Deep Science Publishing
https://doi.org/10.70593/978-93-49910-41-6_7

DeepScience
Open Access Books

Chapter 7: Enhancing fraud detection with machine learning and pattern recognition

7.1. Introduction

The shift from traditional fraudulent paper-based transactions to the electronic world means that the discovery of new fraudulent schemes is essential. The volume of transactions occurring on a global scale is simply huge, and traditional methods of fraud detection, such as handcrafted deterministic rules or reward models, are very time-consuming and unscalable. Many such rules can generate a mountain of false positives and significantly increase the cost of detecting the real "needle." For that reason, modeling fraudulent patterns has gained importance lately. Particularly, data mining, as a means of extracting implicit but useful relationships and patterns between various data attributes, has proven to be increasingly effective in the detection of fraudulent attempts (Phua et al., 2010; West & Bhattacharya, 2016).

The use of intelligent data analysis and data mining has increased due to the proliferation of different electronic means of monetary transactions. A variety of financial crimes, such as corporate, association, and consumer fraud, money laundering, loan scams, as well as various terrorist activities, all of particular interest to banks, benefit from the early detection of fraudulent intentions. The use of data mining seems to be particularly relevant in credit card transactions, which are growing rapidly in electronic business, but the percentage that is inspected by hand is low. Consequently, this new unbalanced proportion has increased the number of unchecked potentially costly fraudulent events and unlawful commerce. The huge amount of electronic transactions, with their representative complexity of procedures, makes this economic activity particularly ripe for the application of data mining (Abdallah et al., 2016; Bahnsen et al., 2016; Fiore et al., 2019).

7.1.1. Purpose and Scope of the Study

The research explored how an integrated multiple model fraud detection approach improves detection accuracy of false positives while providing practical operational complexity. This was performed using both supervised and unsupervised learning models. The data was part of a large bank that has a presence in many countries, with thousands of ATMs, and processed many transactions within a day. The bank's ATM transactions were used to distinguish fraudulent from non-fraudulent transactions because of a massive fraud rate. This study provides several significant contributions. Although some other studies combined unsupervised and supervised models for a single fraud detection task, the current study pioneers the development of a multiple model fraud system to improve detection accuracy, balance classification scores, and show how to handle different features suitable within an operational banking context.

The bank can monitor all bank transactions from various channels, but the sighting of a fraudulent card swiped inside a bank branch, even with many security measures in place, could have a significant impact on the issuing bank. The challenge is to detect in real-time what typically happens on card transactions within a few seconds without having any idea whether the client is legitimate or fraudulent. The bank has to make quick decisions on whether to curb the transaction or accept the applied charge. What makes this complex situation is the frequency of transactions and the daily attempts by each card at different ATMs worldwide. With thousands of ATMs situated in various countries and executing many transactions a day, the bank needs a quick detection system that is both able to tag the small number of frauds in the millions of daily transactions and calm the clients during numerous card rejections.

7.2. Understanding Fraud

Understanding the concept of fraud within your business is the first step in putting prevention measures in place. While fraud can come in many forms, including employee theft and financial statement fraud, the most common type is asset misappropriation. This refers to someone stealing or misusing the company's resources. It comes in many forms, such as skimming, fake refunds, billing or payroll schemes, theft of money or inventory, or even the use of company property for personal purposes. One area where fraud flourishes is with staff expenses. Expense fraud can happen at every level, from hourly staff to middle management and the top executives. One of the common behaviors seen with expense fraud is vague descriptions.

It is not just internal sources of fraud that are the problem. As customers, businesses need to be able to notice, detect, and prevent potential cases of fraud. Payment fraud can come in many different forms, including counterfeit cards, lost or stolen cards, account

balances, and unauthorized transactions. By the time a chargeback occurs, it is usually too late for real-time fraud monitoring. However, there are many ways that businesses can use technology to create smarter payments. Data analysis is a potent tool. Any fraudulent activities left undetected within your customers' business will more than ever erode your top line.

7.2.1. Types of Fraud

Fraud can be committed in many forms. The types of fraud often encountered can be broken down into three categories: corruption, misappropriation of assets, and fraudulent financial statements.

In corruption, fraud involves a wrongdoer who misuses his or her influence in a business transaction for dishonest gain. This usually means that the wrongdoer uses his or her authority in the business to extort funds from third parties in exchange for certain services or facilities. In the fraud triangle model employed in criminology, corruption schemes are usually an abuse of power or position. These problems occur when an employee uses his or her authority in a social economy to create a deception. Corruption can be found in various environments.

The next type of fraud is the misappropriation of assets. In this type of fraud, an employee steals an employer's resources by utilizing the employee's unauthorized access to a company's resources to steal money, equipment, or any other assets in the company. This type of wrongdoing is often the simplest way to benefit oneself criminally without mistakenly damaging a company's underlying financial reporting systems since the inappropriate actions are likely to result in the taking of company resources rather than using them to capture and render the financial results of the corporation. Misappropriation also triggers a wide range of schemes that involve paybacks, fictitious billings, shell companies, and more, all intended to make it easier for an employee to transfer ownership of the company's assets to his or her control. The final category of fraud, namely, fraudulent financial statements, is a more difficult, less well-defined form of corporate wrongdoing. In fraudulent financial statement wrongdoing, a company's financial accounts are manipulated by the executive management in order to show the firm in a better or more controlled situation than is the case. In other words, the firm's financial results are knowingly misstated by those in a position of authority, thereby deceiving financial statement users and investors into believing that the company has achieved growth or recovered from previous problems. Management and the board of directors of a company are also expected to defraud the company while involved in fraudulent financial statement wrongdoing.

