Banking Fraud detection using Machine Learning

Authors**
Prateek Dahiya| Harmeet Patel| Parth Patel | Jugraj Singh| Gurdatt Singh
Data Analytics for Business, Zekelman School of Business and Information Technology

Table of Contents

Table 1. Abbreviations.

ML	Machine Learning
ANN	Artificial Neural Network
QA_1	Quality Assessment Compliance ID 1 in Table 2

***Table 2**. Quality Assessment of the project.

S/N	**Quality Assessment**	**Evaluation**	**Quality Assessment Compliance ID**
1	Is the business problem clearly defined?	Yes	QA_1
2	Is the background of the business problem presented?	Yes	QA_2

3	Is market research done for performance benchmarking?	Yes	QA_3
4	Are metrics to measure the success of the project clearly defined?	Documented, In review	QA_4
5	Is the project process flow clearly defined?	Yes	QA_5
6	Are the conclusion and recommendations clearly stated?	Pending	QA_6

Business Problem

Identification and forecasting of fraudulent instances in banking transactions commonly referred as Banking Fraud.[*QA_1]

Overview

Financial fraud is any kind of deceptive tactics which is used for gaining financial benefits illegally. It has recently become a widespread problem across the globe. Digitization of services like Conventional services such as such as e-commerce, healthcare, payment and banking systems has made financial fraud more easier to commit by fraudsters. With new ways of accessing, buying, and exchanging money online, the threat of fraud to all banks, organizations, and individuals has never been higher.

> "Banks globally are seeing an increasing trend in scams." [1]
>
> -KPMG

Manual verifications and inspections are imprecise, costly, and time consuming for identifying such fraudulent activities. With the advent of artificial intelligence, machine-learning-based approaches can be used intelligently to detect fraudulent transactions by analysing a large number of financial data.[*QA_2]

Analytical Methods

Methods already used in the industry[*QA_3]:

1) **Support Vector Machine (SVM)**
   SVM is a supervised ML method that seeks a maximum margin hyperplane forclassifying input training data into two categories. SVM is capable of classifyingnew data points based on a

labelled training set for each class. Based on the reviewedliterature, several researchers investigated SVM techniques for fraud detection. [2]

2) **Fuzzy-Logic-Based Method**
Fuzzy logic (FL) is an effective conceptual framework for addressing the issue of representing the data in a context of uncertainty and ambiguity. It is a logic that shows that methods of thinking are not accurate but estimated. The Fuzzy combinations offer effective concepts for handling complex modelling in a new and better way. Several methods based on the FL have been used for fraud detection. To detect anomalous behaviours in credit card transactions, the FUZ-ZGY hybrid model, based on the fuzzy and Fogg behavioural models. A system based on fuzzy logic was employed to track the historical activities of the merchant, and the Fogg behavioural method was employed to characterize the customer's behaviour along two different but related dimensions: the ability to commit fraud and motivation. Another fuzzy-based method was proposed in to detect fraud in credit cards by categorizing the fraud transactions and non-fraud transactions with decreased false positives. The method used fuzzy c-means clustering and the ANN model. The model was evaluated on synthetic data and the results showed that the combination of clustering techniques and learning mechanisms help in reducing false positives.[3,4]

3) **Hidden Markov Model (HMM)**
The HMM is a dual embedded random method often used to perform more complex random processes better than the traditional Markov model. The method used is more effective for credit card fraud detection[3]. A similar approach was used to achieve internet banking fraud detection by disclosing the right users and monitoring their illicit behaviours.

4) **Artificial Neural Network (ANN)**
ANN is an information-processing technique inspired by biological neural network behaviour. ANN is very powerful when there is the availability of a large volume of data [95–97]. Several ANN-based methods have been proposed for fraudulent detection in the financial sector. The model increased the detection rate and reduced false negative costs at various instances.[5]

5) **KNN Algorithm**
The K-nearest neighbours (KNN) algorithm is a convenient, straightforward supervised ML technique that is powerful in addressing both regression and classification processes. The class label is usually determined by the KNN model using a small set of the nearest samples. The KNN model is a type of non-parametric model that is used for both classification and regression tasks and that can locate similar neighbourhoods that are closest to a given sample point in a dataset and create a new sample point based on the distance between two samples of data. Although it worked well on many datasets, the performance of this technique is likely compromised by unbalanced datasets. The Euclidean distance is one of the most well-known techniques for calculating distance.

