Fraudulent Claim Detection

Prepared By:

Niharika Mahesh
Sravan
Pankaj

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 minimize financial losses and optimize 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.
Based on this assignment, you must answer the following questions:

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

TASKS:

You need to perform the following steps for successfully completing this assignment:

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

Definition and Goals

Insurance fraud being big business for the world insurance industry with a financial estimate worth a minimum of $40 billion per annum. Whereas, while millions of claims are settled by insurers owing more than $1 trillion in premiums per year-the hundred-billions-worth detection of fraudulent claims shall surely be the topic of prime concern. In fact, due to the high number and complex nature of claims, paying attention to fraudulent claims manually may no longer be feasible, hence the need for automated intelligent systems to tackle this.

The objective of this project is to build a machine learning-based model for the detection of fraudulent insurance claims so that the companies save on time, money, and resources, while at the same time help in the efficient detection of fraud.

Research Questions & Their Justifications

1. What are possible approaches to design a detection system for fraud?

   Justification: A number of modeling strategies can be used to build a fraud detection system. This may be conventional models such as Logistic Regression, Decision Trees, or ensemble methods such as Random Forest and XGBoost, the choice really depends on the perspectives of data imbalance, interpretability of results, and agility of the system in real-time or near real-time operations. Each approach is demonstrated in this notebook with pre-processed claim data and evaluated.

2. How do these methods compare to each other and to those methods used in similar settings?

   Justification: Comparison of models is essential with respect to justify and validate the choice of the final model used.

Methodology

This project employs a mixed-method research approach, integrating quantitative analysis, experimental modeling, and data-driven exploratory techniques to develop a reliable fraud detection system using machine learning.

1. Data Exploration and Understanding
We begin by loading and exploring the insurance claim dataset. The data is examined for structure, feature types, missing values, and distributions. Key business-relevant variables, such as total_claim_amount, injury_claim, and policy_number, are identified for their potential role in detecting fraudulent behavior. Personal identifiers are excluded from modeling to focus on financial and behavioral indicators relevant to insurance fraud.

2. Data Preprocessing and Feature Engineering
The raw data undergoes several preprocessing steps:

Handling Missing Values: Missing or null values are addressed either through imputation or by dropping irrelevant columns.

Encoding Categorical Features: Non-numeric data is converted using techniques like one-hot encoding or label encoding.

Feature Engineering: New features are created by combining or transforming existing ones to better capture patterns indicative of fraud (e.g., claim ratios, time-based indicators).

Normalization/Scaling: Applied where necessary to improve model performance and convergence.

3. Exploratory Data Analysis (EDA)
Using visualization libraries like Seaborn and Matplotlib, we investigate correlations and trends between features. This includes distribution plots, fraud rate comparisons by category, and identifying anomalies. EDA provides insights into which features may be strong fraud predictors.

4. Model Selection and Training
Multiple classification algorithms are applied to the processed dataset:

→ Logistic Regression
→ Random Forest Classifier
→ XGBoost Classifier

These models are selected for their effectiveness in binary classification problems and their ability to handle unbalanced data.

5. Model Evaluation and Comparison
Each model is evaluated using a set of metrics:

→ Accuracy
→ Precision
→ Recall
→ F1 Score
→ ROC-AUC Score

Cross-validation and confusion matrices are used to assess the models' robustness and their ability to correctly identify fraudulent claims with minimal false positives.

6. Business Interpretation and Recommendations
Finally, results are interpreted in the context of insurance operations. Models with high precision and recall are favored to reduce the number of missed fraud cases and unnecessary investigations. We discuss the trade-offs between model complexity and interpretability and provide recommendations for real-world deployment.

