Fraudulent Claim detection Model

By Abhijay Giri & Abdul Azeez Sheik

Objectives

The objective of this assignment is 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.

The key goals are:

● How can we analyse historical claim data to detect patterns that indicate fraudulent claims?
● Which features are most predictive of fraudulent behaviour?
● Can we predict the likelihood of fraud for an incoming claim, based on past data?
● What insights can be drawn from the model that can help in improving the fraud detection process?

Data Pipeline

1. Data Preparation
2. Data Cleaning
3. Train Validation Split 70-30
4. EDA on Training Data
5. EDA on Validation Data (optional)
6. Feature Engineering
7. Model Building
8. Predicting and Model Evaluation

Data Understanding

The dataset contains information on orders placed through Porter, with the following columns:

The insurance claims data has 40 Columns and 1000 Rows. Following data dictionary provides the description for each column present in dataset:

Column Name	Description
months_as_customer	Represents the duration in months that a customer has been associated with the insurance company.
age	Represents the age of the insured person.
policy_number	Represents a unique identifier for each insurance policy.

Column Name	**Description**
policy_bind_date	Represents the date when the insurance policy was initiated.
policy_state	Represents the state where the insurance policy is applicable.
policy_csl	Represents the combined single limit for the insurance policy.
policy_deductable	Represents the amount that the insured person needs to pay before the insurance coverage kicks in.
policy_annual_premium	Represents the yearly cost of the insurance policy.
umbrella_limit	Represents an additional layer of liability coverage provided beyond the limits of the primary insurance policy.
insured_zip	Represents the zip code of the insured person.
insured_sex	Represents the gender of the insured person.
insured_education_level	Represents the highest educational qualification of the insured person.
insured_occupation	Represents the profession or job of the insured person.
insured_hobbies	Represents the hobbies or leisure activities of the insured person.
insured_relationship	Represents the relationship of the insured person to the policyholder.
capital-gains	Represents the profit earned from the sale of assets such as stocks, bonds, or real estate.
capital-loss	Represents the loss incurred from the sale of assets such as stocks, bonds, or real estate.
incident_date	Represents the date when the incident or accident occurred.
incident_type	Represents the category or type of incident that led to the claim.
collision_type	Represents the type of collision that occurred in an accident.

Column Name	**Description**
incident_severity	Represents the extent of damage or injury caused by the incident.
authorities_contacted	Represents the authorities or agencies that were contacted after the incident.
incident_state	Represents the state where the incident occurred.
incident_city	Represents the city where the incident occurred.
incident_location	Represents the specific location or address where the incident occurred.
incident_hour_of_the_day	Represents the hour of the day when the incident occurred.
number_of_vehicles_involved	Represents the total number of vehicles involved in the incident.
property_damage	Represents whether there was any damage to property in the incident.
bodily_injuries	Represents the number of bodily injuries resulting from the incident.
witnesses	Represents the number of witnesses present at the scene of the incident.
police_report_available	Represents whether a police report is available for the incident.
total_claim_amount	Represents the total amount claimed by the insured person for the incident.
injury_claim	Represents the amount claimed for injuries sustained in the incident.
property_claim	Represents the amount claimed for property damage in the incident.
vehicle_claim	Represents the amount claimed for vehicle damage in the incident.
auto_make	Represents the manufacturer of the insured vehicle.
auto_model	Represents the specific model of the insured vehicle.

Column Name	Description
auto_year	Represents the year of manufacture of the insured vehicle.
fraud_reported	Represents whether the claim was reported as fraudulent or not.
_c39	Represents an unknown or unspecified variable.

Below are our findings on the same:

- **Business requirement**
- **Problem Statement**
- **Analysis on the dataset**
- **Model Comparison**
- **Addressing the problem statement**
- **Assumptions and Observations**

- **Business requirement**

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. 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.

- **Analysis on the dataset**

Our analysis followed the following step by step process:

1. Data Preparation: This involved getting a sense of the data and identifying the dependent (target) and independent variables
2. Data Cleaning: This involved checking for columns with null values, addressing columns with ? As values in them, dropping empty columns, tackling invalid values, fixing data types – numerical, categorical, date-time
3. Train-Test Split: This involved splitting the data into independent variables (X) and dependent / target variable (y) and conducting the train-test split
4. EDA on Training Data: We split the data into numerical and non-numerical – on the numerical we perform univariate analysis (through graphs), correlation analysis (to identify cases of multi-collinearity), class balance (between fraud and non-fraud cases), likelihood of fraud across categorical variables (to club a few and reduce features for later)
5. Feature Engineering: In feature engineering, we started with the resampling of data in training set in order to resolve the class imbalance (else the model will only predict

in one direction), created features like ageofvehicle, net capital gain / loss, and removed some redundant columns like age, individual claims (since we already had total claim amount), along with grouping a few instances within a categorical variable to reduce number of features while retaining max information, dummy variable creation for training (resampled) and test data, feature scaling using standard scaler for only numerical variables etc.