6) **Bayesian Method**
The Bayesian model (BN) is a particular type of graphical model that takes into account both independent and conditional relationships between various variables. A directed graph's nodes and edges are used by the BN. The Bayesian model is a particular type of graphical model that takes into account both independent and conditional relationships between various variables. A directed graph's nodes and edges are used by the BN. This model is very

powerful in searching anonymous probability computations. Based on the reviewed literature, we explored different papers on the two main types of Bayesian methods, namely, the Bayesian belief network and Naive Bayes (NB). NB is an ML model that is based on the Bayes theorem and is used to predict membership probabilities per class. It predicts a given data point label based on the probability that belongs to a particular category. Some researchers utilized the NB model for financial fraud detection.

7) **Decision Tree**
A decision tree (DT) is an ML technique that is used for creating decision supporttools in the trees of inner nodes, which represent binary options over the features.For many years, there have been several methods based on the decision trees that arewith different accuracy metrics. The results indicated that DT performed better than theexisting approaches with a high degree of accuracy. A study was conducted in forauto fraud detection by using an ML technique. The authors compared three differentmethods including NB, DT, and RF methods, and the result proved that DT outperformedother methods.[3]

8) **Clustering**
Clustering is an unsupervised learning method that involves grouping identical instances into the same sets. Although Clustering techniques are popular in financial fraud detection, they were, however, implemented considerably less than classification techniques in the industry.[3]

9) **Logistic Regression**
Logistic regression (LR) techniques are mainly applied in binary and multi-class classification problems. It operates by performing regression on a set of variables. It is typically a useful technique for describing patterns and clarifying connections between numerous dependent binary variables. In line with a review article by Abbasi et al. the logistic regression method is one of the most used machine learning (ML) techniques for detecting financial misstatement models. A majority of the algorithms used LR techniques for financial fraud detection.[3]

Performance Evaluation Metrics Used for Financial Fraud Detection Using Machine Learning Models

To evaluate the performance of a model, the evaluation metric is very important in financial fraud detection. However, there are no specific evaluation measures that are strictly used for evaluating ML techniques for fraud detection. In recent times, several performance evaluation metrics have been employed by different researchers that include accuracy, precision, recall, F1 measure, false-negative rate (FNR), the area under the curve (AUC), specificity, etc. [*QA_3]

Challenges

Imbalanced Dataset Virtually, most financial transaction datasets comprise millions of transactions, and all of them share a common issue, namely imbalanced datasets. On other hand, the number of fraudulent financial transactions is far fewer than non-fraudulent ones. This issue is generally caused as a result of the fact that the rate of actual fraud transactions out of all transactions is nominal. The problem of imbalanced data distribution generally affects the efficiency of machine learning models. Therefore, training models for detecting fraudulent activity that is very minimal requires extra consideration.

To address the issue of imbalanced data, some studies have applied oversampling approaches[3]. The others attempt to introduce approaches that may effectively work with extremely imbalanced data. Some have utilized the oversampling method for future work and some avoided oversampling as it may lead to choice-based sample biases. Moreover, the only applied oversampling method was the SMOTE (Synthetic Minority Oversampling Technique). Because of this, it could be deduced that future studies could consider employing other oversampling techniques, as well as under-sampling techniques.[3]