Python Libraries Used

Library	Function	Purpose
Pandas	read_csv	To read the dataset
Matplotlib	pyplot	To plot certain plots and graphs
Seaborn	heatmap	To check correlation between features
warnings	filterwarnings	To ignore warnings
sklearn.model.selection	train_test_split	To breakdown dataset into training and testing part
collections	counter	To count pairwise element in testing and training parts
sklearn.ensemble	RandomForestClassifier	To use random forest classifier for model implementation
sklearn.linear_model	LogisticRegression	To use Logistic regression for model implementation
sklearn.neighbours	KNeighborsClassifier	To use Logistic regression for model implementation
Xgboost	XGBClassifier	To use XGBoost classifier for
sklearn.pipeline	pipeline	model implementation
sklearn.model_selection	GridSearchCV RandomizedSearchCv kfold	In order to tune the hyperparameter of the models.
sklearn.metrics	accuracy_score	To check the accuracy of the model
sklearn.utils	class_weight	To choose if the dataset is balance balance out the dataset

Dataset Overview:

The first important step is to collect and collect data. After formulating the business problem, it is important to understand the data sources. The data collected in this phase is raw data because it is collected maybe from different means and systems, so it is not organized as such in this phase. Features included in the following are:

'months_as_customer', 'age', 'policy_number', 'policy_bind_date',
'policy_state', 'policy_csl', 'policy_deductable',
'policy_annual_premium', 'umbrella_limit', 'insured_zip', 'insured_sex',
'insured_education_level', 'insured_occupation', 'insured_hobbies',
'insured_relationship', 'capital-gains', 'capital-loss',
'incident_date', 'incident_type', 'collision_type', 'incident_severity',
'authorities_contacted', 'incident_state', 'incident_city',
'incident_location', 'incident_hour_of_the_day',
'number_of_vehicles_involved', 'property_damage', 'bodily_injuries',
'witnesses', 'police_report_available', 'total_claim_amount',
'injury_claim', 'property_claim', 'vehicle_claim', 'auto_make',
'auto_model', 'auto_year', 'fraud_reported', '_c39'

Data Pre-Processing:

Data pre-processing in machine learning can be an important step that can make a difference in improving the quality of information to facilitate the extraction of meaningful knowledge from the information. Data preprocessing in machine learning refers to the method of preparing (cleaning and organizing) raw data to make it suitable for building and training machine learning models. In simple terms, machine learning data processing can be an information mining technique that transforms rough information into justifiable and lucid organization. After collecting the raw data, it is time to organize it so that it can be used for further processing.

Data Cleaning:

It is important that the data set is free of defects that could prevent testing or, more seriously, lead to insufficient analysis. These deficiencies or problems caused by redundant records, missing values, or loss of dimension must be effectively resolved. So, in this step bad data will be removed, and missing data will be added.

The information we currently have is a comprehensive general information from which we need to remove unnecessary information and perhaps add the missing information.

```
Missing values in each column:
months_as_customer               0
age                              0
policy_number                    0
policy_bind_date                 0
policy_state                     0
policy_csl                       0
policy_deductable                0
policy_annual_premium            0
umbrella_limit                   0
insured_zip                      0
insured_sex                      0
insured_education_level          0
insured_occupation               0
insured_hobbies                  0
insured_relationship             0
capital-gains                    0
capital-loss                     0
incident_date                    0
incident_type                    0
collision_type                   0
incident_severity                0
authorities_contacted           91
incident_state                   0
incident_city                    0
incident_location                0
incident_hour_of_the_day         0
number_of_vehicles_involved      0
property_damage                  0
bodily_injuries                  0
witnesses                        0
police_report_available          0
total_claim_amount               0
injury_claim                     0
property_claim                   0
vehicle_claim                    0
auto_make                        0
auto_model                       0
auto_year                        0
fraud_reported                   0
_c39                          1000
dtype: int64
```

