International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (online): 2347-5552, Volume-11, Issue-3, May 2023
https://doi.org/10.55524/ijircst.2023.11.3.19
Article ID IJIRD-1257, Pages 96-100
www.ijircst.org

Prediction of Financial Crime Using Machine Learning

Indurthy Meghana[1], Bitra Pavan Venkatesh[2], Gaddipati Keerthi Ganesh[3], Nadendla Sumanth[4], and Redrouthu Tarun Teja[5]

[1,2,3,4,5] Department of Computer Science & Engineering, PACE Institute of Technology and Sciences,
Ongole, Andhra Pradesh, India

Correspondence should be addressed to Indurthy Meghana; meghana.i@pace.ac.in

ABSTRACT- The purpose of data analytics is to uncover previously unknown patterns and make use of such patterns to help in making educated decisions across a wide range of contexts. Because of advances in modern technology and the fact that credit cards have become an easy target for fraudulent activity, the incidence of credit card fraud has considerably increased in recent years. Credit card fraud is a significant issue in the industry of financial services, and it results in annual losses of billions of dollars. The development of a fraud detection algorithm is a difficult endeavor due to the paucity of real-world transaction datasets that are available due to confidentiality concerns and the very unbalanced nature of the datasets that are publicly available. Use a dataset from the real world in conjunction with a variety of supervised machine learning algorithms to identify potentially fraudulent credit card transactions. In addition, make use of these techniques to create a super classifier through the use of ensemble learning methods. Determine which variables are the most significant and could perhaps lead to a higher level of accuracy in the identification of fraudulent credit card transactions. In addition, we evaluate and discuss the performance of a number of other supervised machine learning algorithms that are currently available in the literature in contrast to the super classifier that can be implemented.

KEYWORDS- Financial crimes, machine leanring, Classification, Accuracy.

I. INTRODUCTION

One of the most fascinating developments in modern technology is known as machine learning (ML). A rapidly developing subfield of data science, machine learning examines how computers might educate themselves through observation and experimentation. One of these challenges consists of providing assistance with the process of developing forecasts for financial data [1]. To put it another way, ML refers to computer programs that automatically improve their overall performance as a result of their accumulation of experiences [2]. It offers the ability to produce systems that are capable of automatically adapting to individual users and customizing themselves to their preferences.

Credit card fraud can be separated down into three distinct categories: crimes committed against the card itself, those committed against merchants, and those committed against the Internet. Frauds connected to credit cards include, but are not limited to, the use of false or counterfeit cards, shoulder surfing, account takeover, and the use of cards that have been lost, stolen, or otherwise compromised. Frauds committed by merchants can include merchant collusion, which occurs when a person working for the merchant takes the credit card information of the merchant's customers. Frauds committed via the internet during online transactions, such as site-cloning, building bogus merchant sites, snooping, man-in-the-middle attack [MITM], and denial-of-service attack, are the most serious of all the different types of fraud. Due to the fact that the traditionally employed AVS (Address Verification System), CVM (Card Verification Method), and other security mechanisms have been shown to be weak against these assaults, customers are unable to make full use of the credit card transaction system [3].

Criminal action or the unauthorized use of any system is what we refer to when we talk about credit card fraud. An exponential rise in the number of people choosing to make their purchases online can be attributed to the proliferation of e-commerce as well as the convenience of conducting business and making payments online.

Credit cards are by far the most frequently used means of payment for internet purchases. There are two distinct types of purchases that can be made using a credit card: 1) a physical card and 2) an electronic card. It is considered to be a "Physical purchase" when the cardholder is present in the same physical location as the card while the transaction is being processed. Web payment gateways like PayPal and Alipay handle the majority of financial transactions online, particularly those involving credit cards.

Every card holder has their own unique pattern, which comprises information about the total amount of transactions, specifics about the things that were purchased, information about the seller, the date of the transaction, and more.

It will be the most efficient approach to fight fraud in financial transactions conducted over the internet. If the system detects even the slightest deviation from the cardholder's typical behavior, an alert will be sent to the user's device, which will cause the transaction to be halted. The purpose of the system to detect fraud is to identify fraudulent activity precisely and well in advance of its commission. The objective is to conduct as accurate and thorough a false fraud detection as possible.

