ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Machine Learning in Financial Fraud Detection

Mayuresh Patil

Digital Solution Arch, IT Consulting, Fenchurch St, Billingsgate, UK

ABSTRACT: Many financial companies are able to identify fraudulent financial products transactions so that customers are not charged for items that they did not purchase. We can be tackled with machine learning. This paper intends to illustrate the modelling of a data set using machine learning with Financial products Fraud Detection. The Financial products Fraud Detection Problem includes modelling past financial products transactions with the data of the ones that turned out to be fraud. The model is then used to recognize whether a new transaction is fraudulent or not. We here is to detect100% of the fraudulent transactions and incorrect fraud classifications. We have achieved accuracy of93% using logistic regression and 92% using naive Bayes and 93.4% using decision tree and we step into deep learning, we used ANN achieved better accuracy then all other algorithms of 99.69%.Data mining tools can be used to spot patterns and detect fraud transactions. Through data mining, factors leading to fraud can be determined. The performance is analysed based on the parameters of the Total Running Time and the Accuracy. The most commonly used fraud detection methods are Neural Network (NN), rule induction techniques, fuzzy system, decision trees, Support Vector Machines (SVM), Artificial Immune System (AIS),genetic algorithms, K- Nearest Neighbor algorithms.

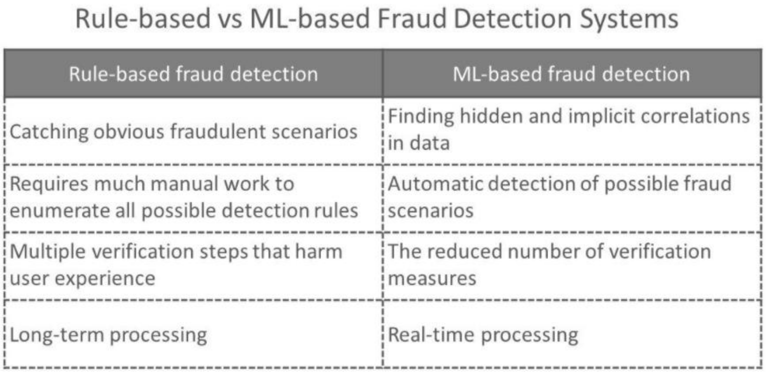
KEYWORDS: Data mining techniques , Fraud detection , machine learning algorithm , multi-level clustering, Data mining, Fuzzy logic, Machine learning,

I. INTRODUCTION

Financial products transactions which has fraud are unauthorized and unwanted usage of an account by someone other than the owner of that account. The fraudulent practices can be analysed to minimize it and protect against similar occurrences in the future. Fraud detection is monitoring the activities of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. This is a problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is challenging from the perspective of learning, as it is characterized by various factors such as class imbalance.. The transaction patterns often change their statistical properties over the course of there is rapid increase in the financial products transaction which as led to substantial growth in fraudulent cases. Many data mining and statistical methods are used to detect fraud. fraud detection techniques are implemented using artificial intelligence, pattern matching. Detection of fraud secure methods are very important. Financial products frauds are increasing heavily because of fraud financial loss is increasing drastically. To reduce these losses prevention or detection of fraud must be done. There are different types of frauds occurring as technology is growing rapidly. So there are many machine algorithms are used to detect fraud now days hybrid algorithms, artificial neural network is used as it gives better performance. Banking sector fraud had accelerated as online and offline transaction. As transactions increase, mode of payment, focus has been given to recent procedure methodologies to handle the fraud. There are many fraud detections solutions and software system which prevents frauds in businesses like financial products products , retail, e-commerce, insurance and industries. Data mining technique is one notable and common methods employed in determination banking sector fraud detection downside. It’s not possible to be sheer sure concerning actuality intention associated right behind an application or transaction. In reality, to hunt out doable evidences of fraud from the accessible data usingmathematical algorithms is the best effective possibility. Fraud Detection Data Mining Techniques and Models can be found useful in the banking sector, mostly for the purpose of detection of fraud happening through

DOI: 10.15680/IJIRCCE.2017. 0508060

the various techniques used by the fraudsters. By using all the different types of data mining models and techniques, ever-increasing fraudulent activities, which are of a major concern for the business as well as the customers and banks, happening can be detected and also reported.

