ACFE™
Association of Certified Fraud Examiners
Indonesia Chapter #111

Asia Pacific Fraud Journal
E-ISSN: 2502-695X, ISSN: 2502-8731
Volume 6, Issue 2 (July-December) 2021
*Asia Pacific Fraud Journal now has been accredited "**SINTA 3**" by Ministry of Research and Technology of The Republic of Indonesia (No SK. 85/M/KPT/2020).*
Availaeble online at: http://apfjournal.or.id/index.php/apf

Machine Learning Algorithms in Fraud Detection: Case Study on Retail Consumer Financing Company

✉Nadya Intan Mustika, Bagus Nenda, Dona Ramadhan
PT. Adira Dinamika Multi Finance, Tbk, Indonesa

ARTICLE INFORMATION

Article History:
Received August 02. 2020
Revised November 07, 2020
Accepted December 1, 2021

DOI:
10.21532/apfjournal.v6i2.216

ABTRACT

This study aims to implement a machine learning algorithm in detecting fraud based on historical data set in a retail consumer financing company. The outcome of machine learning is used as samples for the fraud detection team. Data analysis is performed through data processing, feature selection, hold-on methods, and accuracy testing. There are five machine learning methods applied in this study: Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM). Historical data are divided into two groups: training data and test data. The results show that the Random Forest algorithm has the highest accuracy with a training score of 0.994999 and a test score of 0.745437. This means that the Random Forest algorithm is the most accurate method for detecting fraud. Further research is suggested to add more predictor variables to increase the accuracy value and apply this method to different financial institutions and different industries.

Keyword: Fraud Detection, Machine Learning, Predictor Variable.

1. INTRODUCTION

The era of the industrial revolution 4.0 currently is characterized by the development of digital technology. Digitalization is a necessity that every company must adopt to compete in today's business. Technological developments have been used by all people and stored in the form of data. Therefore, the industrial era 4.0 provides abundant data in various forms.

This abundant data can be used by companies to make business innovations. Likewise, fraud risk management can be helped by analyzing the available data. The urgency of data analysis is increasing, especially during the pandemic, when field activities are limited. One of those affected by the limitation of activities due to the pandemic is the internal auditor team who has to carry out a remote audit. One method that can be used in this remote audit is to perform data analysis.

Fraud is an act that violates the rules committed by internal and / or external parties of a company with the aim of obtaining personal or group benefits. These fraudulent acts are not directly detrimental to the company (Bank Indonesia, 2011). One of the industries that are at risk of experiencing fraud is the financial industry. This risk is getting higher, especially for financial institutions that focus on retail financing, due to the large number of transactions. Otoritas Jasa Keuangan (OJK) has developed a fraud risk management strategy includes several stages such as prevention, detection, investigation, reporting, sanction, monitoring, evaluation, and follow-up.

✉Corresponding author :
Email: nadya1403@gmail.com

There are many methods and strategies that can be used in the fraud detection process. Basically, every fraud incident has certain characteristics, both environmental conditions and a person's behavior. Therefore, fraud can be detected if there is a good understanding of the characteristics of each fraud condition that may arise in the company. Classification analysis can be used to detect fraud based on the characteristics that arise. The results of this classification can be used as a sample in carrying out the process of observation or field inspection to reveal the presence or absence of fraud in these conditions.

Given that there is a large number of transactions in the retail financing business, which means that the availability of data is also abundant, the fraud detection process can use this data to classify a transaction whether it meets the conditions as fraud or not. It is very difficult to perform large amounts of data analysis manually. Therefore, the data analysis process must be assisted by smart or intelligent machine technology to improve human capabilities (Nayak & Dutta, 2017). The intelligence demonstrated by machines is usually referred to as artificial intelligence, which is the study of computer science. Artificial intelligence is broadly used to solve different issues such as business problems, robotics, natural language, mathematics, games, perception, medical diagnosis, engineering, financial analysis, scientific analysis, and reasoning (Rahardja et al., 2017; Russell & Norvig, 2016).

One form of artificial intelligence is the machine learning method. This method (machine learning) can be defined as computer applications and mathematical algorithms that are adopted by means of learning derived from data and producing predictions in the future (Goldberg & Holland, 1988). Machine Learning can be applied in the fraud detection process by classifying whether a transaction is categorized as fraud or not.