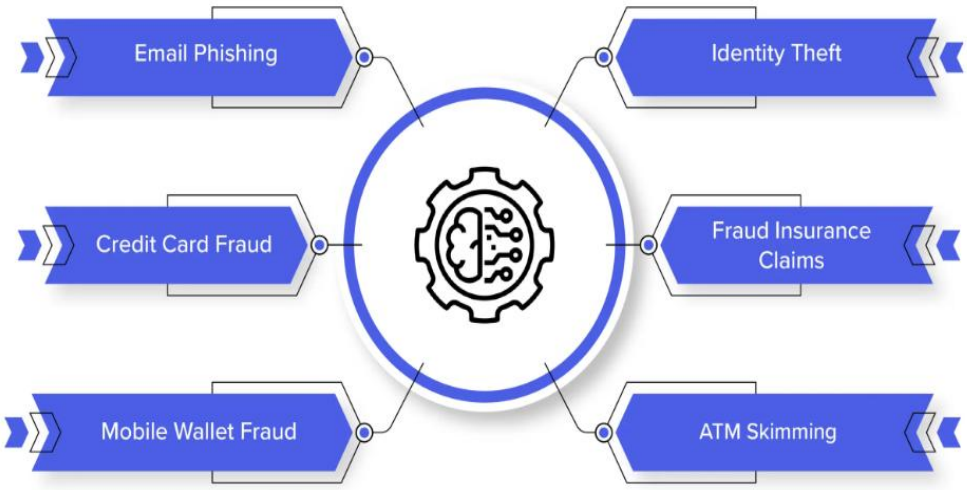

Fig 7 . 1 : Financial Fraud Detection Using Machine Learning

7.2.2. Impact of Fraud on Businesses

There is an old saying that a crooked businessman is a fool's best friend, and a businessman will easily recognize it, for a good businessman feels the burden of a rogue's business. A good businessman's trade cannot tolerate a stray businessman. That is, their costs shift to the businessperson least qualified to detect them and thus most vulnerable to bearing the costs, the businessman who did nothing to cause or control them. Sometimes, we refer to such outsiders as non-customers. In honest trade, businesses prosper if and only if they serve any combination of customers, producers, and society. No customer's business is favored over any producer's business, for neither is in a position to pass costs on to someone else. Quality is a decisive competitive advantage in the absence of shoddy trade.

In fraudulent trade, the focus suddenly changes. The only way for businesses to make any under-the-table money, in addition to various legitimate fees, is to make someone else pay for their otherwise uneconomic deals. Businesses prosper by shifting as many costs as possible to non-customers and, if they can get away with it, to any producer foolish or desperate enough to be eligible for the booby prize. Money skimmed off cost shifting can be additional income or lavish compensation. Transferring mismanagement costs to non-customers can keep seemingly inefficient businesses in operation indefinitely. Unsatisfactory results are simply expected; no one else seems to be any better. These shoddy businesses thrive while frustrating the efforts of more purportedly efficient companies. These deadbeats mix with frauds posing as victims, but cost shifting

does increase the deadbeats' burden on other businesses. The time is cheap, but the profits are forever.

7.3. Traditional Fraud Detection Methods

Traditionally, credit card fraud detection was handled by manually specifying rules used for detecting fraud. This is useful because people have a natural ability to interpret the results and then generalize the rules that are tracked into the fraud detection system. In the 1990s, these rules became codified into simple expert programs to emulate human expertise. As the number of credit card transactions grew, these systems became less effective and, in practice, tended to have very high false positive rates, which led to wasted effort and destruction of consumer confidence. These early fraud detection algorithms suffer from their reliance on the characteristics of historical fraud.

Modern fraud detection systems use a combination of supervised machine learning algorithms, together with clamping mechanisms and expert systems containing a set of rules written by the humans within a bank's fraud detection team. However, some real-time systems are heavily reliant on the rules and overreact or over detect using the expert flags. Therefore, in practice, we find that existing systems are highly dependent on human experts to tweak the detection system and can be very dependent on human expertise.

7.3.1. Rule-Based Systems

Amid the development of data analysis techniques, there are still many institutions that develop rule-based systems in order to process an extensive amount of information. These types of systems are very useful for handling many types of fraud situations. Rule-based methodologies require a team of experts to decide on the rules that allow the system to detect fraud, especially when dealing with financial information. These rules are associated with performance that is characteristic of specific types of fraud. With the aid of a team of expert auditors and accountants, it is possible to outline the rules to handle the whole organization. Rule-based systems are relatively easy to develop. This is why so many companies use this type of methodology. Unfortunately, there are many different types of fraud that result in significant financial loss each year. Some of them are difficult to detect and may not have a consistent indicator. Because of this, many professional fraud auditors argue that developing these characteristics is difficult and may lead to many missed opportunities. Even with a lot of support, important details about the data may be lost, which might have some effect on the detection of fraudulent activity.

7.3.2. Statistical Methods

The previous methods are designed for unsupervised learning, which assumes that the normal samples are available as well as abnormal samples. In many prediction and classification applications, supervised learning is conducted only when a training set is given, and the test sample would be determined as normal or abnormal in accordance with the previous sample statistics. The training samples only have the labels of normal or abnormal, and sometimes they do not change with time. We call such a situation training set based supervised learning or TSBSL. However, one issue is that the distribution of the test samples is different from that of the training samples. In that case, the performance of the classification model, induced from the training samples, is far from being satisfactory. Therefore, both the generalization of the classification and the wide applicability of the statistical model are two key issues for many real-world applications, particularly those using statistical and machine learning methods.