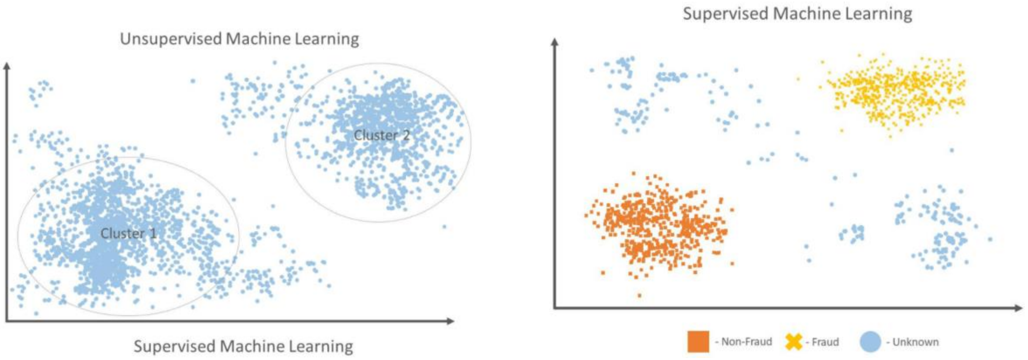


Rule-based vs ML-based Fraud Detection Systems

Rule-based fraud detection	ML-based fraud detection
Catching obvious fraudulent scenarios	Finding hidden and implicit correlations in data
Requires much manual work to enumerate all possible detection rules	Automatic detection of possible fraud scenarios
Multiple verification steps that harm user experience	The reduced number of verification measures
Long-term processing	Real-time processing

II. RELATED WORK

Many studies implemented to detect fraud using supervised, unsupervised algorithms and hybrid ones. Fraud types and patterns are evolving day by day. It is important to have understanding of technologies behind fraud detection. Here discuss machine learning models, algorithms and fraud detection models used in earlier studies. In many models are implemented for fraud detection. In every model different algorithm are used. Detection of financial fraud for new frauds will be problematic if new data has drastic changes in fraud patterns. Logistic Regression algorithm (LR) is implemented to sort the classification problem. Using Gaussian Mixture Models fraudulent cases are discretized. To balance data synthetic minority oversampling is used. Sensitivity analysis is used calculate economic value. This research work has been initiated with the literature study. After reviewing relevant literature, we get to know about research efforts made to overcome the targeted problem and to determine what all data mining techniques have been applied to achieve high accuracy in insurance fraud detection of health care data. In this way people may not fully analyzed the report data for its huge amount and wild range, which caused many shortcomings in judgment. In recent years, data mining method has been widely used in fraud detection to reduce the errors caused by experts' judgment, including Internet fraud detection Yao, J. et al[1] Financial statement fraud has been a difficult problem for both the public and government regulators, so various data mining methods have been used for financial statement fraud detection to provide decision support for stakeholders. The purpose of this study is to propose an optimized financial fraud detection model combining feature selection and machine learning classification. The study indicated that random forest outperformed the other four methods. As to two feature selection methods, Xgboost performed better. And according to our research, 2 or 5 variables are more acceptable for models in this paper. Panigrahi, P. K et al[2] The proposed framework provides a systematic process for the auditors in discovering internal financial frauds. The auditors can use their own experience and investigation skills and integrate with tools and techniques available in different software. The suggested data structures of fraudulent transactions assist the auditors in preparing the data for application of various techniques using software. Chen, Y.-J. et al[3] This study considers the characteristics of variety and value of big data used in finance and economics to develop a big data-based fraud detection approach for the financial statements of business groups to more precisely detect the financial statement fraud of business groups, and thus reducing investment losses and risks and enhancing investment decision making benefits for investors and creditors. Rawte, V., et al[4] Fraud is widespread and very costly to the healthcare insurance system. Fraud involves intentional deception or misrepresentation intended to result in an unauthorized benefit. It is shocking because the incidence of health insurance fraud keeps increasing every year. In order to detect and avoid the fraud, data mining techniques are applied. This includes some preliminary knowledge of health care system and its fraudulent behaviors, analysis of the characteristics of health care insurance data. Data mining which is divided into two learning techniques viz., supervised and unsupervised is employed to detect fraudulent claims. But, since each of the above techniques has its own set of advantages and disadvantages, by combining the advantages of both the techniques, a novel hybrid approach for detecting fraudulent claims in health insurance industry is proposed. In Machine Learning, problems like fraud detection are usually framed as classification problems —predicting a discrete class label output given a data observation..



