Interim Progress Report (IPR)

# Introduction and overview

Detection of fraud affects a wide range of businesses, particularly financial sector, insurance, government organisations, and law enforcement. Fraudulent efforts have grown considerably in recent years, emphasising the importance of fraud detection. Despite the efforts of several institutions, fraud results in the loss of enormous sums of money each year. So the amount of fraudulent operations is so tiny, detecting these scams is challenging.

In the insurance industry, around 10% of claims contain various types of dishonesty, resulting in at least 25% of insurance expenditures. Fraud may take many forms, from inflated losses to unintentional spending. Because there are so many distinct types of fraud, it's more hard to spot them.

Online wallets in varying forms are on the rise all around the entire globe. The amount of transactions handled by payment providers is rapidly increasing. In 2021, PayPal, for example, took $579 billion in total payments. Along with this change, there has been a significant surge in securities fraud in various banking systems.

The role of cyberspace & cyber-crime teams is to prevent unwanted online financial fraud. Numerous banks and investment organizations have specialized teams of hundreds of analysts working on automated systems to monitor transaction information via their products and flag those that are possibly fraudulent. As a result, to be effectively and efficiently equipped to handle computer hackers cases, it is critical to investigate how to face the issues of identifying fraudulent required fields in enormous volumes of data.

**1.2 Aim and Objective**

The objective is to develop a model that can accurately predict which payments are likely to be fraudulent. The standard method of detecting fraud is to create algorithms based on fraud signs. A fraud judgment might be made in two ways based on these heuristics.

In other instances, guidelines would be defined to determine if the matter needed to be investigated.

A checklist with ratings for the different fraud indications might be created in other circumstances. The sum of these ratings and the claim's worth will decide if the case has to be investigated.

This research involved a few months of work in creating a framework for detecting fraud in financial transactions. We desire that the research's conclusion make it easier to analyze & notice fraudulent transactions.

The project's three essential aims are as follows:

* To research the literature on financial fraud detection & better comprehend the many facets of the issue.
* Employing supervised ML procedures, tackle the concern of financial fraud detection using a publicly available selection dataset.
* To evaluate several categorization approaches to resolve which is most appropriate for this application.

The eventual aim is to develop a system and programmes that include the analytics and machine learning principles covered in the curriculum. The quality of the categorization findings and the scope of the analysis are critical to the research's triumph. We anticipate that the latest research will offer a baseline for future research & development in this area and a knowledge foundation for students interested in learning about the complexities of fraud prevention.

**1.3 Problems**

The most serious problem that insurance providers confront is fraud, which results in massive losses for them that are often irreversible. Because combating fraud cases, particularly in insurance firms, is a difficult process, the key objective is to avoid fraudulent acts at all costs. According to reports, between 22% and 34% of auto insurance claims are believed of being fraudulent, although only 3% of situations are punished. The first step in minimizing fraudulent situations is to discover them, which is complex and just not "expense" because extensive and time-consuming examinations may irritate genuine clients (Gupta, 2022).

Increasing investigative expenses also make it more difficult to discover fraud. As a result, businesses fail to conduct necessary investigations, resulting in a slew of potential hazards. Because manual fraud detection is costly and inefficient, we now need to examine the fraud before approving the claim. Several artificial intelligence & ML approaches are effective in identifying fraud (Hanafy, 2022).

This project aims to develop a model that can predict insurance fraud. The difficulty with ML fraud detection would be that scams are significantly less prevalent than legitimate insurance payments.

Given the variety of fraud types and the low number of confirmed frauds in regular sampling, detecting financial fraud is difficult. When developing detection algorithms, the expense of false warnings must be weighed against the cost of loss avoidance. Machine learning approaches improve forecast accuracy, allowing loss controllers to cover more territory with fewer false positives.

Insurance fraud refers to various unethical behaviors that a person may engage in to obtain a favorable outcome from an insurance company. This might include arranging the event, exaggerating the circumstances, including the notable characters and the incident's cause, and lastly, exaggerating the magnitude of the harm.

