Data Analytics for Business

Zekelman School of Business and IT

Banking Fraud detection using Machine Learning

Prateek Dahiya, Harmeet Patel, Parth Patel, Jugraj Singh, Gurudatt Singh

Associated Professor
Prof. Manjari Maheshwari

Table of Contents

Table 1. Abbreviations.

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

*__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?	Yes	QA_4
5	Is the project process flow clearly defined?	Yes	QA_5
6	Are the conclusion and recommendations clearly stated?	Yes	QA_6

Abstract

Financial fraud, seen as deceptive tactics to gain financial gain, has recently increased. Manual checks and inspections are inaccurate, costly and time consuming. With the advent of artificial intelligence, machine learning-based approaches can be intelligently used to detect fraudulent transactions by analyzing large amounts of financial data.

Business Problem

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

Introduction

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. It has recently become a widespread problem across the globe.

> "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

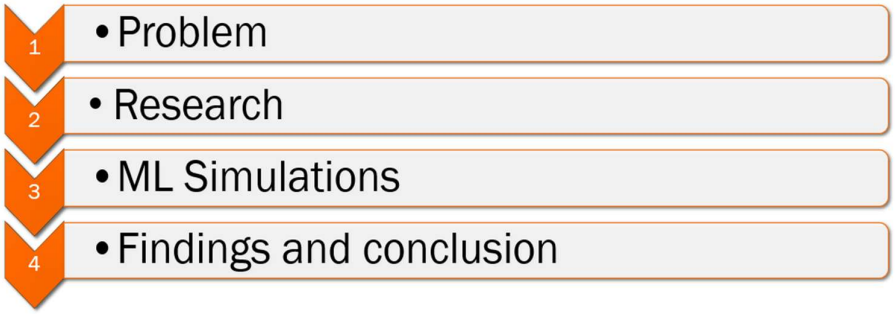
Financial fraud can occur in a variety of sectors such as corporate, banking, insurance, and taxes sectors. Money laundering, financial transaction fraud, and other forms of financial crime have recently become more and more of an issue for businesses and industries. Despite several initiatives to curtail financial fraud, it continues to negatively impact society and the economy since significant sums of money are lost to fraud every day. Several methods for detecting fraud were first introduced many years ago. The majority of old procedures are manual, which is not only time-consuming, expensive, and inaccurate, but also unworkable. More and more studies are being done to limit transactional fraud in banks; however, they are ineffective in reducing losses brought on by fraudulent activities. With the advancement of the artificial intelligence (AI) approach, machine learning and data mining have been utilized to detect fraudulent activities in the financial sector. Both unsupervised and supervised methods can be employed to predict fraud activities.

Research Methods

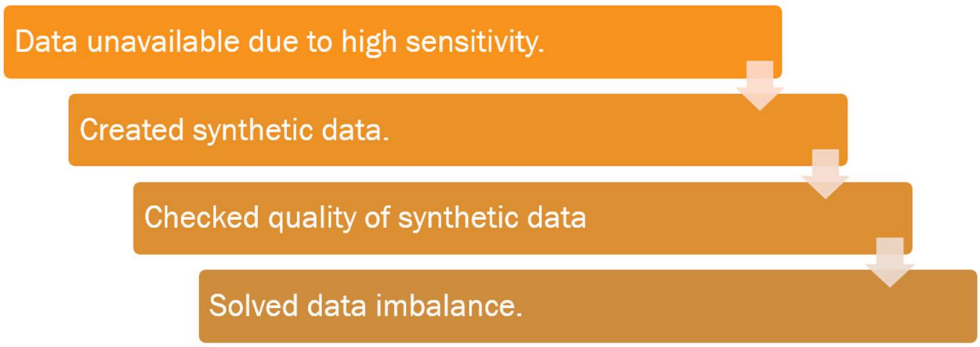
Many methods were considered to achieve the goals of this project which were to detect and predict fraudulent banking transactions. Exhaustive internet research was done to find journals, websites, blogs and research papers so as to get an idea of the best practices being used in the market to detect banking fraud. Also, experienced professionals both from the Industry and college were contacted to share their experience on the topic. This approach helped immensely to set up initial performance metrics which then enabled to judge the success or failure of the project quantitatively.

