CREDIT CARD FRAUD DETECTION SYSTEM

PRIYADARSHINI COLLEGE OF ENGINEERING

Dr. P. N. FALE

YASH BANDIWAR [1] PARTIK PILLEWAN [2] GAURAV GOKHALE [3] GAURAV TARALE [4]

Abstract -
In this review article, we'll talk about how machine learning can be used to spot credit card fraud. It is more crucial than ever to have precise mechanisms in place to spot fraudulent conduct given the rise in online transactions. The authors suggest applying machine learning techniques to pre-process data sets and analyze them to precisely identify fraudulent credit card transactions. While minimizing false positive fraud classifications, the goal is to identify 100% of fraudulent transactions. To accomplish this, the study focuses on employing anomaly detection methods on modified credit card transaction data.

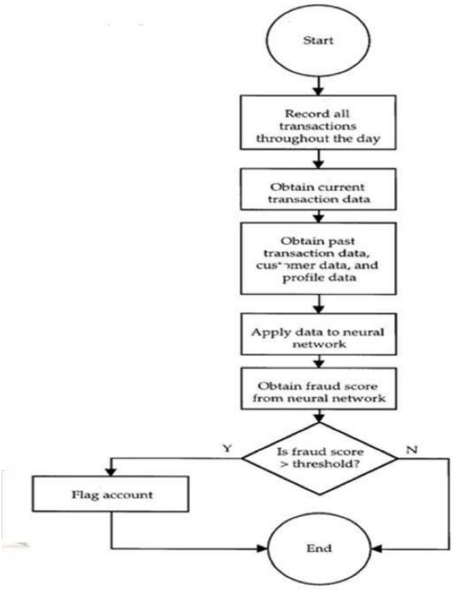
I. INTRODUCTION

The unauthorized use of another person's credit card to make purchases or withdraw cash is referred to as credit card fraud. The fraudster obtains the credit card information using dishonest methods, such as stealing the actual card, breaking into the account of the cardholder, or duping the cardholder into disclosing their information. Prior to receiving a statement or alert of suspicious behaviour, the card issuing authority and the legitimate cardholder are unaware that the card is being used fraudulently. The cardholder may suffer financial losses because of this illegal action, which could also lower their credit score.

Credit card fraud detection is, technically speaking, a procedure that examines user behaviour and transactions to detect and stop unauthorized use of credit cards. Credit card issuers must put in place efficient fraud detection methods due to the rise in credit card fraud that has accompanied the expansion of e-commerce. Since they can examine enormous datasets of authorized transactions to find suspect behaviour, machine learning methods are frequently utilized for this purpose. complaints of possibly fraudulent transactions are generated by the algorithms, and these complaints are then investigated by human experts to ensure their veracity. The machine learning algorithms are constantly updated and taught to increase their accuracy over time based on the feedback from the investigators. Users' actions are observed.

By using fraud detection, fraudulent actions including intrusion, defaulting, and the unauthorized use of credit cards can be stopped.

We'll employ a machine learning base model to recognize the various kinds of anomalies. The process flow diagram is shown in Figure 1.


Start
Record all transactions throughout the day
Obtain current transaction data
Obtain past transaction data, customer data, and profile data
Apply data to neural network
Obtain fraud score from neural network
Y
Is fraud score > threshold?
N
Flag account
End

Design of proposed algorithm

II. LITERATURE REVIEW

In their model, Prajal Save et al. [18] combined Luhn's and Hunt's algorithms with a decision tree. To detect if an incoming transaction is fraudulent or not, Luhn's algorithm is employed. The input, which is the credit card number, is used to validate credit card numbers.

The degree of outlierness and address mismatch are used to evaluate how far each incoming transaction deviates from the typical profile of the cardholder. The Bayes Theorem is used in the final phase to determine whether the general belief has been reinforced or weakened.

by employing a sophisticated combination heuristic to combine the computed likelihood with the original suspicion of fraud.