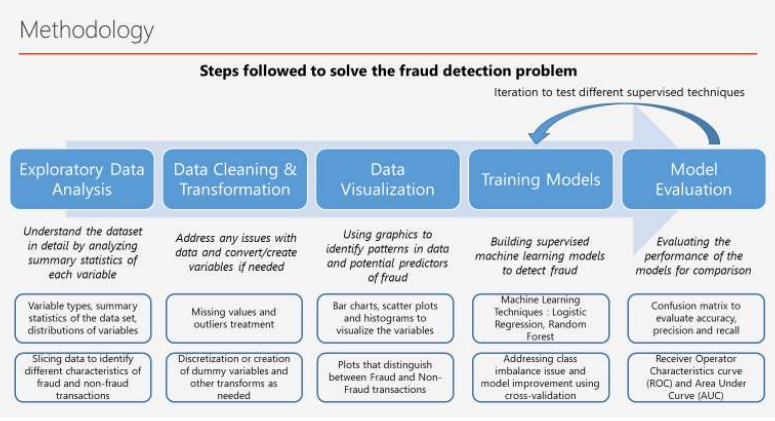
**1.4 Research Methodology**

In this study, the conventional ML method was used. The tagged class variable in the discovered dataset was utilized as the prediction variable in machine learning models.

* We employed exploratory analysis to examine the data set and identify potential fraud predictors.
* We witnessed the distinction between fraud and non-fraud transactions using several visualization tools.
* We used two supervised ML approaches – Logistic Regression and Random Forest – to address the fraud detection problem.
* To remedy the dataset's class imbalance, we also explored under-sampling.
* The models were constructed via cross-validation to minimize overfitting and provide consistent performance.
* To compare the performance of the models, performance metrics such as the Confusion Matrix and Area Under Curve (AUC) were utilized.

Jupyter notebook was used to conduct this analysis, which was done in Python. The ML models were run using built-in libraries and techniques. Functions were developed when needed to make specific analyses or visualizations easier.

The figure below depicts the whole method that was followed throughout the project.



**Figure 1 Methodology structure**

# Background research

Fraud detection is an issue that affects numerous businesses, including banking and finance, insurance, government organizations, and law enforcement. Fraudulent efforts have grown considerably in recent years, emphasizing the importance of fraud detection, regardless of various actions. Annually, substantial quantities of money are lost due to fraud in financial institutions. Detection of these forgeries. It is challenging because the number of fraudulent actions is relatively low.

In this [1], authors offers a blockchain-based architecture for a level of readiness and data sharing across the insurance network's numerous interacting actors, Distributed ledgers peer-to-peer interface that allows for validating treatment claims in a secure, immutable, and transparent manner. Additionally, discuss how technology and intelligent models might be combined to better business operations. It will demonstrate how these capabilities may be used to construct a system that eliminates fraud in automotive, healthcare, including life insurance payments, among several other areas. There study will examine the many forms of fraud prevention and detection within insurance claim systems and how they are classified using various machine learning algorithms. In addition, it discusses the future of Fraud Detection inside the Insurance Payout System.

In this [2], authors categorize the sequence of insurance claims, we propose two versions of LSTM-based ML algorithms. Autoencoders can produce feature significance that provides knowledge into the models' predictions whenever models are put into reality. This method is based on the notion that data outliers are fake. The procedures were developed and tested using a dataset we created using data from just a Swedish insurance firm, with the few flagged frauds used exclusively for analysis and testing. The experiments demonstrate legislature capability, and additional research reveals that the combined effect of autoencoders and LSTMs is efficient yet performs similarly to the baselines used. This research serves as a starting point for practitioners interested in learning essential areas of anomaly detection in fraud detection.

In this [3], authors employing machine learning techniques aims at detecting auto vehicle fraud. In addition, the accuracy will be examined using a confusion matrix. It could aid in the calculation of specificity, recall, & accuracy.

In this [4] , authors study describes a detecting fraud approach that uses data to forecast & evaluate fraud tendencies. They use the Nave Bayesian Identification & Judgement Tree-Based techniques to create classifiers. The method is described briefly and how it may be used to identify fraud. Both strategies make use of the same data. The classifier predictions are analyzed and interpreted. Bayesian Nave Visualization, Decision Tree Visualization, & Rule-Based Categorization are all used to assist prediction accuracy. They test strategies for detecting fraud in the automotive insurer.