We introduce the notion of sample label type to address the change of the test sample distribution in relation to that of the training samples. We also present that TSBSL is the direct statistical modeling method for the sample label type problem and address the problem of label discrepancy, resulting from the different distributions between the test and training samples. To solve the previous problems, we propose a method of learning for TSBR where the inductive bias would be reduced by using basis functions. Our objective is to perform supervised learning robustly by using extensive empirical evidence and a fine-tuning process. Our main contributions are twofold. First, we initialize the labels for the test samples by using the functions of label dependence. Second, the method of learning does not require solving the underlying ground truth and could treat the images of finite samples through the integration concept of statistical learning.

7.3.3. Limitations of Traditional Methods

Traditional fraud detection methods have three central drawbacks: the need for detailed information on previous fraud activities; the high cost and low efficiency of the detection processes; and the heterogeneous characteristics of diverse fraud models. Each of these drawbacks can limit the scope of anti-fraud activities. However, recent advances with pattern recognition techniques and machine learning models can facilitate the completion of fraud tasks, such as classification and verification. The use of traditional techniques for the detection of fraud and the prediction of fraud risk is important because of the large volumes of money involved in fraud activities. However, the application in a company can be difficult because traditional techniques have a number of limitations. The three central drawbacks are: the need for detailed information on previous fraud activities; the high cost and low efficiency of the detection processes; and the

heterogeneous characteristics of diverse fraud models. One potential solution to these issues is to use data mining technology to enhance existing fraud detection methods, as it can be applied to each of the three types of fraud. The pattern recognition technique has been widely used to solve problems, such as classification and verification, which can facilitate the completion of fraud tasks.

7.4. Machine Learning in Fraud Detection

Machine learning has applications in a number of different domains where pattern recognition is important. More and more, machines today are learning from data and carrying out activities that have previously been done by only humans or by computers following the commands of humans. From a computer science perspective, a machine learns when it can carry out a task from experience and achieve an improvement in performance. The primary reason that the use of machine learning in fraud detection is growing is that fraud-detection problems grow in difficulty and complexity every day, data are changing every day, data are unstructured, and the amount of data is growing. Although experts can classify data using judgment, such judgments are unconstrained, leaving the analyst prone to unconscious biases and concealment from the user.

Companies also want to reduce the errors they make. Machine learning has the ability to learn how to differentiate between pristine and suspicious transactions in two main ways. One way uses supervised learning, where refined data are used to train the machine how to associate certain patterns with the pristine and suspicious transactions. Unsupervised learning acts without training or guidance and involves the application of algorithms for testing the ideas. Working in an unsupervised learning framework is a bit of a free-for-all because algorithms are allowed to do what they like with the data, without restrictions, and the aim is typically to investigate possible indicators of fraudulent behavior. Unsupervised learning through machine learning is a particularly popular fraud-detection means in the fields of financial crimes and cybersecurity because of the techniques' abilities to monitor the vast majority of transactions without being clouded or restricted by assumptions.

7.4.1. Overview of Machine Learning

The objective of this book is to demonstrate how such techniques can be used in the area of fraud detection. Most of the work, particularly in terms of advances in efficiency and effectiveness, has been achieved through the use of technologies such as personal computers, networks, intelligent systems, databases, data warehousing, data mining, and expert systems. In recent times, accounting and business assurance work has been further enhanced by a new generation of techniques and package solutions that have been

created to use high-speed computer-based technologies to detect suspicious and fraudulent practices. Because fraud detection and prevention is the focus of this book, a detailed examination of such components is not undertaken. These may include intelligent systems technology, machine learning and neural networks, heuristics and attribute indicators, genetic and fuzzy logic algorithms, and other heuristic searching techniques commonly used in artificial intelligence.

However, what can be stated is that each of these components has played an important role in shaping the fraud detection environment in terms of methods that are recognized as valid and structured. These techniques have formalized the steps and built the frameworks in determining what needs to be searched for, with the build-up of experience and expertise quite often used to develop more informative rule statements that have greater recognition capability. The implementation of established features of the search mechanism model across platforms, as platform constructs, provides a natural mechanism for encouraging reusable components and technology improvements across all domains. Because the key to machine learning's success may be due to its ability to detect complex, difficult-to-investigate, and more serious fraud, the choice of platform may also emphasize the incorporation of more sophisticated searching and pedagogic mechanisms.

7.4.2. Supervised vs Unsupervised Learning

There are two broad categories of machine learning tasks: supervised learning and unsupervised learning. In a supervised learning task, the model is trained using a set of examples from which the model learns patterns or characteristics that can then be used on new examples. One of the key advantages of supervised learning is the ability to use models in classification tasks on new data. In unsupervised learning, the model is trained using a set of data without examples of the outcome we wish to predict. The goal of the model is then to group the data into natural clusters. The most common applications of unsupervised learning relate to reducing dimensionality, finding natural clusters of input data, and detecting anomalies such as fraud. In the context of the credit card fraud dataset, the outcome information is missing; we don’t know a priori the fraud status of individual transactions in the dataset.

In the case of the fraud detection dataset, the lack of outcome information precludes the use of standard steps used in supervised learning, such as model fitting by minimizing prediction errors. Instead, the task for unsupervised learning focuses on uncovering patterns and relationships inherent in the data distribution. The output for unsupervised learning in the credit card fraud detection problem is to detect unusual and suspicious activity associated with fraudulent use. During the model-building step, this question is informed by the data and the probability distribution under consideration. The likelihood

estimating distribution step of model building is crucial in fraud detection. Indeed, an analyst who can understand, assess, and interpret the risk derived from a model will have a better assessment of the appropriateness of the chosen model. An output of an unsupervised learning task is the anomaly score. The anomaly score for a transaction is the probability of the transaction conditionally on its parent cluster, and this probability is used in the final algorithm output.

