.
3. Remove high-cardinality columns.
4. Delete rows where the umbrella_limit contains negative values.

5. Replace ambiguous entries (?) with 'UNKNOWN' in specific columns.
6. Convert data types of policy_bind_date and incident_date.

Feature Engineering

Feature engineering involved creating new features to improve model performance:

- **incident_month:** Extracts the month from incident_date.
- **incident_day_of_week:** Extracts the day of the week.
- **policy_age:** Calculates the number of days between policy_bind_date and incident_date.
- **injury_claim_ratio, property_claim_ratio, vehicle_claim_ratio:** Ratios of individual claim amounts to total_claim_amount.

Handling Categorical Features

- insured_occupation: Categories with less than 6% of total occurrences were replaced with 'Other'.
- insured_hobbies: Categories with less than 5% frequency were grouped into 'Other'.

Model Building

Two models were built:

1. Logistic Regression
2. Random Forest

Logistic Regression

- Feature Selection using RFECV
- Model Building and Multicollinearity Assessment
- Model Training and Evaluation
- Finding the Optimal Cutoff
- Final Prediction and Evaluation

Random Forest

- Get Feature Importances
- Model Evaluation
- Cross-Validation
- Hyperparameter Tuning using Grid Search
- Final Model and Evaluation

Model Evaluation

Both models were evaluated using various metrics, including accuracy, sensitivity, specificity, precision, and F1-score.

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score
Logistic Regression	0.73	0.75	0.72	0.49	0.59
Random Forest	0.83	0.79	0.84	0.64	0.71

Conclusion

How can we analyse historical claim data to detect patterns that indicate fraudulent claims?

- Exploratory Data Analysis (EDA): Initial patterns are identified by visualizing data. This includes:
    - Comparing the frequency of categorical feature values (e.g., incident_type, collision_type, incident_severity) between fraudulent and non-fraudulent claims. A higher proportion of fraud for a specific category (e.g., 'Major Damage' for incident_severity) indicates a pattern.
    - Examining distributions of numerical features (e.g., total_claim_amount, age, number_of_vehicles_involved, witnesses) for fraudulent versus non-fraudulent claims using box plots. Significant

differences in medians or ranges (e.g., higher claim amounts or fewer witnesses in fraudulent claims) highlight patterns.

- Feature Engineering: New features are created to uncover more complex patterns:
    - Temporal features like policy_age (time between policy start and incident). A pattern might be that newer policies are more frequently associated with fraud.
    - Ratio features like injury_claim_ratio (proportion of total claim from injury). An unusually high ratio for a minor incident type could be a pattern.
- Machine Learning Model Interpretation: Models learn and quantify these patterns:
    - Logistic Regression: Features selected by RFECV and those with significant p-values and influential coefficients in the model summary (e.g., a high positive coefficient for a dummy variable like incident_severity_Major Damage) are identified as components of fraudulent patterns.
    - Random Forest: The model provides feature importance scores. Features with high importance (e.g., total_claim_amount, incident_severity, policy_age) are key elements in the patterns the model uses to distinguish fraudulent claims.

Which features are most predictive of fraudulent behaviour?

Based on the analysis, the features most predictive of fraudulent behavior are identified by both Logistic Regression (via RFECV and statistical significance) and Random Forest (via feature importance scores).

Considering the Random Forest model generally performed better, its feature importances are particularly insightful. The key predictive features include:

- Incident-Specific Features:
    - incident_severity: Different levels (Major Damage, Minor Damage, Total Loss, Trivial Damage) are highly indicative.
    - witnesses: The number of witnesses.
    - incident_hour_of_the_day: The time of the incident.
    - number_of_vehicles_involved.
    - bodily_injuries.
- Claim Amount Features:
    - total_claim_amount: The overall amount claimed.
    - injury_claim, property_claim, vehicle_claim: The breakdown of the claim amount.
    - Engineered ratios like injury_claim_ratio, property_claim_ratio, vehicle_claim_ratio.
- Policy and Customer Features:
    - policy_age: An engineered feature representing the policy's age at the time of the incident.
    - months_as_customer: Duration the customer has been with the insurer.
    - age: Age of the insured.
    - policy_deductable.
    - umbrella_limit.

    - capital-gains and capital-loss.

- Vehicle Features:

    - auto_year: The manufacturing year of the vehicle.

The Logistic Regression model, after RFECV and VIF filtering, also highlighted many of these, such as months_as_customer, policy_deductable, umbrella_limit, capital-gains, various incident_severity levels, witnesses, and property_claim_ratio.

What insights can be drawn from the model that can help in improving the fraud detection process?

1. High-Impact Features: The model highlights strong indicators of fraud, such as incident_severity, total_claim_amount, policy_age (engineered), number of witnesses, and capital-gains/loss. This enables investigators to prioritize reviews based on the most critical factors.

2. High-Risk Scenarios: By examining certain claim types—like those involving 'Trivial Damage' or cases where no police were contacted—that frequently appear in fraudulent activity, the company can identify recurring fraud patterns and flag suspicious claims more quickly.

3. Implement Automated Prioritization: The model assigns each claim a fraud score. High-risk claims are escalated to investigators, while low-risk claims are processed faster. This improves efficiency and reduces manual workload.

Both the Logistic Regression and Random Forest models were developed to classify insurance claims as fraudulent or legitimate. The Random Forest model performed slightly better than the Logistic Regression model on the validation data, as indicated by the higher accuracy and F1-score. Both models show reasonable performance in classifying fraudulent claims. The Random Forest model achieved an accuracy of 0.83, a sensitivity of 0.79, a specificity of 0.84, a precision of 0.64, and an F1-score of 0.71 on the validation data, suggesting it is the better performing model.