In this [5] author Illustrate that a Laplacian matrix's primary non-principal eigenvector is related to a bi-class classification strength metric that can be used to rank anomalies. They proposed SRA delivers an aberration ranking, whether about the class labels or concentrating on two primary patterns by presently only the first non-principal eigenvalues of the Laplacian matrix. The ranking connection can be chosen established on whether the smaller class's cardinality (positive or negative) is sufficiently big. They show that our suggested SRA outperforms unusual conventional case fraud detection approaches using a vehicle insurance claim data set and disregarding labels while producing ranking. Eventually, they offer that, while the proposed SRA performs well for a few resemblance metrics in the car insurance industry.

In this [6], authors focuses on determining the efficacy and verifiability of the most well-known ML algorithms for fraudulent prediction. An anonymously insurance company's automotive data claims were used to test the supervised approach. We want to suggest a method for improving the validity of AI  outcomes. The research concluded that Random-forests outperforms all other techniques tested.

In this [7], authors This research aims to offer a unique DL approach that uses unstructured variable significance to acquire pragmatic insights into the behaviour of an insured individual. It provides the framework for understanding how minimal effort may be used to get insights into an authorized person's dishonest conduct. They present a new latent inconsistent design using two important unsupervised DL  algorithms, notably fully convolutional or the nonlinear autoencoder, after a preliminary assessment of the limits of conventional fraud investigative techniques. The characteristics of each model are addressed to help the reader understand how models may be customized for fraud protection and how the findings should be interpreted. Achievement assessments are done qualitatively and quantitatively, with a more significant focus on critical review. Various measures are employed to evaluate better the area of knowledge of the fraud detection setup.

# Ethical , legal professional and social issues

## Ethical issues:

An ethical fraud analyst will conduct a comprehensive inquiry. Lawyers will never suspect an insurance coverage of fraud unless he/she first obtains enough facts to reject a civil action or compel a prosecutor to pursue a criminal prosecution or both.

## Legal issues:

Both criminal & civil laws apply to fraud. The most evident distinction is who files the legal lawsuit. Offenses can only be brought by government prosecutors, although a victim of fraud can initiate a civil case. A people who commit fraud may be legally punished as well as civilly sued.

Fraud must be shown "further than a reasonable doubt" in a criminal situation. If guilty of the felony, a person may face jail time, probation, and fines.

## Professional issues:

Insurance fraud detection has proven to be difficult, owing to the fact that both the demographics of fraudulent & legitimate behaviors fluctuate, or the data sets employed are extremely biased. The variables utilized and the approach employed to identify fraud have an influence on the functioning of fraud detection.

## Social issues:

 It is  critical problem to comprehend since it has a serious influence on claims-related losses and insurance costs, particularly for companies. Increased legal expenses have an effect on insurers' claim payouts, loss of control, including, consequently, how much customers pay for coverage. Summary of the progress to date

## For written work

I have done write chapter 1 introduction part and chapter 2 literature reviews

## For practical work

Collect a data set from open source and performing data analysis and data visualization technique

# Project plan:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Activity** | **Starting date** | **Target date** |
| 1 | Choose research area. | Done |  |
| 2 | Literature search (find & confirm research gap). | Done |  |
| 3 | Decide on research question and/or hypothesis. | Done |  |
| 4 | Write the literature review draft. | Done |  |
| 5 | Scope out research (initial thesis plan). | Done |  |
| 6 | Write the proposal’s draft. | Done |  |
| 7 | Submit the proposal. | Done |  |
| 8 | Proposal approved. | Done |  |
| 9 | Literature review. | Done |  |
| 10 | Writing the report. |  |  |
| 11 | Submit the report. |  |  |
| 12 | Experiment the selected tools on identified dataset. |  |  |
| 13 | Develop the comparison criteria. |  |  |
| 14 | Develop the selection algorithms. |  |  |
| 15 | Feedback the entire thesis. |  |  |
| 16 | Submit results and findings. |  |  |

# References:

Bhowmik, R., 2011. Detecting auto insurance fraud by data mining techniques. *ournal of Emerging Trends in Computing and Information Sciences 2, no. 4 (2011):.*

Gomes, C. Z. J. a. H. Y., 2021. Insurance fraud detection with unsupervised deep learning.. *ournal of Risk and Insurance 88.3.*

Gupta, A. a. M. C. L., 2022. Comparative Analysis of Numerous Approaches in Machine Learning to Predict Financial Fraud in Big Data Framework.. *Soft Computing: Theories and Applications. Springer, Singapore, 2022. 107-123..*

Hanafy, M. a. R. M., 2022. Classification of the Insureds Using Integrated Machine Learning Algorithms: A Comparative Study.. *Applied Artificial Intelligence .*

Hansson, A. a. H. C., 2022. Insurance Fraud Detection using Unsupervised Sequential Anomaly Detection..