7.4.3. Common Machine Learning Algorithms

Common patterns used within machine learning for fraud detection are unsupervised and semi-supervised learning. Unsupervised learning encompasses training a model with only input data and no corresponding output data. It is permitted to act on its own to detect underlying patterns or features. Appropriate algorithms are clustering and autoencoder models. Clustering permits a model to group similar data points into clusters. Automating the process to find fraud patterns is suitable for a use case where the financial institution doesn't have fraudulent labels to begin with. A disadvantage of clustering is that it requires multiple iterations. Further, dimensionality is a concern that needs to be addressed; it is important to reduce the number of features during preprocessing to prevent overfitting. In addition, if only parts of the dataset are being iterated on, then the algorithm might skip smaller and sparser clusters containing fraud transactions.

Deep learning, essentially an automated feature engineering tool, has emerged as a practical choice often combined with supervised models. Autoencoder models can create new features by learning abstract representations from the input layer, composed of the original dataset, and passing through nonlinear transformable hidden layers. It is especially useful due to its high level of connections and capability to handle large-scale fraud detection datasets. The two distinct autoencoder architectures are the basic architecture and the generative adversarial autoencoders. The first minimizes the reconstruction errors between the dataset values and the decoded input data; the output should be as close as possible to the input data by reducing the amount of noise required in the temporary hidden layer. Both classifiers and the basic model are simultaneously trained and help in improving accuracy in predictions. The generative adversarial architecture can significantly reduce false negative detection results while training.

7.5. Pattern Recognition Techniques

A more powerful approach to design and develop a fraud detection system is to use a combination of different techniques such as knowledge discovery in databases, design and implementation of decision rules, artificial learning systems, and expert systems.

These systems can use a variety of techniques to uncover potential fraud that a single technique would not have the power to uncover. We concentrate on a system that uses expert rules and a machine learning system. The text also explores the use of pattern recognition techniques with the developed machine learning system that can improve the detection of insurance fraud. Pattern recognition systems are systems capable of identifying and recognizing patterns contained in raw data. Pattern recognition topics are concerned with the classification of linear algebra techniques, statistics, and probability.

Our goal here is to introduce five pattern recognition techniques that can be used with fraud detection algorithms and our decision model. The presented techniques are cross-validation, genetic algorithms, backpropagation of neural networks, SVM, and Bayesian classifiers. Out of these techniques, logistic generalized regression was already described and is used in an enhanced design of fraud detection algorithms. Cross-validation and genetic algorithms are found in the fraud detection algorithm, while backpropagation of neural networks, SVM, and Bayesian classifiers are discussed. In the next subsections, we look into these pattern recognition techniques with a focus on fraud detection features that can be a great assistance.

7.5.1. Introduction to Pattern Recognition

In a rather broad context, information can be classified into: (1) data; (2) a model or an estimated model; (3) a feature; (4) a decision or an explicit inference; and (5) a result or goal. The objective is to provide techniques to classify and identify patterns based on data. We are mainly interested in data that are subject to validation, cleansing and transformation, enhancement, visualization, query, mining, and decision making. Ideally, the identified patterns can eventually lead to some level of understanding and the creation of a model, which ultimately contributes to the advancement of science and technology.

There are several levels of concepts and techniques involved in pattern recognition. Generally speaking, pattern recognition is the subarea of machine intelligence that is concerned with the development of algorithms, protocols, and visual capacities, and the competence to be engaged in detecting, recognizing, classifying, and organizing patterns. These concepts and techniques can be classified into three major groups: (1) grouping or cluster analyses that partition existing patterns; (2) classification that generates criteria to automatically recognize patterns; and (3) determination of the best range of associations or distance among members of the group. Since fraud detection can be considered a binary classification or a clustering task depending on the applied technique, clustering techniques are occasionally used to identify fraud patterns instead of probability profiling management and standard rule-based management.

7.5.2. Feature Extraction Methods

In this part, we evaluate the effect of feature selection and comparison methods, which are tested on fraud detection datasets in the initial stages. The feature selection methods run on the entire dataset and calculate the common attributes to stay in the set, providing an adequate evaluation of the variables. We selected filtering methods that evaluate the power of features such as the Chi-Squared, the ReliefF, and the Information Gain; the algorithms of Wrapper and Embedded, and Rank filters are otherwise classified. The comparison methods classify both features and eliminate those that generally do not oversize the distribution plus the criteria specified. For comparison purposes, the classifier methods used on the training set to plan the model of the validation set calculate the importance of the characteristics. We selected comparison approaches such as the Decision Tree, Gradient Boost, Random Forest, and Extreme Gradient Boost for the same reason, all in well-known classifiers. The fraud classification experiments were carried out on four public datasets of different dimensions, so we could conduct the investigation thoroughly.

We believe that our postulates are considered the main contributions and that empirical aspects should be tested on many distinct fraud datasets. We would expect a high number of fraud attributes using the rule-based methods for any fraud dataset. Otherwise, postulate 2 can be false according to the sequential coverings of a class. Therefore, the detection stability of fraud generally decreases fraud detection availability when recursive estimation rules are used. We expect that all feature selection methods improve the predictive power of the fraud detection set similar to various standard fraud classifiers. Our feature selection process consists of five widely used filter methods, Wrapper and Embedded filter methods in various fraud dataset applications based on machine learning classifications. In addition, we investigate four different comparison methods and four traditionally established classifiers tested on a variety of metrics.