III. PROPOSED SOLUTION

Behavioural analytics use machine learning to understand and anticipate behaviours at a granular level across each aspect of a transaction. The information is tracked in profiles that represent the behaviours of each individual, merchant, account and device. These profiles are updated with each transaction, in real time, in order to compute analytic characteristics that provide informed predictions of future behaviour. Profiles contain details of monetary and non-monetary transactions. Non-monetary may include a change of address, a request for a duplicate card or a recent password reset. Monetary transaction details support the development of patterns that may represent an individual's typical spend velocity, the hours and days when someone tends to transact, and the time period between geographically disperse payment locations, to name a few examples. Profiles are very powerful as they supply an up- to-date view of activity used to avoid transaction abandonment caused by frustrating false positives Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule or policy with an aim to attain unauthorized financial benefit. Numerous literatures pertaining to anomaly or fraud detection in this domain have been published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates have revealed that techniques employed in this domain include data mining applications, automated fraud detection, adversarial detection. In another paper, Suman, Research Scholar, GJUS&T at Hisar HCE presented techniques like Supervised and Unsupervised Learning for financial products fraud detection. Even though these methods and algorithms fetched an unexpected success in some areas, they failed to provide a permanent and consistent solution to fraud detection. A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of financial products transaction data set of one certain commercial bank. Outlier mining is a field of data mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e. the transactions that aren't genuine. They have taken attributes of customer's behaviour and based on the value of those attributes they've calculated that distance between the observed value of that attribute and its predetermined value. Unconventional techniques such as hybrid data mining/complex network classification algorithm is able to perceive illegal instances in an actual card transaction data set,based on network reconstruction algorithm that allows creating representations of the deviation of one instance froma reference group have proved efficient typically on medium sized online transaction. There have also been efforts to progress from a completely new aspect. Attempts have been made to improve the alert feedback interaction in case of fraudulent transaction. In case of fraudulent transaction, the authorised system would be

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

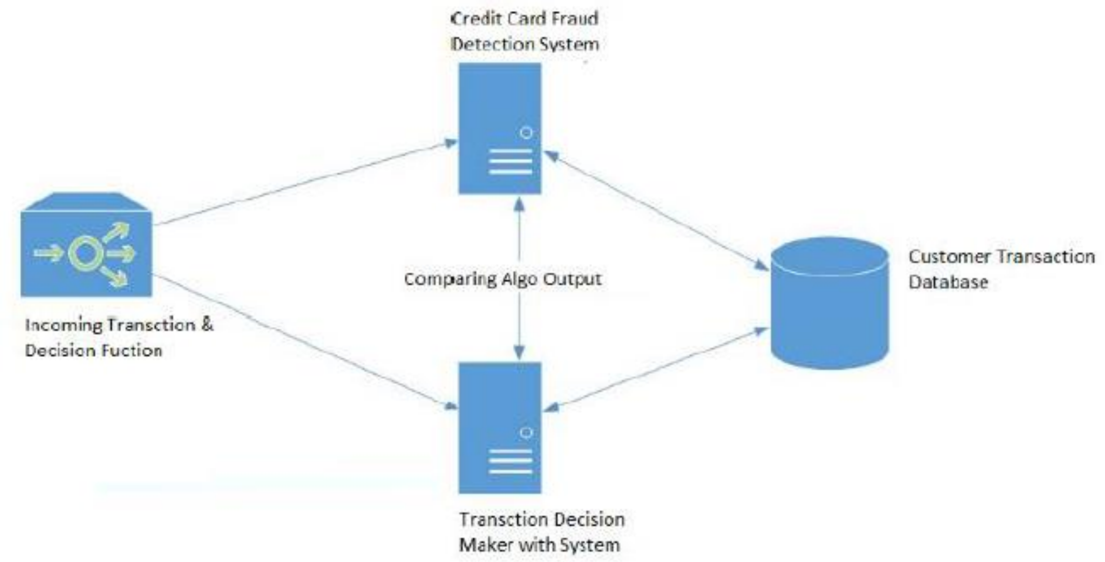
Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