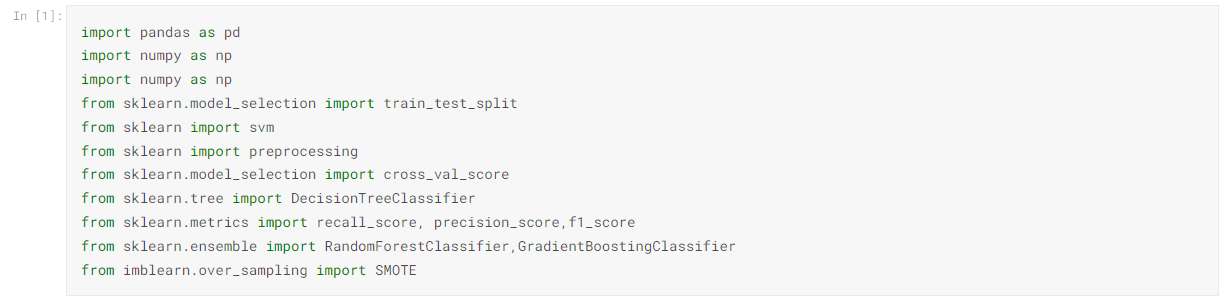
Itri, B. Y. M. Q. M. a. B. O., 2019. Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS.*

Nian, K. H. Z. A. T. T. C. a. Y. L., 2016. Auto insurance fraud detection using unsupervised spectral ranking for anomaly.. *The Journal of Finance and Data Science 2, no. 1 .*

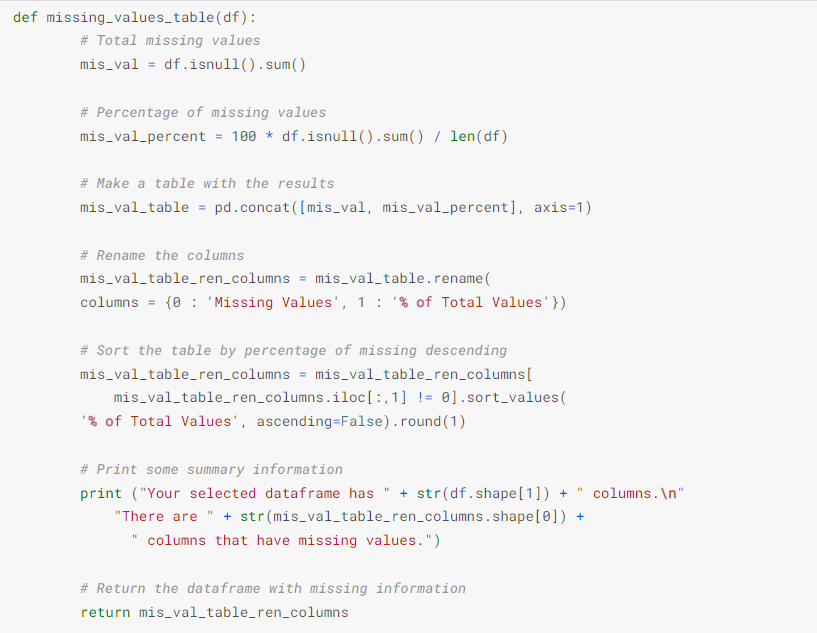
Roy, R. a. K. T. G., 2017. Detecting insurance claims fraud using machine learning techniques.. *In 2017 international conference on circuit, power and computing technologies (ICCPCT), .*

Vyas, S. a. S. S., 2022. Fraud Detection in Insurance Claim System: A Review.. *In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 922-927.*

# Appendix:



Import libraries



Checking missing values