First, the characteristics of the transaction records need to be completely ordered, and then the values of each attribute need to be categorized. We build a logical graph of BP (LGBP) based on them, which abstracts and encompasses all of the various transaction records. We characterize the transactional behaviors of users and the diversity of those behaviors by defining, on the basis of LGBP, the path-based transition probability and diversity coefficient. In order to capture the temporal aspects of transactions, you should also design a state transition probability matrix. After that, construct a BP for each individual user. For the purpose of determining whether or not an incoming transaction is legitimate, a BP-based fraud detection method that takes into account the idea drift problem has been presented [4].

The scam is uncovered through the use of outliers. In the supervised method, the models are utilized to distinguish between fraudulent and non-fraudulent behavior in order to locate the outlier. The practice of clustering is applicable in the field of engineering as well as in a variety of scientific fields, including psychology, biology, medicine, computer vision, communication, and remote sensing. When conducting clustering, one observes a group of patterns as a result of abstracting the underlying structure. The patterns have been clustered based on the number of shared characteristics that are more than those shared by any other pattern group. In order to satisfy a wide range of needs, a number of alternative clustering methods have been proposed. The structure of abstraction serves as the foundation for clustering methods, which can be broken down further into hierarchical and partitioning subcategories. In order to develop a physical plan, the catalyst optimization algorithm is applied. The structure of Catalyst can be thought of as a tree made up of node objects [5].

Methods of supervised learning focus on analyzing a variety of previous transactions, which are recorded by the cardholder or the credit card provider, in order to determine whether or not a new purchase is fraudulent. For this strategy to work, you will need a dataset that has been segmented between fraudulent and legitimate observations. Methods of unsupervised learning necessitate the structuring of unlabeled data into groupings of conceptual similarity referred to as clusters. They base their work on the assumption that out of the ordinary occurrences signify fraudulent transactions. Clustering makes it feasible to discover different data distributions, which enables one to apply different prediction models to each of those distributions.

Semi-supervised ones combine the aforementioned procedures in order to profit from recalling previous fraudulent transactions and applying unsupervised techniques in order to identify future fraudulent transaction patterns. Semi-supervised ones are also known as hybrid ones. A hybrid technique that includes multiple machine learning algorithms, such as SVM, MLP, random forest regression, autoencoder, and isolation forest, can be used to identify fraudulent credit card transactions. This hybrid technique is one option.

Our work has made the following significant contributions.

- The fraud detection team’s heavy workload of data processing is eliminated by machine learning. The findings aided the team’s research, insights, and reporting.
- The following ML methods were built and evaluated: SVM, KNN, DT, NB, and LR. Using fresh data retrieved from the form, banks may discover the default behavior of consumers and anticipate whether a person would commit fraud or not. The rest of the paper is organized as follows: Section 2 discusses the related work.

II. RELATED WORK

This issue was handled by our team in the publication that we published by employing the method of loan default prediction using ML algorithms. The anticipation of loan repayment is a topic of discussion in the banking and financial industries. In today's highly competitive economic system, a credit rating has evolved into an extremely important component [6].

Academic interest has increased in recent years as a result of recent advancements in data science and artificial intelligence. At the moment, its primary concerns are credit risk analysis and loan projections. The rising demand for loans necessitates the development of more accurate credit scoring and loan prediction systems. Determining an individual's credit rating has been the subject of research for several decades. In the past, experts were used, and current models continue to rely on the opinions of experts, but the focus these days is on automating as much of the process as possible. ML algorithms and neural networks are currently being utilized for the purposes of credit rating and risk assessment. This subject area has experienced a number of noteworthy achievements, paving the way for forthcoming research [7].

The paper [8] examined a variety of methods, including SVMs, KNNs, artificial neural networks (ANN), logistic regression (LR), ANN with stochastic gradient (SGD), boosting, RF, naive Bayes (NB), and others, and came to the conclusion that there is not a single method that is superior to the others. The authors conducted research on credit ratings for mortgages and arrived at the following findings [9]: credit applications that do not match the standards are typically declined due to the default risk.