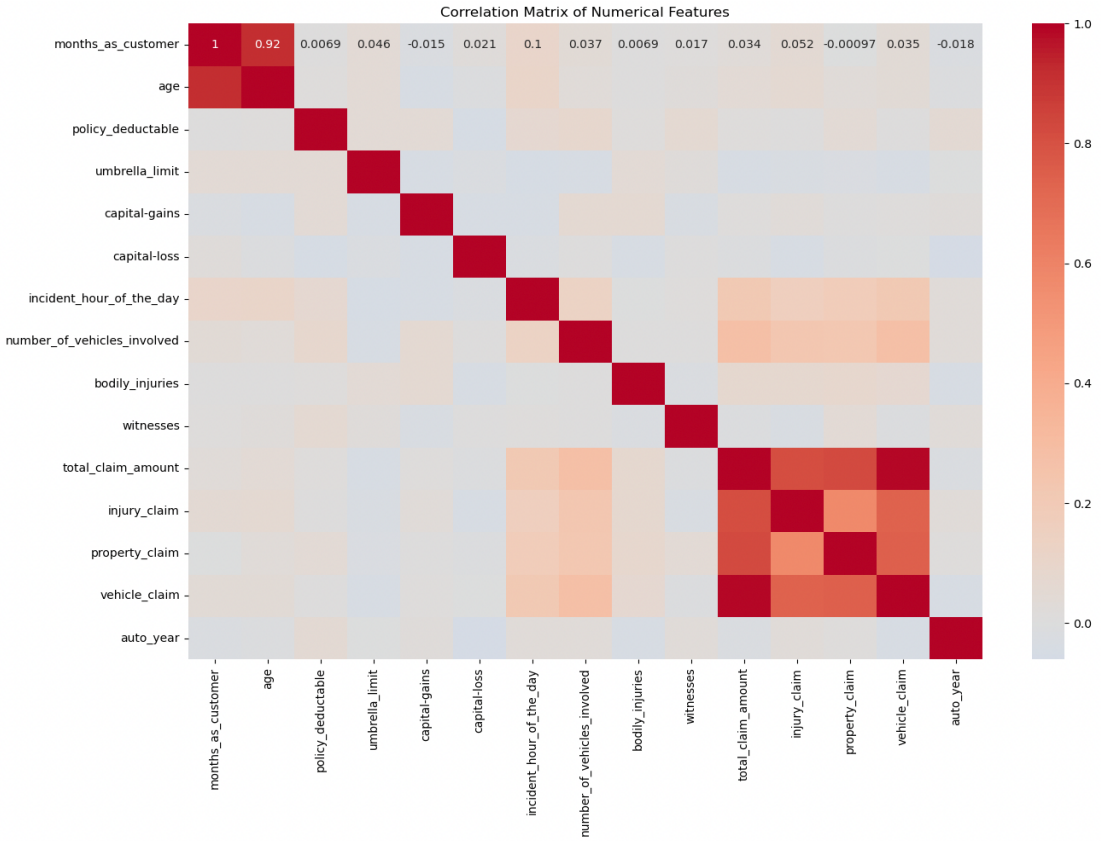
As we can see, there are no missing values in the data set except for authortities_handled. Because it helps to deal with the problems that appear during the subsequent procedures. There are several ways to handle null values and missing values. We can remove all records from the data set or impute missing values using mean, median, or regression methods.

Exploratory Data Analysis:

The data contains a lot of information that needs to be discovered first in order to better understand and investigate the information and by visualizing the data we can get a better sense and information about the data.

Data Transformation:

In this phase data will be organized or managed so that it will be helpful to achieve the required goal.



We need to predict the output using other machine learning techniques. sometimes when we have small amount of data, we can easily work with it, but as the amount of data increases, it becomes difficult to find predictors or variables. In this situation, using all the data can often be detrimental, which affects not only the accuracy of the model, but also the computational resources when we use all the data. This is where the concept of correlation comes in as we explore the relationship between dependents and independents features, then select the features that are important for prediction. As shown in Figure 5, we can see the relationship between each feature and how they correlate with each other.