6. Model Building: Model building was carried out using both Logistics Regression and Random Forest including keeping important features and also doing hyperparameter tuning
7. Predicting and Model Evaluation: Predictions were made using the best version of the logistic regression model as well as the random forest model (next slide)

- **Model Comparison**

Model	**Data Type**	**Accuracy**	**Sensitivity (Recall)**	**Specificity**	**Precision**	**F1 Score**
Logistic Regression	**Training**	**90.4%**	**88.8%**	**92%**	**91.7%**	**90.3%**
Logistic Regression	**Validation**	**77.7%**	**79.8%**	**71.4%**	**79.8%**	**84.1%**
Random Forest	**Training**	**87.9%**	**84.5%**	**87.9%**	**87.5%**	**85.9%**
Random Forest	**Validation**	**79.1%**	**80.3%**	**75.7%**	**90.5%**	**85.1%**

While logistic regression model performed better on the training data, on the validation / test data, random forest gave better accuracy and other metrics as well – ensuring a more stable model

- **Addressing the problem statement (Evaluation and Conclusion)**

1. How can we analyse historical claim data to detect patterns that indicate fraudulent claims?
Ans: Using multiple available features we can put together a classification model (logistic regression / random forest etc.) which can help develop a clear picture of which features are important and which are not in determining fraudulent behaviour and that can be used to predict future instances of fraud – this requires us to make sure that the data is representative of both fraud and non fraud cases
2. Which features are most predictive of fraudulent behaviour?
Ans: Using different models (Log Reg vs Random Forest), we got different features as important (in order of importance):
Log Reg (based on value of coefficient) – Hobbies – Chess, Hobbies – Crossfit, Hobbies – Golf, Incident Severity (These did not make too much business sense except for the Incident Severity features)
Random Forest (based on feature importances) - Incident Severity, Hobbies – Chess, Total Claim Amount (made more business sense in Incident Severity and Total Claim Amount)
3. Can we predict the likelihood of fraud for an incoming claim, based on past data?
Ans: Ideally with a model like this we should get a better idea of what is the likelihood of fraud given a new claim which comes up – yes if we do the right kind of feature engineering and also have a base data with balanced values for fraud and non-fraud cases then we should be able to put together a better model as well
4. What insights can be drawn from the model that can help in improving the fraud detection process?
Ans: Insights that can be drawn is – we see hobbies playing a role in being important factors when it comes to fraud detection – while it may not hold too much business sense maybe clubbing hobbies by their nature (indoor / outdoor, risky / less risky etc.) may help put things into perspective and give better features. In case of random forest we had some business relevant features coming into picture like incident severity, total claim amount which can indicate fraudulent behaviour based on certain threshold values (above a certain value of total claim amount the probability of fraud may go up)

- **Assumptions and observations**

Across the document we have taken certain considerations as a group in order to make sure we are able to build the best model:

- Decided to drop the rows where authorities contacted is missing - in insurance industry it is important to know if authorities have been notified
- umbrella_limit has a negative value which does not make sense (row should be dropped)
- collision_type the rows with question mark may have to be dropped because if we do not have enough info about the type of collision then that is not very helpful (rows should be dropped)
- property_damage - the rows with question mark we may need to check if we use mode or if we drop certain columns. the problem is if we drop column then we lose the data for the rest - need to make sure if this is relevant or not

- Dropping the column c39 since it is empty
- Negative values in capital loss make sense so we will let it be
- Some columns have ? - which can be interpreted as unknown. let us keep it as is instead of my initial thought of dropping it
- There are a couple like incident hour of day and auto year which can be left as integers itself because making them date time might take away the essence of the datapoint
- We also created a feature called age of vehicle at time of incident for that i will first convert auto_year to a full date
- Dropping a few features which will not make sense in the modeling
- Dropping the following from X = policy number, policy bind date, insured zip, incident date, auto year, manufacture year
- In correlation matrix: Here we see two kinds of relations in terms of features - one is that total_claim_amount is highly correlated with other claim amounts and hence others can be scrapped later and we keep only total_claim_amount
- In correlation matrix: Second is age and months as customer are highly correlated which again makes sense - older a person more likely that they have spent more time as a policy holder
- The ratio of false to total in the main data set is roughly 73% and so it the case in the training data set - hence the training data set is representative of the main data
- But there is still a bias towards N as compared to Y and in later steps we will be using random oversampler as well
- There is some variation (in fraud %) (it is minimal in some but it exists) across all categorical variables and hence we will retain them
- The ones which have very little variation are - police report available, gender but not deciding to drop them even though they show very little variation
- In addition to age of vehicle at time of accident, we also created - One more combining capital gains and loss – net capital gain / loss
- Dropping columns like injury_claim, property_claim, vehicle_claim, age because they were all highly correlated with the variables like total claims amount and months as customer
- Did not create dummy variable for target (Dependent) variable - here the point to note is that if the target variable (fraud reported - is already in boolean format then there is not point of turning it into dummy variables