Applicants with lower incomes have a better probability of being accepted and of timely loan repayment if they apply for financial aid. Exploratory data analysis was used in the research that was carried out by the authors of [10].

The primary purpose of the study was to classify and assess the individuals who applied for loans. The authors observed, through the use of seven different graphs, that the vast majority of people who applied for loans preferred short-term loans. The authors of [11] asserted that SVMs are capable of outperforming LR and RF based on the benchmarking models presented in the publication. In addition, by using models like as LR, RF, GB, etc., they demonstrated the necessity of doing data quality checks, such as data analysis and cleaning, prior to the modeling process. According to the research presented in the study, there are two important factors that need to be taken into consideration when financing a loan: the algorithm and the selection of features. In their study [12], the authors created a model for determining the likelihood of defaulting on a loan by employing data mining techniques. Specifically, they employed three algorithms, namely J48, NB, and Bayes net. J48 was determined to be the most effective algorithm for the task on account of the high accuracy (78.37%) and low mean absolute error

(0.34) that it possessed. The model that Aditi Kacheria and her colleagues developed relied on NB modelling. They increased both the quality of the data and the accuracy of the classification by using KNN and binning. The KNN algorithm was utilized to deal with the missing data, and binning was utilized to get rid of any irregularities. According to the findings of their research, the majority of the local banks in the Czech and Slovak Republics utilize logit-based models.

Credit card transactional details such as account numbers, card types, types of purchases, places and times of transactions, client names, merchant codes, transaction volumes, and so on are included in the datasets that are associated with credit cards. Multiple researchers made use of these data as a variable in order to determine whether or not the transaction was legitimate or fraudulent, as well as to identify outliers that needed additional investigation [13].

III. PROPOSED APPROACH

This study was founded on the analysis of data regarding bank loans that was acquired through Lending Club. Lending Club is an online banking platform that provides borrowers with the opportunity to acquire loans from willing investors at a cheaper interest rate. The dataset was derived from a real-world, publicly available source, which can be accessed at https://www.kaggle.com/datasets/wordsforthewise/lending-club. This platform is a type of marketplace that facilitates the creation of peer-to-peer connections between borrowers and investors. It does this by helping to match borrowers and investors. The purpose of this research was to discover the individuals who are requesting for loans despite having a poor credit history. On the Jupyter Notebook, which is an open-source platform, we employed the Python programming language and its associated libraries. Not only is the Jupyter Notebook easy to use, but it also has a greater level of performance. In this research, we evaluate the accuracy of several alternative algorithms and design a system that is capable of performing the early prediction of the behaviors of consumers with a greater accuracy and f-score.

The primary objective of our study was to find a way to reduce the amount of false-positive results, also known as Type-1 error and, depending on how accurate the results were, f-score and false-positive parameters. We decided to go with the algorithm that had the best track record of accurately forecasting whether or not the customer would be able to pay, and then we took the appropriate steps to reduce the potential for loss as much as possible.

A representation of our suggested system can be found in Figure 1, where it is shown as a model.

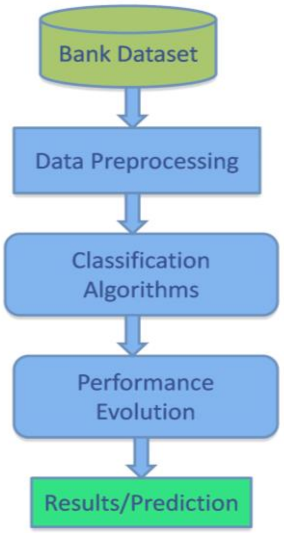


Figure 1: Proposed approach

A. Logistic Regression

Problems with classification can be solved by employing a method of supervised learning known as logistic regression. When dealing with classification problems, the dependent variables are typically discrete or binary values like 0 or 1. Logical regression employs categorical variables in its algorithm, such as 0 or 1, Yes or No, True or False, Spam or Not Spam, etc. The predictive analytic approach that it employs is predicated on the concept of probability. In spite of the fact that it is a type of regression, the application of logistic regression is not the same as the application of the algorithm for linear regression. An expensively sophisticated form of fun.