alerted and a feedback would be sent to deny the ongoing transaction. Artificial Genetic Algorithm, one of the approaches that shed new light in this domain, countered fraud from a different direction. It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

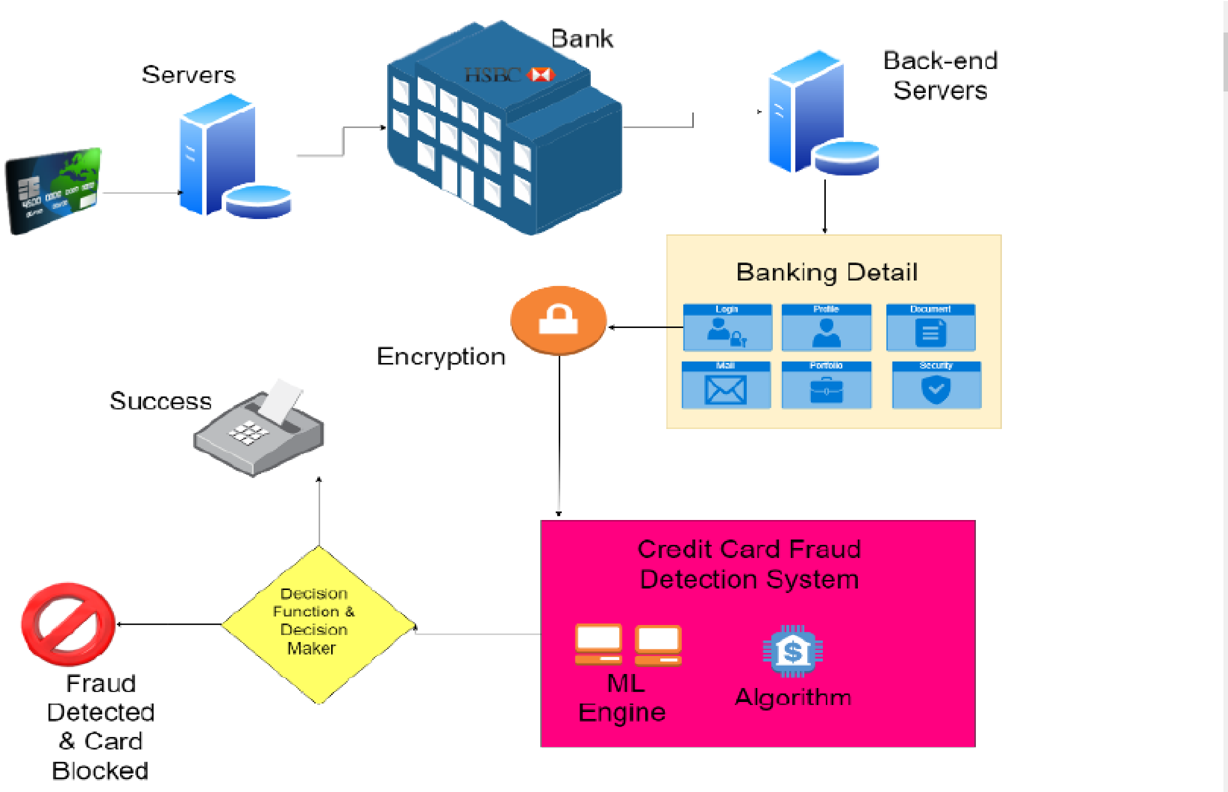
IV. METHODOLOGY

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers. The basic rough architecture diagram can be represented with the following figure