Shirgave (2019), uses machine learning to conduct fraud detection in the credit card business. This is of course different from business patterns and fraud patterns in retail financing companies. The use of Machine Learning for fraud detection in retail financing companies has not been widely used, because it is quite difficult to find patterns of fraud.

This study aims to classify fraud incidents based on historical data. This research is conducted through case studies on retail consumer financing transactions. The research method is to find a Machine Learning algorithm with the highest level of accuracy. Machine Learning analysis output will be used as a parameter to determine field inspection samples in order to find field facts about the incident of fraud. This is quite important to do considering the large volume of transactions.

2. LITERATURE REVIEW AND HYPOTHESIS

Definition of fraud

Tuanakotta (2013), stated that fraud is an illegal display described as misleading, disguising or destroying trust.The purpose of this activity is to obtain cash, abundance or administrations, to keep away from installment or administrative losses, or to obtain benefits for private issues. ACFE defines fraud as a deliberate act aimed at persuading another person to act that is detrimental to that person.

Fraud is a part of operational risk. In the scientific categorization, operational risk of fraud can be categorized into Internal Fraud and External Fraud (BCBS, 2002). Ramadan (2020), explained that this distinction is based on the fraudster. Internal parties such as employees, which is fraud committed by personwithin the organization who abuses his/her power or assets for individual benefit (ACFE, 2020). Meanwhile, external partiessuch as customers and business partners, and a combination of the two parties.

Remote Audit

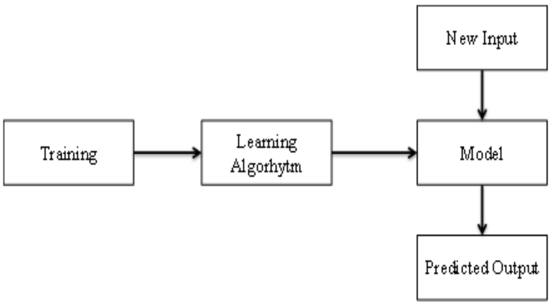
Remote audit is an innovation and business transformation that has been going on for several years. The development of technology is one of the factors that can support the remote audit process. There are several latest technologies that can be utilized in the remote audit process, such as the use of live video streaming, drones, and artificial intelligence. The implementation of technological developments for remote audit elements, such as planning, document checking, field inspection, remote interviews, and closed meetings can be done using video teleconferences and other available technology platforms such as Microsoft Teams, Skype, and Zoom (Litzenberg& Ramirez, 2020).

Remote audit is a solution to new challenges, especially when the inspection team has limited space due to the pandemic. There are pros and cons from various parties regarding this remote audit process. Several positive things were found when implementing the remote audit process, namely reducing travel costs, expanding the use of specialists, increasing the use of existing technology to strengthen documentation and reporting, mitigating audit burdens on operational facilities, and improving organization and confirmation of required documentation (Litzenberg& Ramirez, 2020). Apart from simplifying the audit process, remote auditing has limitations that also need attention. Lack of direct personal interaction opens up opportunities for fraud. This can increase the chances of submitting manipulated documents and omitting relevant information.

Machine Learning

Machine learning described as computer applications and mathematical algorithms that are adopted by means of learning derived from historical data and producing predictions in the future (Goldberg & Holland, 1988). Learning process is an attempt to acquire intelligence that is taken through two stages: training and testing (Huang, Zhu, & Siew, 2006). Here is the Machine Learning Algorithm:

Figure 1. **Machine Learning Algorithms**



Source: Processed data

Recent research has revealed that machine learning methods are divided into Supervised, Unsupervised, and Reinforcement Learning (Somvanshi& Chavan, 2016). When historical data with data classifications are available, direct machine learning output classifications can be determined. In this case, the output classification is divided into Fraud and Non Fraud categories. This method of determining fraud can be solved using the Supervised Learning method.

Supervised Learning

The supervised learning method is based on a data set collection with label. This collection of data set is used to summarize the characteristics of the distribution of attitude dimensions in each type of application so as to form an attitude model from the data (Amei et al, 2011). Moreover, supervised learning will use regression and classificationmethod. Brownlee (2016) revealed that the regression problem is when the output variable is aexact value, such as price or sales number. At the same time,the classification problem is when the output variable is categorical, such as a one or zero, yes or no..Supervised learning has several popular algorithms such as Logistic Regression,Back-propagation, Neural Network, Decision Tree,Random Forest, Naive Bayesian, Rocchio Method, Linear Regression,k-Nearest Neighbor, and Support Vector Machines (SVM).