7.5.3. Applications in Fraud Detection

Fraudulent activities have become a major concern for many businesses that are faced with a wide range of fraudulent behaviors in various forms, such as disputed insurance claims, faked trading volumes, identity theft, etc. It is very difficult for an enterprise to manually inspect all the potentially fraudulent activities covertly occurring. Computer-based fraud detection or forensic analysis is the process of automatically identifying illegal behaviors from the many observations collected from different operational systems. The proactive nature of fraud detection is a great advantage since fraud can inflict enormous damage to both individuals and enterprises when left unchecked. Forensic analysis can be seen as the reactive approach to fraud detection, detecting fraud and what caused it to occur after damages have already been made. Traditional fraud

detection techniques, such as rules-based detection and anomaly detection in the financial sector, have had limited success due to the highly evolving nature of fraud and the many different forms it can take. Alternatively, classification, clustering, association rule mining, pattern recognition, and artificial intelligence and machine learning techniques can be applied to all fields of fraud detection. The most popular approach to applying machine learning in fraud detection is to use a mixture of unsupervised learning for profiling the normal behavior and supervised learning for training fraud detection rules or models, which can help identify whether a new case of transaction is indeed fraudulent. This is especially beneficial in situations where the nature of fraud may change rapidly.

7.6. Data Collection and Preparation

Fraud is one of the most common criminal activities with a very high impact on economic and social life. These usual activities are financed with many different resources, and the proceeds from fraud may be used to fund criminal and/or terrorist organizations. Detection and deterrence of fraud are essential for organizations to maintain their economic and social stability. Many approaches have been developed to

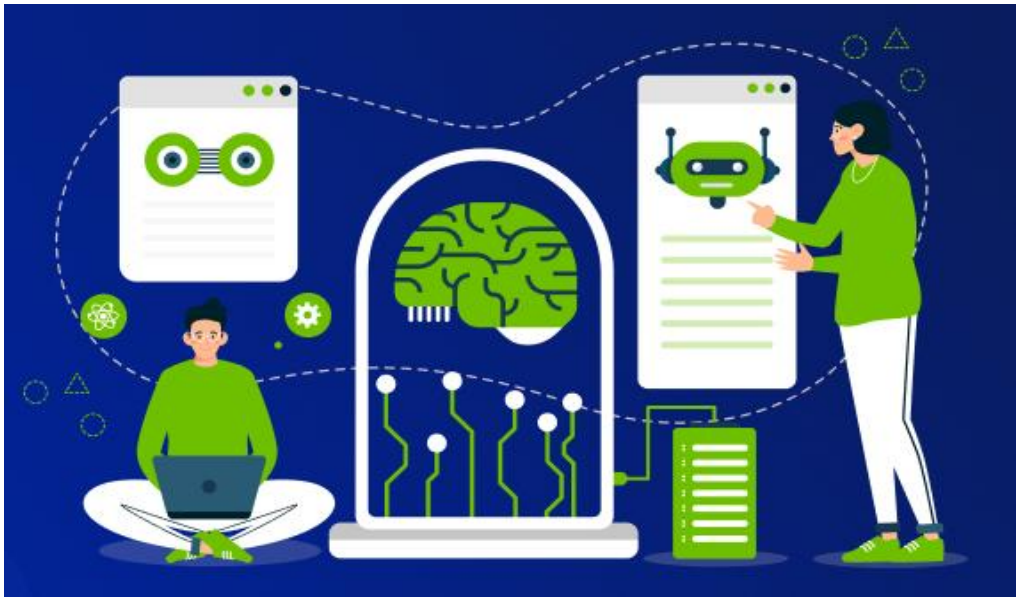

Fig 7 . 2 : ML for Fraud Detection

detect this type of crime, which include statistics, business rules, and data mining, among others. In this work, we describe a fraud detection tool using data mining (clustering and classification with neural networks). The results from the fraud classifier are compared

with results from a real business case. An economic gain in the classification part using different sampling techniques is obtained.

In this section, we describe how data was collected, prepared, and pre-processed before it could be used. The database was composed of different data sources, including operational, transaction, customer data, and known fraud and known non-fraud samples. The fraud and non-fraud samples included different types of fraud, both from external and internal crime. Another important source of data was the scoring results used by the banks to manage their clients. This data, associated with the business rules, helped to identify the features that best contribute to the detection of fraud. In addition to the data pipeline, this involved merging due to the initial and consecutive fraud events. All these steps are part of the data preparation process.

7.6.1. Sources of Data

Fraudulent financial activities originate from two sources: within the organization itself or through external sources. Internal fraud originates within the organization and is usually instigated by employees. External fraud, on the other hand, originates outside of the organization, often involving individuals engaged in consumer financial activities. Fraudulent activities often leave a pattern of behaviors and interactions as they unfold. This section discusses some of the sources where such patterns could be found. If firms can have copies of such patterns, they can employ machine learning tools to enhance these patterns and find other sources of fraud that are not common. For example, telephone numbers that have a large number of long-distance calls may be linked to stolen credit card fraud. These patterns should make it easier for investigators to find and examine the suspicious activities. Telephone numbers that have a large number of short-duration calls can be linked to auto repair fraud. These patterns should make it easier for investigators to find and examine the suspicious behaviors. These patterns will be easy to understand and easy to explain.

There are known patterns of fraudulent activities as they unfold. Credit card fraud often involves cards from certain areas. Therefore, the geographic patterns and the area codes to which the cards are calling should be analyzed. Systems that resemble telecommunications fraud often involve different marks of modems. The manufacturer can identify the marks of the modem that is involved. The boxes that resemble telecommunications fraud often make calls to the same number. If the number is called during the middle of the night, it could suggest that they are using the phone lines without the expectation of being caught. These address patterns should be investigated to evaluate the probability that the owner of the phone number is aware that the number is being used for fraudulent reasons. The use of area codes and exchange codes should be investigated for fraudulent reasons, as should logs of the volume of calls and the times

when the fraud may occur. If the same phone number is calling two locations near each other, tools can be used to evaluate if the calls are being forwarded elsewhere. Finally, traffic analysis tools can be used to evaluate the pattern of calls that the telephone number is making. These types of tools are being utilized with good success. The sources of the data can include: own and third-party data systems, digital photos, telephone metadata, video/audio data, data for mobile applications, traffic violation data, Point of Sale and Retail Transaction datasets, Travel/Admission transaction datasets, ownership/supplier manufacturing transaction data, and employee access level data.