J. Vimala Devi et al. Three machine-learning techniques were described and put into use to find fraudulent transactions. The performance of classifiers or predictors is assessed using a variety of metrics, including the Vector Machine, Random Forest, and Decision Tree.

Either these measures depend on or don't depend on prevalence.

Additionally, similar methods are employed in mechanisms that detect credit card fraud, and the outcomes of these algorithms have been contrasted. supervised algorithms by Popat and Chaudhary [20] were presented. Some of the techniques utilised include Deep Learning, Logistic Regression, Nave Bayesian, Support Vector Machine (SVM), Neural Network, Artificial Immune System, K Nearest Neighbour, Decision Tree, Fuzzy Logic Based System, and Genetic Algorithm. Algorithms for detecting credit card fraud show which transactions are likely to be fraudulent.

To perform prediction, grouping, and outlier detection, we compared machine learning techniques. Shiyang Xuan and others, 21 The Random Forest classifier was used to train the behavioural characteristics of credit card transactions. The following categories are employed to train the characteristics of legitimate and dishonest behaviour: both a random forest based on CART and a random forest based on random trees. Performance measurements are computed to evaluate the model's efficacy.

Geetha S. and Dornadula [5] The transactions were combined into appropriate groups using the sliding-window method, and then some features from the window were extracted to discover trends in cardholder behaviour. There are features like the maximum amount, minimum amount of a transaction, average amount in the window, and even the length of time that has passed. Sangeeta Mittal and others, 22 Some well-known supervised and unsupervised machine learning techniques were chosen to assess the underlying issues. From traditional to contemporary supervised learning methods have all been taken into consideration. These include of Bayesian methods, hybrid algorithms, deep and traditional neural networks, tree-based algorithms, and so forth. It has been evaluated how well machine learning algorithms can spot credit card fraud. Numerous well-known algorithms in the supervised, ensemble, and unsupervised categories were assessed on various criteria.

It is determined that unsupervised algorithms perform better across all measures both in isolation and in contrast to other approaches because they are better at handling dataset skewness. Akila and Deepa [17] Different methods, including the Anomaly Detection Algorithm, K-Nearest Neighbour, Random Forest, K-Means, and Decision Tree, were employed to identify fraud. Based on a specific scenario, multiple strategies were provided, and the ideal algorithm for spotting dishonest transactions was projected. The system generated a fraud score for that specific transaction using a variety of criteria and algorithms to forecast the outcome of fraud. An approach for detecting fraud using deep networks has been presented by Xiaohan Yu et al. The article presented a deep neural network approach for identifying credit card fraud.

III. METHODOLOGY

For instance, systematic literature reviews are a form of technique that conducts a literature review on a particular subject and may be used to spot fraud. The main objective of a systematic review in this situation is to locate, assess, and interpret the literature-based studies that answer the authors' research objectives. Finding research possibilities and gaps in the area of interest is a secondary objective. In this study, we made an effort to go through Kitchenham's suggested actions of planning, carrying out, and reporting analysis iteratively. [28]

3.1 Selection of rudimentary Studies

Keywords were entered into the search engine to highlight primary research for selection, and they were then picked to support the growth of research that aims to help answer the study's questions. The only valid Boolean operators were AND and OR. The search criteria were (machine-learning OR machine learning) AND "fraud detection." One of the systems investigated was the IEEE Explore Digital Library.

- Science Direct - Elsevier - Google Scholar - Website The title, keywords, and abstract were all searched for, per the search platforms. We carried out the searches and reviewed all of the earlier research on March 28, 2021. The outcome of these searches refined using the criteria described in Section 3.2, resulting in a collection of results that could be run.

3.2 Inclusion and Exclusion Criteria

The inclusion of case studies, opinions on how to construct a hybrid method to strengthen current procedures, and modern technology fraud detection might all be taken into consideration for this SLR. English must be used for all writing and reading on papers. Any Google Scholar results are scrutinised before submission, as if Google Scholar had the power to reject publications of inferior calibre. The most current iteration of a sample must be submitted for this SLR.