In this study, five classification algorithms will be compared: K-Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine. K-Nearest Neighbor is a technique of object classification based on the class that closest to the object. Decision Tree Algorithm can be used to classify a sample of data, whose class is not yet known, into existing classes. Attribute data must be categorical data. If it is continuous, the attribute must be discretized first. Similar to the Decision Tree, the Random Forest algorithm is also used to classify a sample. Random Forest is carried out by merging trees by conducting training from the sample data owned. Random Forest uses a Decision Tree to carry out the selection process. One of the classification methods widely developed today is the Support Vector Machine (SVM). Prasetyo (2014) stated that the basic concept of this method is to maximize the hyperplane boundary that separates a data set. Logistic Regression is a type of regression that analyze one or more independent variables with the dependent variable expressed in categorical scale, so that the output will be in the form of category such as 0 (null) or 1 (one).

3. METHODS

Data

The data set t used in this research are historical data for the period of 2015-2019 regarding the characteristics and levels of fraud in retail financing company X. The data used consist of 26 variables, where all variables are in categorical form. The number of objects of observation is 46,536 transactions in retail financing company X. This study is assisted by the Python Jupyter Notebook software.

Research Methodology

The variables used in this study consist of 25 predictor variables which are the features that arise and 1 response variable which is the occurrence of fraud. The predictor variables can be grouped into two: the customer profile and the financing profile. Details of the variables used in this study can be seen in Table 1.

The analysis steps are as follows:

a. Collecting historical data on the characteristics and data of fraud incidents.
b. Preprocessing fraud incident data.
c. Dividing the data into training data and testing data.
d. Classifying using training data with the K-Nearest Neighbors (KNN) algorithm, Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine (SVM).
e. Performing analysis and evaluation of the model.
f. Drawing conclusions and giving suggestions.
g. Making predictions with new input data and creating a Confusion Matrix.

4. RESULT AND DISCUSSION

Preprocessing Data

The first step that must be taken before conducting further analysis is conducting data preprocessing. At this stage the steps taken are the detection of the Missing Value process and the Feature Selection

Missing Value Detection

At the data preprocessing stage, the first thing to do is to detect the Missing Value. The following are the results of the detection of missing values for each research variable.

Table 1. **Research Variable**

No.	Variable	Group	Remark	Measurement Scale
1	Y	Output	Fraud Classification	Ordinal/Categoric
2	X1	Customer Profile	Cust_Profile 1	Ordinal/Categoric
3	X2		Cust_Profile 2	Ordinal/Categoric
4	X3		Cust_Profile 3	Ordinal/Categoric
5	X4		Cust_Profile 4	Ordinal/Categoric
6	X5		Cust_Profile 5	Ordinal/Categoric
7	X6		Cust_Profile 6	Ordinal/Categoric
8	X7		Cust_Profile 7	Ordinal/Categoric
9	X8		Cust_Profile 8	Ordinal/Categoric
10	X9		Cust_Profile 9	Ordinal/Categoric
11	X10		Cust_Profile 10	Ordinal/Categoric
12	X11		Cust_Profile 11	Ordinal/Categoric
13	X12		Cust_Profile 12	Ordinal/Categoric
14	X13		Cust_Profile 13	Ordinal/Categoric
15	X14	Financing Profile	Financing_Profile 1	Ordinal/Categoric
16	X15		Financing_Profile 2	Ordinal/Categoric
17	X16		Financing_Profile 3	Ordinal/Categoric
18	X17		Financing_Profile 4	Ordinal/Categoric
19	X18		Financing_Profile 5	Ordinal/Categoric
20	X19		Financing_Profile 6	Ordinal/Categoric
21	X20		Financing_Profile 7	Ordinal/Categoric
22	X21		Financing_Profile 8	Ordinal/Categoric
23	X22		Financing_Profile 9	Ordinal/Categoric
24	X23		Financing_Profile 10	Ordinal/Categoric
25	X24		Financing_Profile 11	Ordinal/Categoric
26	X25		Financing_Profile 12	Ordinal/Categoric

Source: Processed data

Feature Selection

After the Missing Value problem is resolved, the next step is to carry out the Feature Selection process. Feature Selection functions to select the variables that are not good for use and select only a few variables that are best used by reducing the scores and sorting based on the scores obtained. The following are the results obtained from the Feature Selection process (Table 3).