7.6.2. Data Cleaning Techniques

Data samples often need to be cleaned before machine learning and some conventional mathematical pattern recognition can be implemented. Data cleaning deals with the issues of inaccurate, incomplete, or inconsistent data. While the distorted input and output variables of a model can introduce inaccuracy and noise, elsewhere the whole dataset, data elements, and data values can cause model parameters to be defective as a result of inappropriate data quality. Smart, rigorous, and statistically proven data cleaning techniques are required for different types of data. Many data quality tools have been developed, but they often require human intervention and are not widely used in a fully automated manner. On the other hand, fully automated data mining can be used to perform exploratory data analysis at the data cleaning stage. Consistent and integrated data cleaning can be used by combining variable selection with transformation and the handling of incorrect training data as missing attribute values; then missing data can be indicated by a specific value, and affected corresponding data can be transformed correctly. The missing attribute values can be predicted by providing standard deviation variance mean for distribution from prediction models. The constructed statistical datasets are also available for the retrieval of these parameters by the training and testing phases of datasets to create an artificial statistic, which can be rearranged in terms of fitting statistical parameters. The statistical and learning models do generate noise and gross error while data collection and transmission imprecision or approximation can also contribute to data error due to the number of variables related to prediction models.

7.6.3. Data Transformation for ML Models

In section 6.4, we introduce how we label the data, then we explain our transactions and merchants features to prepare for our first set of models. Machine learning models require data in a certain format and structure. The input transaction data needs to be transformed into fixed-size vectors with relevant information for the machine learning algorithms. This data transformation has three core objectives: 1) model input format, 2)

feature processing, and 3) modeling information encoding. There are primarily three categories of features we need to take care of: general transaction features, time-based transaction features, and merchant level features. After extracting these features and labels, we need to assemble the input training dataset for model training; time can be padded for fixed-size input, and during model building, missing features can be omitted or replaced.

Another important thing about these transformations is to handle statistical information separately. For example, are the different count statistics likely costly to pay? In terms of feature processing, many machine learning models work efficiently with normalized features. Categorical features can also be encoded into a more machine learning model-friendly format. Even though most models are capable of dealing with missing input signals, they tend to gain benefits when their inputs are adjusted. Model encoding refers to how to encode modeling information in the provided data structure. For example, if a model can recognize the transaction time for each machine learning model, the transaction time does not need to be encoded, but when it does not recognize it, we may try to evaluate encoding methods. Data integration and labeling entity recognition are other preprocessing techniques that are used to prepare the data for a machine learning model.

7.7. Model Training and Evaluation

Due to the high computing costs during the model training phase, we decided to reduce the set of base algorithms. Prior to this, we filtered those that presented the most difference in performance. This process resulted in the choice of four machine learning methods: Binary Boost, Model Tree, Random Forest, and RTPM PPA. We used the software for this test with the following configuration: 10-fold cross-validation and automatic data partitioning. We note that only those algorithms that were relevant to the application set were available. After this evaluation, the Random Forest method turned out to be the most efficient for this classification. It was then configured for application in three approaches.

In the first approach, an unmodified dataset was used as the input. In the second approach, the algorithm was applied in an attempt to correct the imbalance of the data coming from the previous point. Finally, in the third approach, an active learning methodology was applied, resulting in a new balance validated by humans from the fraud team of the company studied. Unlike other forms of learning, this technique is capable of performing individual case analysis to prioritize unfeasible, unfounded, and misuse-ridden contracts. Active learning presents definitions for its practical functions to assess the benefits of the large set of data points and to label the contrast between them by sharing the richness and reliability of the labeler's expertise. The cross-validation test

was performed before the production cost. The output of this last stage, in its three forms, was considered robust. With acceptable results in both balances, we suggested the company choose the approach that best fits their needs for three distinct contexts.

7.7.1. Training Machine Learning Models

Our method builds on ideas from prior applications of machine learning to financial and insurance fraud. Our main improvements are a novel statistical ensemble of several machine learning models, the use of a machine learning strategy identification pipeline, and the incorporation of several deep learning models. The machine learning package is also crucial to our method achieving peak out-of-sample performance. The broader impact of this work is that it suggests a suite of ideas to improve the ability of organizations to swiftly detect fraud in incoming streams of transactions, especially true in financial services applications.

To train our machine learning models, we collected a labeled training data set with specific contrived machine learning model inputs for both phishing and transaction manipulation contracting offers to professionalize crime. To collect material for the model inputs, we conducted exploratory data analysis preparation steps. Afterwards, we trained the models on more than 11,000 historical phishing and significant manipulation attempts that were used by adversaries, teams of researchers from cyber crime fighting companies, inside threat actors employed by various financial institutions to sabotage trading systems, or researchers conducting experiments. After transforming the published threat intelligence information, we verified that the rules and weights used to find each attempt matched the threat intelligence.

7.7.2. Evaluation Metrics

In fraud detection and in all fields related to anomaly detection, the type of data is usually highly imbalanced. In other words, the number of regular data (or non-fraudulent transactions) is usually very large compared to the number of irregular data, which are the cases we are really interested in. This makes a model fall into a pitfall - a model that predicts all the non-fraudulent transactions while classifying a few fraudulent transactions. The outputs will look very impressive, but if the outputs of a hundred fraudulent transactions are missing, for example, our company and customers will suffer serious consequences.