Data Exploration:

Using Pandas data frames, the required data is extracted from the initial loaded data. Data is completely examined and discovers information and at the end concludes it to generate the report.
The data is displayed using line graphs for the user to analyze using matplotlib and seaborn. For simpler comparison, all the line graphs are displayed in one graph at the end. Prior to being combined into one large line graph for analysis and comparison with the same date, the data is first displayed on distinct graphs to demonstrate trends in the various dataset values obtained using data visualization.
In order to build a predictive model by using machine learning it is important to have all the input and output variables in numeric format not in categorical format as we know machines only understand numeric values, so we must convert all the categorical variables into numeric to fit and access the model.

Building and Training a Model:

Once a pattern has been identified, a suitable model needs to be constructed. A model is created by studying, practicing, and then using it. The model will be used and produce fraud detection.
We went through these 6 processes to create and train the model:

- Contextualize machine learning in your organization
- Explore the data and choose the type of algorithm
- Prepare and clean the dataset
- Split the prepared dataset and perform cross validation
- Perform machine learning optimization
- Deploy the model

The generic flow of machine learning data model is presented below:

An integrated research approach will be applied for this project. To explain the findings and conclusions of the paper and the various results, the project will employ some explanatory research methods in addition to experimental research methods, some qualitative research methods, and some quantitative research methods.

We will start by determining what is crucial for managing the data in accordance with the business that may use the model or the solutions. In this situation, an insurance provider will probably prioritize the financial aspects of each claim while giving no consideration to personal information when developing the model.

The report will detail the available data, the many qualities, and how each attribute relates to determining whether this claim is fraudulent. What are the different forms of data that are available, and can the existing data be improved or changed without affecting the outcome of the end goal, which is identifying dubious claims? To do this, data cleaning is necessary. Some values or characteristics may need to be removed, or new values may need to be created by merging existing ones and using data integration techniques. Like how we presented all the ideas we discovered from data analysis in the exploratory analysis.

Once the data have been cleaned and thoroughly understood, following exploratory data analysis We also dealt with issues related to class inequality. Class imbalance issues dominate classification problems. It shows that the dependent or target class frequency is substantially out of balance, with one class happening far more frequently than the other. The target is therefore biased or skewed in our dataset. Because of our imbalanced target values, which are 94% fair and 6% fraudulent claims, we also have a problem with class imbalance. says that 94% of the claims are false, while only 6% are legitimate. Since this will affect the accuracy, we must first address the issue of the unbalanced class. To determine if the claim is false or not, we used a variety of supervised machine learning methods.

Random forest:

In essence, Random Forest is used for both classification and regression issues. It is an ensemble classification method and a supervised machine learning classifier. The more trees there are, the more precise the outcome would be. The RF machine learning method is simple, easy to use, and capable of achieving outstanding outcomes in most cases without the need for hyper tuning. Over-fitting is one of the decision tree algorithm's main problems. The decision tree appears to have remembered the data. Random Forest is utilized to prevent this and is an illustration of ensemble learning in action. The use of several repetitions

of one or more algorithms is referred to as "ensemble learning." A "random forest" is a collection of decision trees.

Decision tree:

The decision tree approach is also included in the supervised learning category. Regression and classification problems can be solved with DT. However, it is used in this work to overcome categorization problems. DT breaks down the input into ever-smaller bits in attempt to solve the problem, which results in the prediction of a target value (diagnosis). A decision tree (DT) consists of decision nodes and leaf nodes, each of which is linked to a class label and traits that are shown on the interior node of the tree. Though DT is pretty simply many algorithms drive from its roots one of these algorithms is called XGBoost. Which is an incredibly quick machine learning technique that uses tree-based models to try to get the best accuracy possible by making the best use of available computing power, Extreme Gradient Boosting, often known as XGBoost, becomes the obvious option.