B. Support Vector Machine

The support vector machine (SVM) is a technique that can be used for classification and regression problems. On the other hand, it finds widespread application in the categorization problems of machine learning. The goal of the Support Vector Machine (SVM) technique is to determine the optimum line or decision boundary that may divide an n-dimensional space into classes. This will enable us to categorize new data points in the future in a more efficient manner. This optimal decision boundary is referred to as a hyperplane in the scientific community. The SVM is responsible for selecting the extreme vectors and points that are used in the construction of the hyperplane. The Support Vector Machine (SVM) approach gets its start from support vectors, which are used to represent very severe examples.

C. Decision Tree

The training data are continuously segmented depending on a certain parameter as you describe the input and the related output when using decision trees, which are a form of supervised machine learning. The components of the tree known as decision nodes and leaves are the ones that can be used to explain the tree. The leaves are a representation of the various options or outcomes. At the decision nodes, the data are split up into different categories.

D. Random Forest

The training data are continuously segmented depending on a certain parameter as you describe the input and the related output when using decision trees, which are a form of supervised machine learning. The components of the tree known as decision nodes and leaves are the ones that can be used to explain the tree. The leaves are a representation of the various options or outcomes. At the decision nodes, the data are split up into different categories.

E. K-Nearest Neighbor

The approach to machine learning known as k-nearest neighbor is rooted in statistics and can be used for classification and regression. The input for the KNN classification is the k-nearest training samples, and the output is class membership. In KNN, K is a positive integer that is almost always on the low end.

IV. RESULTS AND DISCUSSIONS

In the beginning, the Pandas Dataframe was utilized in order to perform analysis on the dataset that was obtained from the Lending Club website. This provided the specific information regarding the datasets, including their sizes, types, feature information, and so on. Table 3 displays the information on the dataset that was used for this study. We carried out an exploratory data analysis in order to collect information regarding the correlation between the characteristics of the datasets so that we could incorporate that into the visualization of the data.

The decision tree regressor was selected to serve as the behavioral model for this body of work. The interest rate instalment was taken into consideration as the behavioral variable.

It's possible that a credit card issuer will take notice of a cardholder's shift in spending habits over the course of the previous six months, moving from discount to upmarket merchants, for example. It is possible that the credit card holder will engage in fraudulent activity in the not-too-distant future if they are late with the payment of the interest rate instalment. The card issuer takes into consideration additional information, such as whether or not the cardholder has made late payments or is only paying the minimum payment, in order to further limit down the options and develop a more accurate risk profile. Payment delays are related with an increased risk of business failure and personal bankruptcy.

In order to determine whether or not there are any significant correlations between the many variables that make up our dataset, we constructed a correlation matrix. This helps us determine whether or not some of the features in our dataset need to be removed. Additionally, it demonstrates which characteristics are essential for the overall classification. Our correlation matrix was transformed into a really good visual display that was simple to understand thanks to the seaborn library for SNS heatmap, which we utilized for this purpose.

Fig. 2 shows the credit policy description, various classifiers have used, and accuracy results presented in Fig. 3 and Fig. 4. The Fig. 3 Classification Score of NB, DT, and KNN classifiers. Fig. 4 Classification Score of RF, SVM, and LR classifiers.