Project process flow

The project was conceptualised with a simple yet effective process flow strategy.



The problem statement was formalised first which was followed by research and ML simulations; and from those, the findings and final conclusions were interpreted. During the problem formulation, business problem was coded into parameters which could be measured quantitatively. Then the data was procured. Banks generally do not share their transactional data hence, the data was synthesised using a combination of randomising functions in Python, copulas and deep learning libraries specifically coded to generate data by looking at any given sample data. The then created sample data was then test for fit for probability distributions using Kolmogorov–Smirnov tests and Chi square tests.



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 for classifying input training data into two categories. SVM is capable of classifying new data points based on a labelled training set for each class. Based on the reviewed literature, 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. Several ANN-based methods have been proposed for fraudulent detection in the financial sector. The model increases 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. There are 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 support tools 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 are with different accuracy metrics. The results indicated that DT performed better than the existing approaches with a high degree of accuracy. A study was conducted in for auto fraud detection by using an ML technique. The authors compared three different methods including NB, DT, and RF methods, and the result proved that DT outperformed other 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]

Fraud types and ML models

Fraud Type	Description	Technique Used
Financial Statement Fraud	This is a corporate fraud such that the financial statements are illegitimately modified to allow the organizations to look more beneficial.	Support Vector Machine Clustering based method Decision Tree Logistic Regression Naïve Bayes Artificial Neural Network

Credit Card Fraud	Illegitimate use of the card without proper owners' authorization	Support Vector Machine Fuzzy logic Clustering based method Artificial Neural Network Hidden Markov model Decision Tree Genetic Algorithm Artificial Neural Network Naïve Bayes Logistic Regression Random Forest
Auto Insurance Fraud	Fraudulent claims by individuals or organizations to support the relevant expenses of theft or accidental damages.	Support Vector Machine Artificial Neural Network K Nearest Neighbors Naïve Bayes Clustering-based method
Health Insurance Fraud	Fraudulent claims by an individual to get health insurance profits.	Support Vector Machine K Nearest Neighbors
Cyber Financial fraud	Financial fraudulent activities through cyber space.	Artificial Neural Network SVM
Others	Other frauds that are faced in the financial domains include commodities and securities fraud, mortgage fraud, corporate fraud, and money laundering.	Support Vector Machine Decision Tree Fuzzy logic Clustering-based method Hidden Markov model

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

Creation of synthetic data

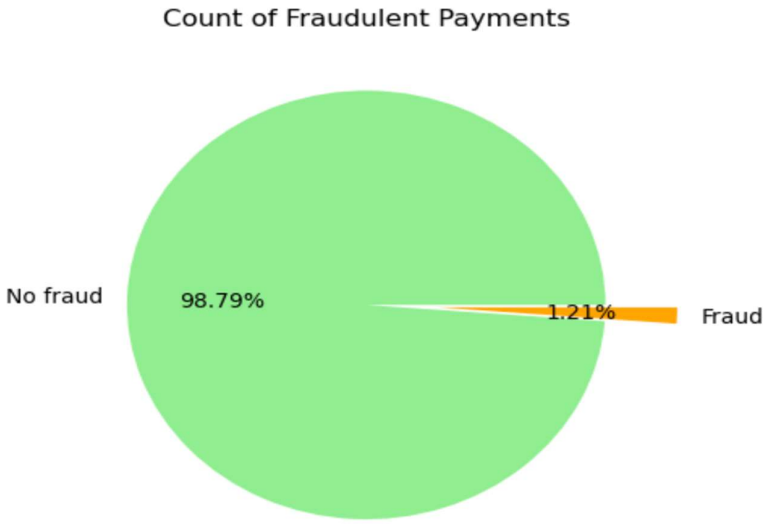
Since banks were hesitant to share their transactional data, it was decided to use synthetic data instead. The synthetic data was designed to mimic real world transactional data. Many options were evaluated to generate synthetic data by using modules and libraries in Python like Faker, np.random etc. However the synthetic data was tested using Kolmogorov–Smirnov (KS) tests and Chi-Squared (CS) tests. KSTest is used to compare the continuous columns, and CSTest compares the discrete columns. Both tests result in a normalized score between 0 to 1, with the target is to maximize the score. With this method, the discrete sample columns are almost similar to the real data. In contrast, continuous columns might have a deviation in distribution.