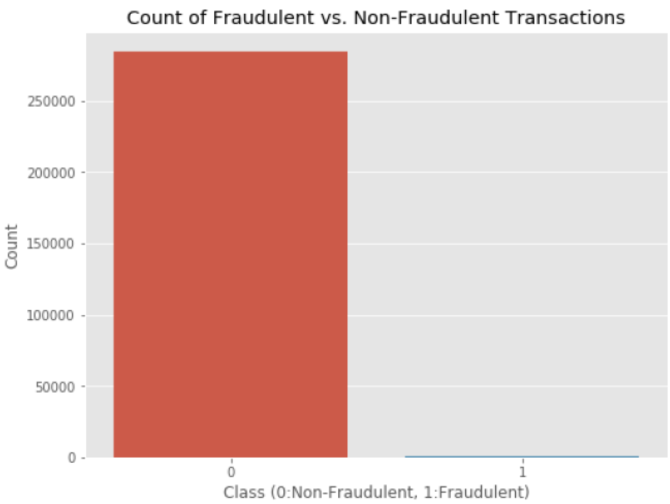

Credit Card Fraud Detection System
Incoming Transction & Decision Fuction
Comparing Algo Output
Customer Transaction Database
Transction Decision Maker with System

When looked at in detail on a larger scale along with real life elements, the full architecture diagram can be represented as follows:

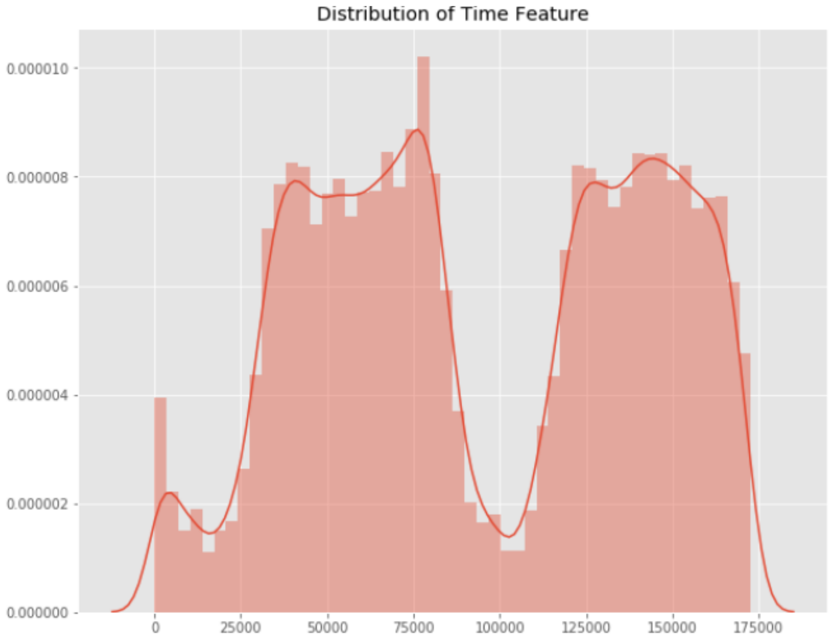
DOI: 10.15680/IJIRCCE.2017. 0508060



First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data. The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one. We plot different graphs to check for inconsistencies in the dataset and to visually comprehend it:



This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.