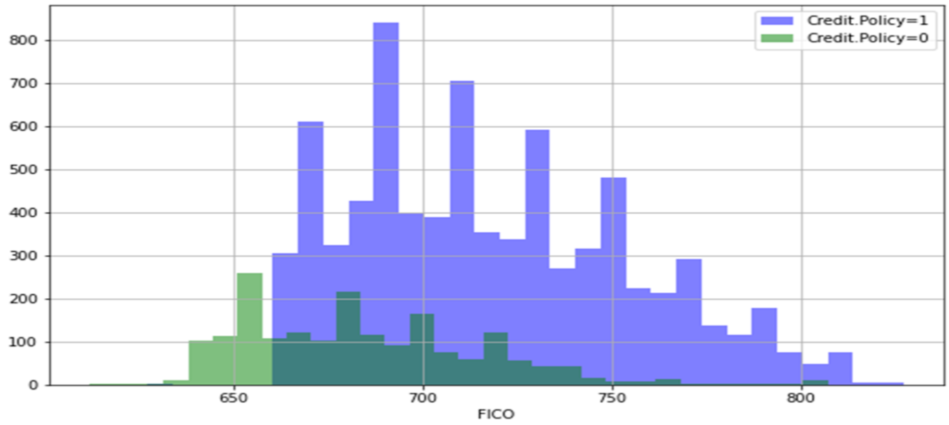
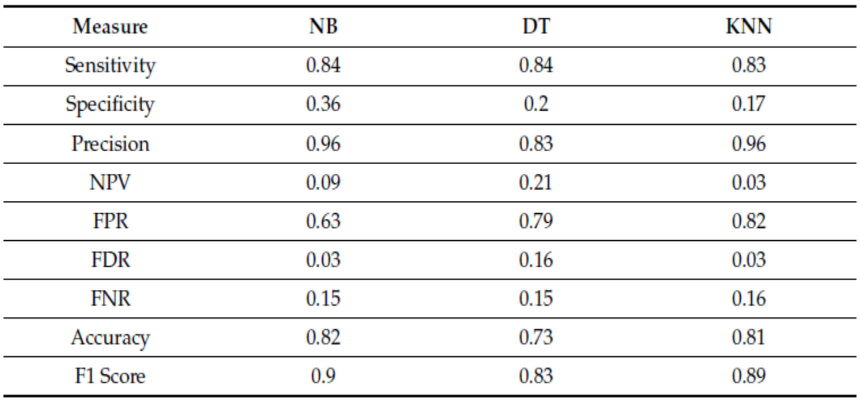
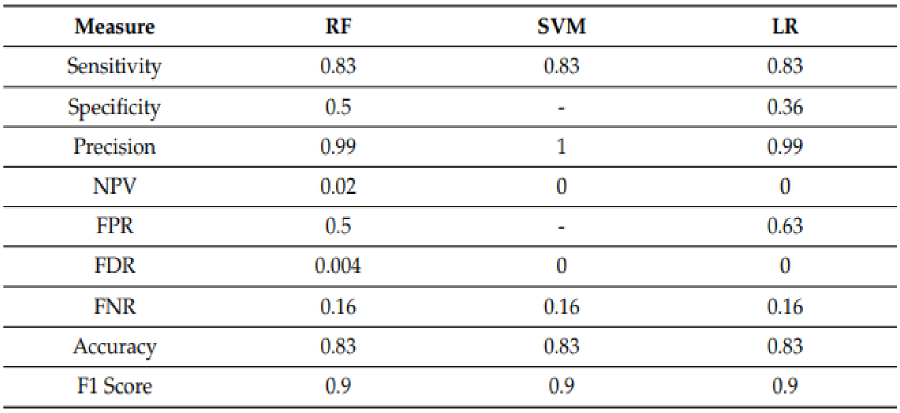


Figure 2: Credit policy description.



Measure	NB	DT	KNN
Sensitivity	0.84	0.84	0.83
Specificity	0.36	0.2	0.17
Precision	0.96	0.83	0.96
NPV	0.09	0.21	0.03
FPR	0.63	0.79	0.82
FDR	0.03	0.16	0.03
FNR	0.15	0.15	0.16
Accuracy	0.82	0.73	0.81
F1 Score	0.9	0.83	0.89

Figure 3: Classification Score of NB, DT and KNN classifiers



Measure	RF	SVM	LR
Sensitivity	0.83	0.83	0.83
Specificity	0.5	-	0.36
Precision	0.99	1	0.99
NPV	0.02	0	0
FPR	0.5	-	0.63
FDR	0.004	0	0
FNR	0.16	0.16	0.16
Accuracy	0.83	0.83	0.83
F1 Score	0.9	0.9	0.9