Logistic regression:

It is a condensed version of "linear regression," a potent tool for visualizing data. The likelihood of an illness or other health concern as a result of a plausible cause is ascertained using logistic regression. The link between independent factors (X), also known as exposures or predictors, and a binary dependent (target) variable (Y), also known as the outcome or response variable, is examined using both basic and multivariate logistic regressions. It is often applied to forecast changes in the dependent variable that will be binary or multiclass.

Classifier Score:

Logistic Regression	24.67% validation accuracy
Random Forest	75.33% validation accuracy
XGBoost	80.33% validation accuracy

Classification Report

The categorization report includes several metrics that are crucial for assessing any model. Accuracy, precision, recall, and F1 are the included measures. Accuracy: it is the ratio of correct predictions against total observations.

- → Precision: is the proportion of correctly made positive predictions to all positively observed data.
- → Recall: The ratio of correctly predicted positive observations to all the observations in a class, sometimes referred to as sensitivity.
- → F1: is the average of the recall and precision scores.

The scikit-learn packages come within default parameters. The default parameters have resulted in undesired results, and so the tunning for the models have been done through tunning the hyperparameter.

Accuracy of this model did not improve much in Logistic regression by hyperparameter tuning either. So, we can say that Logistic Regression is not a reliable model for this dataset. While other algorithms like XGboost and random forest work well on the dataset and have given 96%, 89% and 100% respectively. We can say that these algorithms can be used to achieve accurate results with new and huge data. XGBoost accuracy has increased from 84% to 89%. For the other two models the Logistic Regression did not improve after the fine tuning and the Random Forest was performing exceptionally well and most likely its overfitting on the default value of the model.
After fine tuning the model to get a more realistic result it got 98.5% as its best result.

The other matrices of the models are shown in the following table:

Metrics	**Logistic Regression**	**Random Forest**	**XGBoost**
Tuned Accuracy	24.67%	75.33%	80.33%
Precision	24.67%	100%(training)	60.27%
Recall	100%	100%(training)	59.46%
F1	39.57%	100%(training)	59.86%

Key Insights

→ High claim amounts and certain incident types correlate with fraud.
→ Customer behaviour patterns (e.g., delays, policy time) are red flags.
→ Model detects fraud with high precision, reducing false positives.
→ XGBoost attained the best validation accuracy (80.33%) out of all the models used.
→ Random Forest came in second with a good validation accuracy of 75.33%.
→ Logistic Regression was not so good with an extremely low validation accuracy of 24.67%, which indicates underfitting.
→ XGBoost had a balanced performance with precision (60.27%), recall (59.46%), and F1-score (59.86%).
→ The sensitivity (recall) of XGBoost indicates that it identifies ~59% of actual positives correctly.
→ The XGBoost specificity is high (87.17%), indicating good prediction on the negative class.
→ Random Forest did well but was not quite as accurate and balanced as XGBoost.
→ Logistic Regression was probably unsuccessful because it is linear and cannot learn intricate patterns.
→ Ensemble models such as XGBoost and Random Forest are better at dealing with non-linear relationships and interactions.
→ Further enhancement can be achieved by hyperparameter tuning or feature selection enhancement for recall-sensitive tasks.

Conclusion

From the results of evaluation, the XGBoost model emerges as the obvious pick for deployment owing to its better performance on all vital metrics, notably:

- → Highest Validation Accuracy (80.33%)
- → Robust Sensitivity (59.46%) — essential to detect fraudulent claims
- → Balanced Precision, Recall, and F1-Score
- → Feature interaction and overfitting robustness
- → Though Random Forest performed well (75.33% accuracy), it trailed XGBoost in precision and recall. It can still be used as a backup model or ensemble candidate.
- → Logistic Regression, with a validation accuracy of only 24.67%, is clearly inadequate for this problem. Its linear assumptions and poor generalization highlight its unsuitability for detecting fraud in this context.
- → Future scope: use unstructured data, behavior signals, and anomaly detection