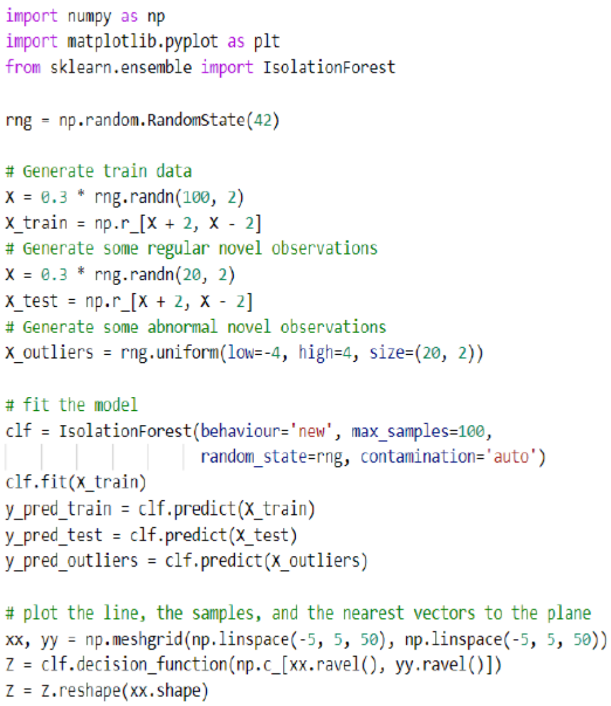


After checking this dataset, we plot a histogram for every column. This is done to get a graphical representation of the dataset which can be used to verify that there are no missing any values in the dataset. This is done to ensure that we don’t require any missing value imputation and the machine learning algorithms can process the dataset smoothly. After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable. The following module diagram explains how these algorithms work together: This data is fit into a model and the following outlier detection

modules are applied on it:
- Local Outlier Factor
- Isolation Forest Algorithm

A. Local Outlier Factor

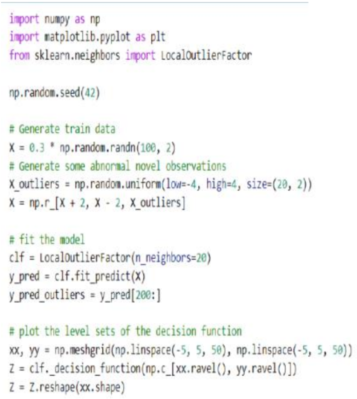
It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbours.
More precisely, locality is given by k-nearest neighbours, whose distance is used to estimate the local data.
The pseudocode for this algorithm is written as:



```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

rng = np.random.RandomState(42)

# Generate train data
X = 0.3 * rng.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * rng.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = rng.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = IsolationForest(behaviour='new', max_samples=100,
                      random_state=rng, contamination='auto')
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)

# plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

B. Isolation Forest Algorithm

The observations by arbitrarily selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designated feature. Recursive partitioning can be represented by a tree, the number of splits required to isolate a sample is equivalent tothe path length root node to terminating node. The average of this path length gives a measure of normality and the decision function which we use. The pseudocode for this algorithm can be written as:



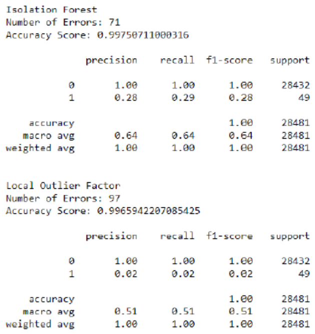
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

np.random.seed(42)

# Generate train data
X = 0.3 * np.random.randn(100, 2)
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
X = np.r_[X + 2, X - 2, X_outliers]

# fit the model
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[200:]

# plot the level sets of the decision function
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf._decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

V. IMPLEMENTATION

This idea is difficult to implement in real life because it requires the cooperation from banks, which aren't willing to share information due to their market competition, and also due to legal reasons and protection of data of their users. Therefore, we looked up some reference papers which followed similar approaches and gathered results.

VI. RESULTS

The code prints out the number of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end and both the results are printed. These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives. Results when 10% of the dataset is used:



```
Isolation Forest
Number of Errors: 71
Accuracy Score: 0.99750711000316

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.28      0.29      0.28        49

    accuracy                           1.00     28481
   macro avg       0.64      0.64      0.64     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor
Number of Errors: 97
Accuracy Score: 0.9965942207085425

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

VII. CONCLUSION AND FUTURE WORK

This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results. While the algorithm does reach over 99.9% accuracy, its precision remains only at 29% when a tenth of the data set is taken into consideration. Choosing the right machine learning method depends on the problem type, size of a dataset, resources, etc. A good practice is to use several models to both streamline assessment and achieve higher accuracy.