We have to find this point of balance in the procedure which describes the trade-off between precision and recall. A model may have a high precision but a terrifying recall, and an extra 0.1 in recall can cost a lot. The F1 score really helps clarify the situation. It

is the harmonic mean of precision and recall. In less technical terms, I can imagine the two goals when I find an error - preventing a non-fraudulent case from being detected and preventing the fraudulent transaction from slipping through - as two eggs. And the eggs are in the fridge since I found the error. I can only take out the eggs one by one. The recall is equivalent to if I am able to take them all out, but the precision is determined by how I take them. If I hold a cuddly panda in front of me, I can retrieve the eggs really slowly to make sure the second egg will suffer no harm, but then perhaps I'd be able to take out the first egg only. Therefore, we always need a tool like the F1 score to help us choose the best precision-recall trade-off.

7.7.3. Cross-Validation Techniques

When working with classification problems, cross-validation is essential to modeling, as it helps ensure the good generalizability of the model by demonstrating that it is not significantly data-specific. During cross-validation, training and validation are performed on different subsets of the data, also referred to by the terms training and testing. In k-fold cross-validation, one of the most common methods, the data is randomly divided into k subsamples, with each subsample retaining equal proportions of the target outcome. The holdout method, a special case of k-fold cross-validation, also makes use of two subsamples.

The decision of which cross-validation method to use can be important and depend on the application at hand. Leave-one-out cross-validation involves tasks being performed k times with just one object placed into the validation set each time. K-fold cross-validation is often preferred since it provides the average performance of the model while running k times faster than the leave-one-out method. In other cases, leave-one-out may be used instead of k-fold if the goal is to identify objects that are incorrectly predicted. In our case, ten-fold cross-validation was selected as a conservative compromise between accuracy and computational cost. The process involves randomizing the data rows and separating them into k equal subsets, with each fold containing m = N/k observations. In ten-fold cross-validation, one fold is chosen as a validation set while the other k – 1 folds are used as training. This process is repeated k times in such a way that each of the k folds is used once as the validation data. Once the k iterations are complete, the cross-validation estimator is calculated as the average or the maximum of the k computed estimates.

7.8. Deployment of Fraud Detection Systems

Having a model that can identify fraud is very important, but it is also very important to have the right user's company, department, or team making decisions based on this

model's outputs. A misuse of this model has huge negative impacts. We proposed an easy-to-implement threshold. Transactions flagged by the model where the difference between early-stage risk and book risk was bigger than our threshold. We established and validated this threshold using a book risk resulting distribution that the company determined to be exhaustive as a reference model. Our threshold allowed us to clearly decrease false positives and ensured that all riskier situations were indeed investigated live and retrospectively. We were able to claim that every single EUR 250k of transacted volume measured was analyzed. The deployment of a machine and pattern recognition-based fraud prevention model is not a technical issue per se. However, the misuse of this model may have catastrophic impacts in the corporate world. A model has no moral values but may have strong ethical impacts, especially when implemented in the banking system where the core of its responses should be driven by carefully established rules and the explicitly determined strategy. We work with R squared near 0.5 for both fraud and quality output. We work with high-powered microlots in a similar way. Our work anticipates about three years of similar works conceived without having access to our insights. A team that performs better than our model has 200 very experienced people, where this team invested decades in mastering and developing rules highly respected inside the company. The model is an excellent support for their analysis, better identifying almost two signals per hour: the need to improve a rule and the transaction deemed more risky.

7.8.1. Integration with Existing Systems

Data mining and machine learning tasks offer the potential to successfully model and detect many types of fraud. Yet organizations concerned about fraud are typically already spending significant amounts on systems and technologies designed to detect or prevent fraud from occurring. As such, there exists an inherent resistance to altering such successful apparatus, even when, over time, its effectiveness appears to be waning. Therefore, however exciting from a research perspective the potential use of intelligent systems for fraud detection is, their relevance will in the short to medium term critically depend on whether they can be meaningfully and seamlessly integrated into existing systems. System integration involves not only the exportability of the knowledge derived from fraud detection models arising from data mining and machine learning tasks, but also the deployability of system outputs and the integration of the new system within business or organizational procedures.

It is widely acknowledged that real-time detection is particularly essential in the combat against fraud. Intelligent fraud detection systems are often explicit knowledge-based systems or rules engines that are able to instantly evaluate extremely complex interactions between fraudulent and non-fraudulent activities or transactions. In contrast,

offline fraud detection systems typically act as filters or preprocessors fed with the produced models or the relevant scoring or ranking.

7.8.2. Real-Time Processing Capabilities

Machine learning technologies are data munchers and are becoming more efficient as data is becoming bigger, wider, and deeper. However, the concern is that machine learning service providers are often seen as black boxes. When we talk about detecting fraud and which methods or techniques to use, professionalism and free sharing are important. Ethical debates are not happening fast enough. Our choice may be to manage and control damages or to rebuild trust if and when the table turns.

Database management systems started to handle indexes in parallel when the business cycle was getting shorter and shorter for more accumulated data. Rule-based data processing systems apply preconditions or constraints to control the ever-growing database so that rules are managed in their concerns and compliance. Massive datasets nowadays are handled by non-SQL based big data platforms or search engines with map reduce, column or wide-column store, in-database map reduce, etc. Reinforcement learning is giving machine learning another round-up run with its capability of real-time processing. With pretrained models or trained models, businesses, experts, and researchers would have the advantage to minimize the evaluation cost and time. In terms of machine learning with models, reinforcement learning doesn't change this part of the process. What reinforcement learning emphasizes is the ability to receive feedback in a dynamic way, updating models immediately after the new observation.

7.8.3. Monitoring and Maintenance

Once we decide on a particular detection method and apply it, it is necessary to provide a monitoring system for keeping the method's effectiveness intact. For example, we may use the method for fraud detection purposes, a misuse detection system. As the system is used, it becomes necessary for periodical reviews to be performed regarding its outcomes and the method's operating environment. Periodical reviews will give us the chance to gather up-to-date information about the observed risk factors at a company. In addition, there are also changes that may happen in the company, which important risk factors may influence. These company changes have to be observed continuously in order to keep them from interfering with the misuse detection method.