Initial test approach for imbalanced data

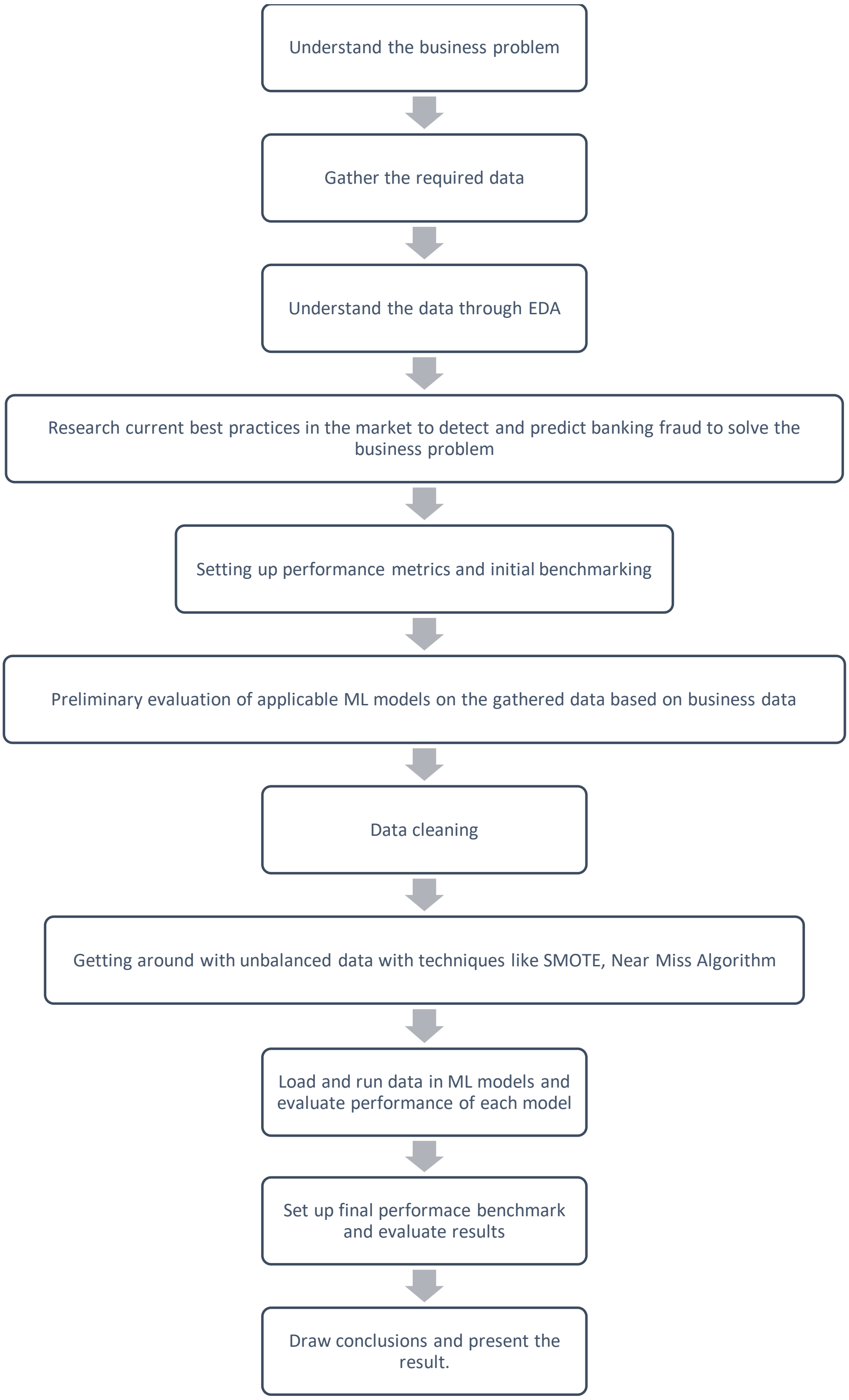
The data set we are going to work on has an Imbalanced Data Distribution, generally happens when observations in one of the class are much higher or lower than the other classes. As Machine Learning algorithms tend to increase accuracy by reducing the error, they do not consider the class distribution.

Imbalanced Data Handling Techniques:
There are mainly 2 mainly algorithms that are widely used for handling imbalanced class distribution.

1) SMOTE
   SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.
   It aims to balance class distribution by randomly increasing minority class examples by replicating them.

2) Near Miss Algorithm
   Near-miss is an algorithm that can help in balancing an imbalanced dataset. It can be grouped under sampling algorithms and is an efficient way to balance the data. The algorithm does this by looking at the class distribution and randomly eliminating samples from the larger class. When two points belonging to different classes are very close to each other in the distribution, this algorithm eliminates the datapoint of the larger class thereby trying to balance the distribution.

Project Pipeline


Understand the business problem
Gather the required data
Understand the data through EDA
Research current best practices in the market to detect and predict banking fraud to solve the business problem
Setting up performance metrics and initial benchmarking
Preliminary evaluation of applicable ML models on the gathered data based on business data
Data cleaning
Getting around with unbalanced data with techniques like SMOTE, Near Miss Algorithm
Load and run data in ML models and evaluate performance of each model
Set up final performace benchmark and evaluate results
Draw conclusions and present the result.

References

[1] https://home.kpmg/xx/en/home/insights/2019/05/the-multi-faceted-threat-of-fraud-are-banks-up-to-the-challenge-fs.html

[2] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[3] https://www.mdpi.com/2076-3417/12/19/9637

[4] https://www.edureka.co/blog/fuzzy-logic-ai/

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7031719/#:~:text=In%20the%20present%20research%20study,is%20used%20for%20fraud%20detection.