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]

Data imbalance and approach for handling imbalanced data



The synthesised data set had an Imbalanced Data Distribution, this 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) Under sampling from majority
2) Over sampling from minority
3) Ensemble
4) SMOTE (Subset of Over Sampling from minority)
   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.

F1 Score over Accuracy

The baseline threshold of the dataset had 98.79% accuracy. In order to overcome this problem, F1 score was used to detect correct classification of fraudulent activities. Also a classification matrix to detect the same was used.
True positive (TP): attacks/intrusions that are accurately flagged as attacks.
• True Negative (TN): normal traffic patterns/traces that are successfully categorized as normal.
• False positive (FP): legitimate network traces that are incorrectly labelled as intrusive.
• False Negative (FN): attacks/intrusions that are incorrectly classified as non-intrusive.

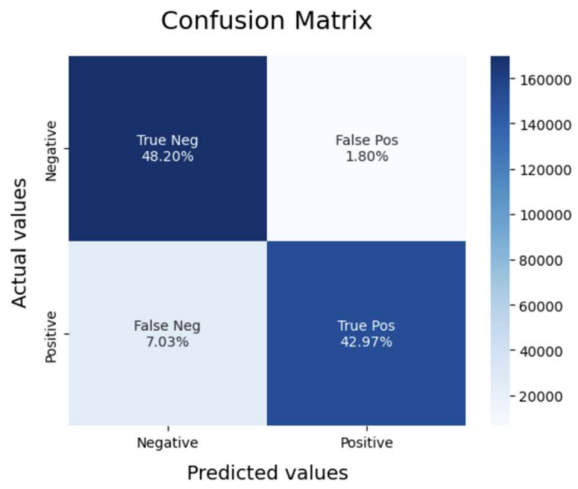
Confusion matrices of the applied ML models

The following machine learning models were applied to the synthetic dataset. The confusion matrices of which are summarised below:

Support Vector Machine (SVM)

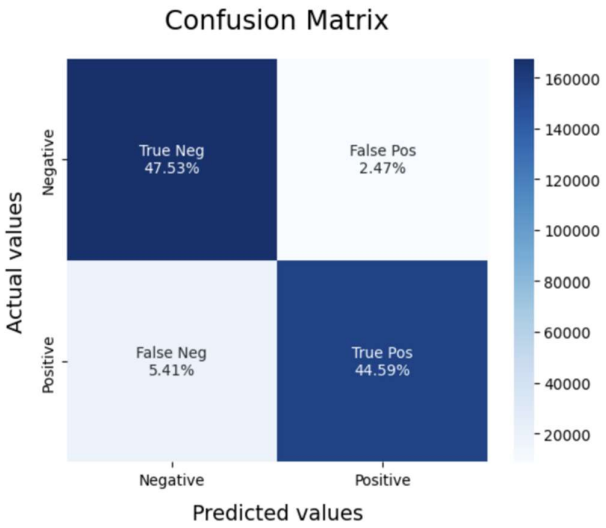
Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

It seeks a maximum margin hyperplane for classifying input training data into two categories. SVM can classify new data points based on a labelled training set for each class.



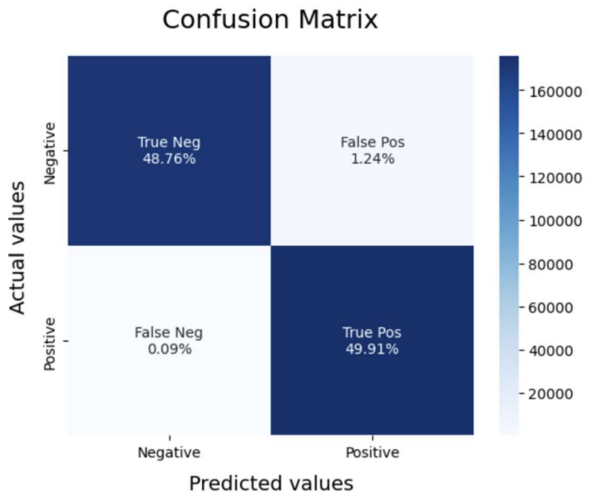
Logistic Regression

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.