After selecting the predictor variable, 19 (nineteen) variables or features are selected to be continued in the next analysis stage.

Hold-Out Method

Based on the results of the Feature Selection, it is found the selected variables or features that will be used in the analysis. The next step is to do the Hold-out Method. Hold-out method is a method that splits data into two types that are training data and test data. The training data is the data used to view the data pattern to be tested, and the test data is the data used to test whether the generated model can predict the data to be tested. In this r, the portion of the testing data is 20% and the portion of the training data is 80%.

Table 2. **Missing Value Detection**

No	Remark	Number of Missing Value
1	Fraud Classification	0
2	Cust_Profile 1	0
3	Cust_Profile 2	248
4	Cust_Profile 3	248
5	Cust_Profile 4	249
6	Cust_Profile 5	4
7	Cust_Profile 6	252
8	Cust_Profile 7	0
9	Cust_Profile 8	3
10	Cust_Profile 9	0
11	Cust_Profile 10	791
12	Cust_Profile 11	4
13	Cust_Profile 12	0
14	Cust_Profile 13	541
15	Financing_Profile 1	0
16	Financing_Profile 2	0
17	Financing_Profile 3	539
18	Financing_Profile 4	0
19	Financing_Profile 5	0
20	Financing_Profile 6	0
21	Financing_Profile 7	0
22	Financing_Profile 8	0
23	Financing_Profile 9	0
24	Financing_Profile 10	0
25	Financing_Profile 11	0
26	Financing_Profile 12	0

Source: Processed data

Model Analysis

After carrying out the data preprocessing stage, the next step is to conduct an analysis to determine the accuracy value using five classification methods: K-Nearest Neighbors (KNN) classification, Decision Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression. To determine which method is more appropriate in the classification analysis, it is done by comparing the accuracy value of each classification method. The higher the accuracy value, the better the resulting classification accuracy, and vice versa. The following is a comparison of the accuracy values of the five methods used.

Based on the resulting accuracy value, it is found that the Random Forest Method is a more appropriate method of classifying fraud data based on data on retail financing company X with the highest accuracy value.

Table 3. **Feature Selection**

No.	Feature	Score
1	*Feature_13*	0.074535
2	*Feature_12*	0.073040
3	*Feature_24*	0.072381
4	*Feature_4*	0.064019
5	*Feature_10*	0.057947
6	*Feature_1*	0.053677
7	*Feature_20*	0.052614
8	*Feature_23*	0.051896
9	*Feature_17*	0.051056
10	*Feature_19*	0.050507
11	*Feature_6*	0.047823
12	*Feature_11*	0.045823
13	*Feature_8*	0.044542
14	*Feature_21*	0.040400
15	*Feature_18*	0.036751
16	*Feature_3*	0.036085
17	*Feature_2*	0.033492
18	*Feature_9*	0.032635
19	*Feature_22*	0.027306

Source: Processed data
Note: Feature is a predictor variable

Table 4. **Accuracy Value Methods**

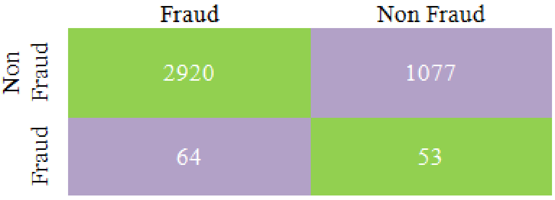
No	Method	Train score	Test score
1	K-Nearest Neighbors (KNN)	0.763671	0.738223
2	Random Forest	0.994999	0.745437
3	Decision Tree	0.994999	0.639633
4	Support Vector Machine (SVM)	0.740743	0.738113
5	Logistic Regression	0.740743	0.738113

Source: Processed data

Validation and Confusion Matrix

After getting a more precise method, the next step is to do the validation process using new input data. The new input data used at this stage is data regarding the characteristics and levels of fraud in retail financing company X in 2020. This validation process is carried out by predicting new input data using the Machine Learning Algorithm with the method that has the highest accuracy value, namely the Random Forest Method. The following is the resulting Confusion Matrix.