For example, PayPal implements express assessment using linear models to separate uncertain transactions from ordinary ones. Then, all transactions that look suspicious are run through an ensemble of three models comprising a linear model, a neural network, and a deep neural network. The three then vote to arrive at the final result with the higher accuracy.
As of today, antifraud systems should meet the following standards:

- detect fraud in real-time
- improve data credibility
- analyze user behavior
- uncover hidden correlations

While these qualities can be offered by machine learning algorithms, they have two serious drawbacks to be aware of. They still require large and carefully prepared datasets for training and still need some features of rule-based engines, like checking legal limitations for cash transactions. Also, machine learning solutions usually require substantial data science skills to build complex and robust ensemble algorithms. This sets a high barrier for small and medium companies to use the technique leveraging internal talent. t. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project

REFERENCES

[1] Yao, J., Zhang, J., & Wang, L. (). A financial statement fraud detection model based on hybrid data mining methods. International Conference on Artificial Intelligence and Big Data(ICAIBD). doi:10.1109/icaibd..8396167.

[2] Panigrahi, P. K. (2011). A Framework for Discovering Internal Financial Fraud Using Analytics. 2011 International Conference on Communication Systems and Network Technologies. doi:10.1109/csnt.2011.74

[3] Rambola, R., Varshney, P., & Vishwakarma, P. (). Data Mining Techniques for Fraud Detection in Banking Sector. 4th International Conference on Computing Communication and Automation (ICCCA). doi:10.1109/ccaa..8777535.

[4] Rawte, V., & Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques. 2015 International Conference on Communication, Information & Computing Technology (ICCICT). doi:10.1109/iccict.2015.7045689.

[5] Verma, A., Taneja, A., & Arora, A. (2017). Fraud detection and frequent pattern matching in insurance claims using data mining techniques. 2017 Tenth International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2017.8284299.

[6] Jayabrabu, R., Saravanan, V., & Tamilselvi, J. J. (2014). A framework for fraud detection system in automated data mining using intelligent agent for better decision making process. 2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE). doi:10.1109/icgccee.2014.6922411

[7] Chen, Y.-J., & Wu, C.-H. (2017). On Big Data-Based Fraud Detection Method for Financial Statements of Business Groups. 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). doi:10.1109/iiai-aai.2017.13.

[8] B.Rajasekhar, B. Sunil Kumar, Rajesh Vibhudi, “Quality of Cluster Index Based on Study of Decision Tree”, International Journal of Research in Computer Science, Vol 2, Issue 1, pp 39-43, eISSN 2249- 8265, 2011.

[1] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, “Random forest for financial products fraud detection”, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),.

[2] Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey, “A Tool for Effective Detection of Fraud in Financial products System”, published in International Journal of Communication Network Security ISSN: 2231 – 1882, Volume-2, Issue-1, 2013.

[3] Rinky D. Patel and Dheeraj Kumar Singh, “Financial products Fraud Detection & Prevention of Fraud Using Genetic Algorithm”, published by International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[4] M. Hamdi Ozcelik, Ekrem Duman, Mine Isik, Tugba Cevik,
“Improving a financial products fraud detection system using genetic algorithm”, published by International conference on Networking and information technology, 2010.

[5] Wen-Fang YU, Na Wang,“ Research on Financial products Fraud Detection Model Based on Distance Sum”, published by IEEE International Joint Conference on Artificial Intelligence, 2009.

[6] Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to Fraud Detection"; Department of Computer Science- Columbia University; 2000.

[7] Salvatore J. Stolfo, Wei Fan, Wenke Lee and Andreas L. Prodromidis;"Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project"; 0-7695-0490-6/99, 1999 IEEE.

[8] Soltani, N., Akbari, M.K., SargolzaeiJavan, M., “A new user-based model for financial products fraud detection based on artificial immune system,” Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on., IEEE, pp. 029-033, 2012.

[9] S. Ghosh and D. L. Reilly, “Financial products fraud detection with a neuralnetwork”, Proceedings of the 27th Annual Conference on System Science, Volume 3: Information Systems: DSS/ Knowledge Based Systems, pages 621-630, 1994. IEEE Computer Society Press.

[10] MasoumehZareapoor, Seeja.K.R, M.Afshar.Alam, ”Analysis of Financial products Fraud Detection Techniques: based on Certain Design Criteria”, International Journal of Computer Applications (0975 – 8887) Volume 52– No.3, 2012