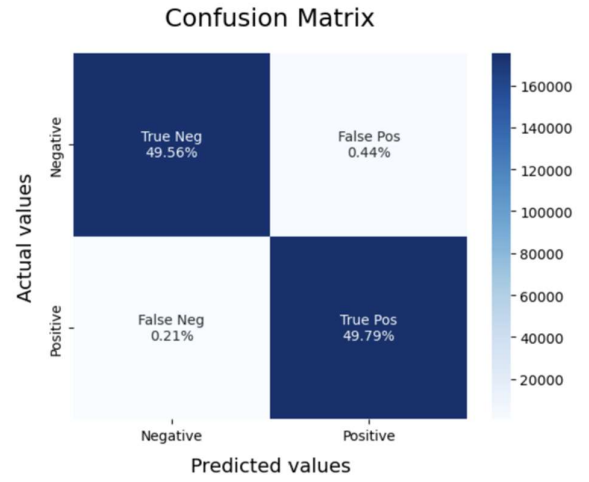
KNN Algorithm

The K-nearest neighbours (KNN) algorithm is a convenient, straightforward supervised ML technique that is powerful in addressing both regression and classification processes.



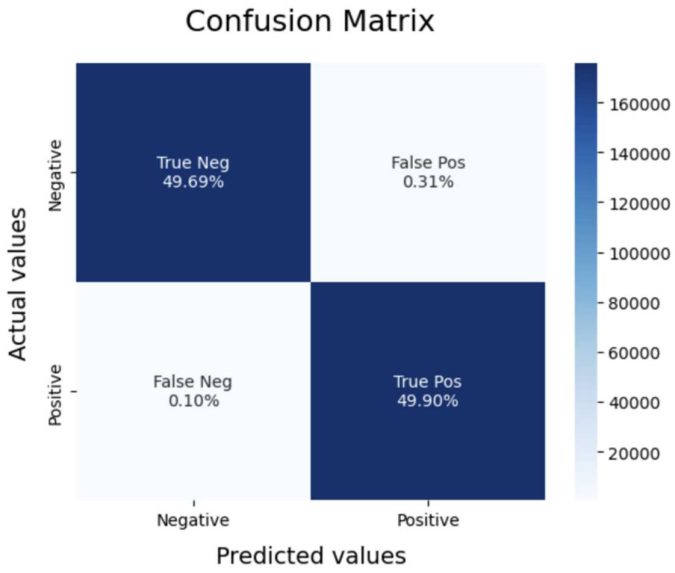
Decision Tree

A decision tree (DT) is an ML technique that is used for creating decision support tools in the trees of inner nodes, which represent binary options over the features.



XG-boost classifier

XGBoost is a distributed gradient boosting library that has been optimised for quick and scalable machine learning model training. A number of weak models' predictions are combined using this ensemble learning technique to get a stronger prediction. Extreme Gradient Boosting, or XGBoost, is one of the most well-known and widely used machine learning algorithms because it can handle large datasets and perform at the cutting edge in many machine learning tasks like classification and regression.