Figure 2. **Confusion Matrix**



	Fraud	Non Fraud
Non Fraud	2920	1077
Fraud	64	53

Source: Processed data

After obtaining the Confusion Matrix as in Figure 4.1, the values for accuracy, precision and recall can be obtained, as seen in Table 4.4.

Table 5. **Accuracy, Precision, and Recall Values**

	Value
Accuracy	0.722654351
Precision	0.730547911
Recall	0.978552279

Source: Processed data

As shown on the table 4.4, the accuracy value generated at the validation stage using new input data is 0.722654351, which means that predictions using the Random Forest Method are quite accurate for the data held.

5. CONCLUSION

Based on the research steps that have been carried out, the results show that the Random Forest algorithm has the highest train score and test score, 0.994999 and 0.745437, which means that the Random Forest algorithm is a more precise algorithm in detecting the level of fraud based on the classification of characteristics that arise from each transaction. In addition, in the validation stage, an accuracy value of 0.722654351 is obtained which proves that the Random Forest algorithm has a fairly good accuracy in predicting data regarding the characteristics and levels of fraud in 2020 at retail financing company X.

It is suggested that further research add more predictor variables that have an influence on the classification of fraud incidents with the aim of increasing the accuracy value of the applied Machine Learning algorithm and apply this method to different financial institutions and different industries.

REFERENCE

Association of Certified Fraud Examiners (ACFE). (2020) Report to the Nation.

Amei, W., Huailin, D., Qingfeng, W., & Ling, L. (2011). A survey of application-level protocol identification based on machine learning. Presented at International Conference on Information Management, Innovation Management and Industrial Engineering.

Amrizal. (2004). Pencegahan dan Pendeteksian Kecurangan oleh Internal Auditor. Jakarta.

A. Said., & V. Torra (eds.). (2019). Data science in practice, studies in big data 46. Springer International Publishing AG, part of Springer Nature 2019.

Bank Indonesia. (2011). Surat Edaran Bank Indonesia No.13/28/DPNP tanggal 9 Desember 2011 perihalPenerapan Strategi Anti Fraudbagi Bank Umum.

Basel Committee on Banking Supervision (BCBS). (2002). Operational Risk Data Collection Exercise.

Brownlee, J. (2016). Master Machine Learning Algorithms: discover how they work and implement them from scratch. Jason Brownlee.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. Machine Learning, 3(2), 95–99.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). "Extreme learning machine: theory and applications." Neurocomputing, 70(1–3), 489–501.

Litzenberg, Roy., & Ramirez, C.F. (2020). Proses audit jarak jauhselama dan setelah covid-19: implikasi jangka pendek dan panjang. The Institute of Internal Auditors.

Nayak, A., & Dutta, K. (2017). Impacts of Machine Learning and Artificial Intelligence on Mankind. Dipresentasikan dalam International Conference on Intelligent Computing and Control (I2C2), 1–3.

Prasetyo. (2014). Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. Penerbit Andi.

Peraturan Otoritas Jasa Keuangan Nomor 35/POJK.05/2018 tentang Penyelenggaraan Usaha Perusahaan Pembiayaan.

Ramadhan, Dona. (2020). Root Cause Analysis Using Fraud Pentagon Theory Approach (A conceptual framework)." Asia Pacific Fraud Journal. Vol 5 No 1 2020.

Sudaryono, & Rahardja, Untung & Roihan, Ahmad. (2017). Design of Business Intelligence in Learning Systems Using iLearning Media. Universal Journal of Management. Vol 5.

Russell, S. J., &Norvig, P. (2016). Artificial Intelligence: A Modern Approach. Malaysia; Pearson Education Limited.

Shirgave, Suresh., Awati, Chetan., More, Rashmi., Patil, Sonam. (2019). A Review On Credit Card Fraud Detection Using Machine Learning. International Journal of Scientific & Technology Research. 8. 1217-1220.

Somvanshi, M., & Chavan, P. (2016). A review of machine learning techniques using decision tree and support vector machine. Dipresentasikan dalam International Conference on Computing Communication Control and Automation (ICCUBEA).

Tuanakotta, Theodorus M. (2013). Mendeteksi Manipulasi Laporan Keuangan. Jakarta: Salemba Empat.