Moreover, improvement in results over time enhances the existing method. Evaluative efforts shall document the method's increase in proficiency and pave the way for that. The method's increase in operating skill can result in changes that improve the results of

the system. If we observe numerous false alerts along our operating system from the viewpoint of low production cost and time, this result becomes vulnerable. We have a chance to write analytical procedures, which are very time-consuming and expensive. We have performed manual procedures for investigation to obtain only accurate results for alert data. These analytical procedures have been applied only to the alerts whose accuracy is confirmed by a misuse. We do not want to do this.

7.9. Conclusion

In this article, we have shown how machine learning can increase the ability of insurance companies to detect fraud. Lawsuits issued because of insurance fraud have generated significant losses. Consequently, the insurance industry is currently striving for a trustworthy fraud detection method. Inspired by cases detected by persuasion, in which inspectors detect fraud by simply tracking down patterns, there is transparency in the decision process that serves to enhance detection and bring confidence to both parties in the analysis. We note that there is a place for both expert systems and discovery systems in insurance fraud analysis. Classic expert systems are appropriate for recognition-based systems, in which incoming claims are checked against specific patterns of interest or against explicit rules. Classic discovery systems are ideal for novel detection-based systems, in which incoming claims are screened for statistical patterns and behavior-based anomalies.

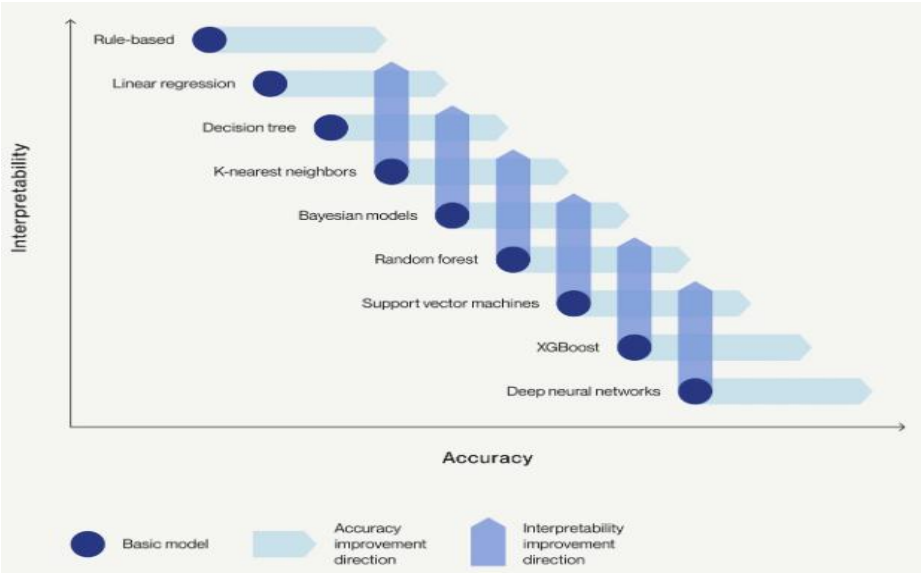

Fig 7 . 3 : Data and Machine Learning in Financial Fraud Prevention

Our analysis has demonstrated that using a neural network can result in an improvement in the performance of the individual fraudulent claimant classifier. By analyzing model performance, we note that some operation points may bring clear financial implications. Moreover, it is worth noting that although our results concern the Brazilian market and are associated with Brazilian insurance fraud, this synthetic detection approach can be used with insurance data from other countries. The tools used in this study are generic, and all the results can be replicated in other countries with slight modifications. Our results are important for the insurance industry since the development of new technologies for fraud detection is one of the main technical tasks for actuaries nowadays. Finally, rule changes can affect historically fraudulent patterns. Our model is open and understandable, and a fraud specialist with domain expertise can incorporate new theories into the detection framework at a small cost.

7.9.1. Key Takeaways and Future Directions

In our discussion of using machine learning and pattern recognition to enhance fraud detection, certain themes arose which are useful to review in a summary. One theme was the superiority of machine learning methods of clustering in spotting potential patterns of fraud across industries. That is not true of individual machine learning models or supervised methods like logistic regression. Another theme was the importance of clustering methods in identifying complex cases of fraud with a pattern recognition basis. Supervised methods have trouble with such cases. While both clustering methods and pattern recognition have a role to play in fraud detection, the fact that a certain estimated model with strong significance has achieved a respected p-value in a behavioral model does not guarantee the company assurance that the fraud will be discovered in the model. But when the models have a behavioral pattern, the likelihood for discovery heightens.

Using pattern recognition to analyze behavioral data can be most helpful in identifying potential financial fraud as it occurs. This discussion was motivated by the widespread deception in financial data and the suspicious circumstances that are often present when fraud occurs. Future research in this area should add to our knowledge of financial statement fraud. Case studies link pattern recognition to situations where fraud is suspected or should be suspected. After all, the fact that the financial numbers are non-behavioral indicates that potential misconduct is afoot.

References

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. Artificial Intelligence Review, 34(1), 1–14. https://doi.org/10.1007/s10462-010-9273-1

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud Detection System: A Survey. Journal of Network and Computer Applications, 68, 90–113. https://doi.org/10.1016/j.jnca.2016.04.007

West, J., & Bhattacharya, M. (2016). Intelligent Financial Fraud Detection: A Comprehensive Review. Computers & Security, 57, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature Engineering Strategies for Credit Card Fraud Detection. Expert Systems with Applications, 51, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection. Information Sciences, 479, 448–455. https://doi.org/10.1016/j.ins.2018.02.060