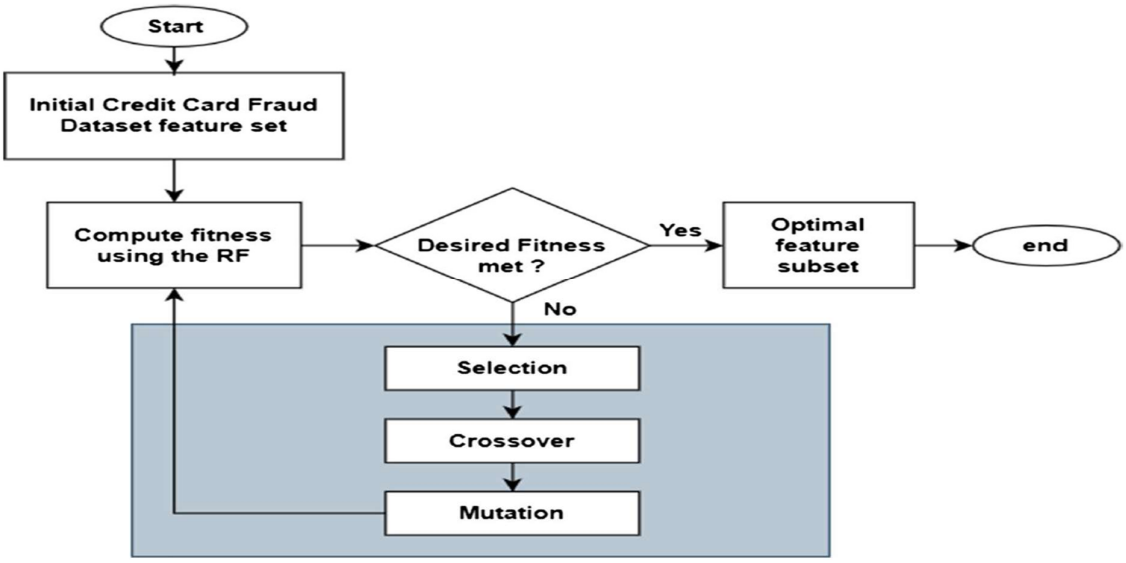


Below are the F1-scores for the models listed in the table and actual model performance:

Model Name	F1 score
Support Vector Machine	0.92
Logistic Regression	0.92
K-Neighbours Classification	0.99
Decision Tree Classification	0.99
XG-boost classifier	0.99
Random Forest classifier	0.98

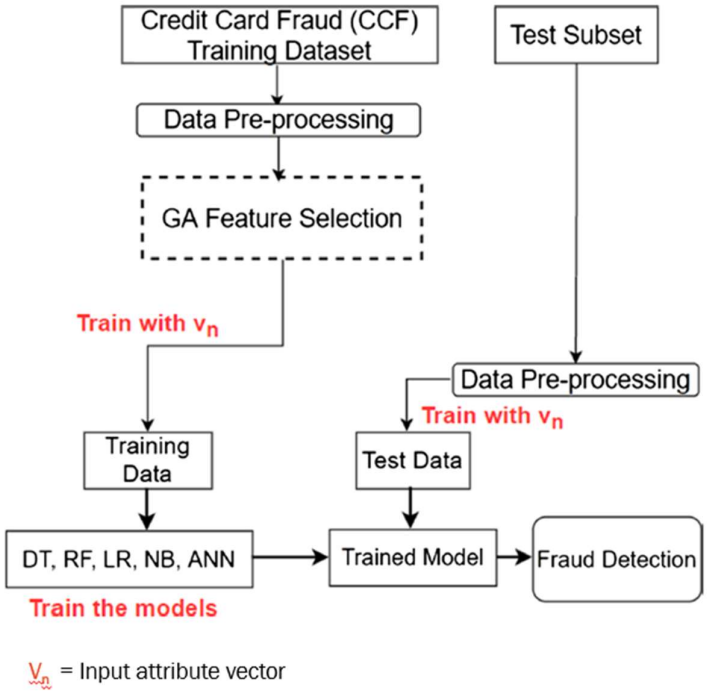
Overfitting of ML models

Most of the ML models which were fed the synthetic data overfitted and were showing similar results for F1 score. Further research is required to find the reasons behind overfitting. One approach can be algorithmic feature engineering to select only certain features in the dataset which show greater chances of detecting fraud. This approach is shown below:



Opportunities

Live transaction data needs live fraud detection. Such a problem can only be solved by using an algorithm which automatically selects critical features for fraud detection and such a feature vector can be put in a training and testing ML algorithm. An example of which is shown diagrammatically below:



V_n = Input attribute vector

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.

Scholarly articles:

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8

https://www.sas.com/en_ca/insights/articles/risk-fraud/fraud-detection-machine-learning.html