Figure 4: Classification Score of RF, SVM and LR classifiers

V. CONCLUSION

Analyses inform the following detailed descriptions. First, we compared the ML-based models to the empirical models and found that the former included behavioral aspects while the latter just included application factors. Then, the study examined the six ML techniques using sensitivity; specificity, accuracy, precision, F1 score, and CAP curve tests and found that the SVM model performed best in the great life cycle model. Finally, this study opens new research avenues that will impact credit and credit risk models. This paper's model can help financial institutions detect fraud. These steps will help financial institutions avoid future scams and fraud. Understanding human behavior in the present has become crucial in any field. Knowing how the behavior of a system, economic unit, or other actor influences the micro or macro level is a fascinating issue. As reported in 2019, Romanian banks face a difficult 2020 [14]. According to official forecasts, 12% of banks could lose money this fiscal year, with customer profitability down 60% from 2019. Most financial institutions are researching their clients' demands and needs inside and outside, but many questions remain. Business banks value long-term customer relationships, therefore they use customer behavior research to predict consumer behaviors and impact product and service design. To keep consumers, banks must satisfy their needs in a competitive climate. Banks also want healthy lending portfolios. This study suggests machine learning can solve these issues.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Almorsy, Mohamed, John Grundy, and Ingo Müller. "An analysis of the cloud computing security problem." arXiv preprint arXiv: 1609.01107 (2016).

[2] Perols, Johan L., et al. "Finding needles in a haystack: Using data analytics to improve fraud prediction." The Accounting Review 92.2 (2017): 221-245.

[3] Patil, Suraj, Varsha Nemade, and Piyush Kumar Soni. "Predictive modelling for credit card fraud detection using data analytics." Procedia computer science 132 (2018): 385-395.

[4] Delamaire, Linda, Hussein Abdou, and John Pointon. "Credit card fraud and detection techniques: a review." Banks and Bank system’s 4.2 (2009): 57-68.

[5] Chaudhary, Khyati, Jyoti Yadav, and Bhawna Mallick. "A review of fraud detection techniques: Credit card." International Journal of Computer Applications 45.1 (2012): 39-44.

[6] Dhankhad, Sahil, Emad Mohammed, and Behrouz Far. "Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study." 2018 IEEE international conference on information reuse and integration (IRI). IEEE, 2018.

[7] Patil, Suraj, Varsha Nemade, and Piyush Kumar Soni. "Predictive modelling for credit card fraud detection using data analytics." Procedia computer science 132 (2018): 385-395.

[8] Zareapoor, Masoumeh, K. R. Seeja, and M. Afshar Alam. "Analysis on credit card fraud detection techniques: based on certain design criteria." International journal of computer applications 52.3 (2012).

[9] Madaan, M.; Kumar, A.; Keshri, C.; Jain, R.; Nagrath, P. Loan default prediction using decision trees and random forest: A comparative study. IOP Conf. Series: Mater. Sci. Eng. 2021, 1022, 012042

[10] Jency, X.F.; Sumathi, V.P.; Sri, J.S. An exploratory data analysis for loan prediction based on nature of the clients. Int. J. Recent Technol. Eng. (IJRTE) 2018, 7, 176–179.

[11] Berrada, I.R.; Barramou, F.Z.; Alami, O.B. A review of Artificial Intelligence approach for credit risk assessment. In Proceedings of the 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP), Vijayawada, India, 12–14 February 2022.

[12] Hamid, A.J.; Ahmed, T.M. Developing prediction model of loan risk in banks using data mining. Mach. Learn. Appl. Int. J. (MLAIJ) 2016, 3, 1–9.

[13] Zareapoor, Masoumeh, K. R. Seeja, and M. Afshar Alam. "Analysis on credit card fraud detection techniques: based on certain design criteria." International journal of computer applications 52.3 (2012).

[14] Madaan, M.; Kumar, A.; Keshri, C.; Jain, R.; Nagrath, P. Loan default prediction using decision trees and random forest: A comparative study. IOP Conf. Series: Mater. Sci. Eng. 2021, 1022, 012042