3.3 Machine learning classifiers

In this project, we used a total of five classifications methods Logistic regression, KNN, Support vector machine (SVM), Decision tree (DT), Category boosting (Cat boots) These classification algorithm methods are widely used for problems such as differential training dataset. Also, it is commonly used in classification learning. That is the reason I compare them in the same training dataset. Also, it can be a cross-sectional comparison with other current studies in the results.

Use logistic regression to detect credit card fraud. Logistic regression is the classical and the best bicategorical algorithm, which is preferred when dealing with classification problems, especially bicategorical ones. The choice of algorithm is based on the principle of simplicity before complexity. Logistic regression is also an excellent choice because it is a recognized statistical method used to predict the outcome of a binomial or polynomial.

A multinomial logistic regression algorithm can regenerate the model. It will be a better classification algorithm when the target field or data is a set field with two or more possible values. The advantage of logistic regression is that he is faster to process and is suitable for bicategorical problems. It is also more straightforward for any beginner to understand and directly see the weights of each feature. Then it is easier to update the model and incorporate new data for different problems (Aihua et al. 2007). Furthermore, it has a disadvantage. There is a limit to the data and the adaptability of the scene. Not as adaptable as the decision tree algorithm. But this is an issue that we can also determine in this project based on the actual situation whether the logistic regression has a better ability to adapt to an extensive data set of credit card transactions(Ng and Jordan 2002).

Regression General Step

Finding the h-function (i.e., the prediction function) Constructing the predictive function h(x), the logistic function, or also known as the sigmoid function, we generally the first step is to build the predictive process, where the training data for the vector, as well as the best parameters. The basic form of the function shown in figure 1

$$g(z) = \frac{1}{1+e^{-z}}$$

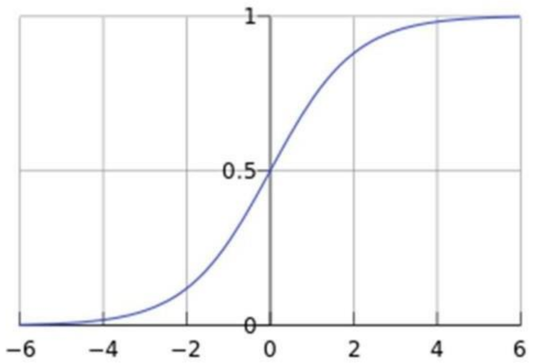


Figure 1. Logical function expression

K- nearest neighbor

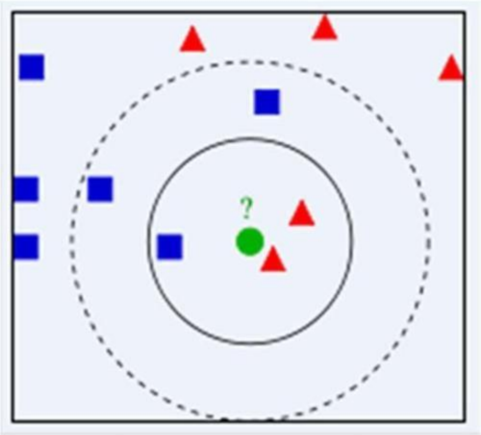
Initially proposed by Cover and Hart in 1968, KNN is a theoretically mature method that is one of the simplest of the data mining classification techniques The term K nearest neighbors mean K nearest neighbors which says that its closest K neighboring values can represent each sample. The nearest neighbor algorithm is a method of classifying every record in a data set.

The implementation principle of KNN nearest neighbor classification algorithm is: to determine the Category of unknown samples by taking all the examples of known types as a reference and at the same time calculate the distance between the new models and all the available pieces, from which the nearest K has known examples are selected, , according to the rule of majority-majority-voting, the unknown samples(Bunsen et al. 2014) and the K nearest models belong to a category with more categories(Duman et al. 2013).

$$d((x_1,\ldots,x_n),(y_1,\ldots,y_n)) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p}$$

The K value of the KNN algorithm in 'scikit-learn' is adjusted by the n neighbors parameter, and the default value is 5.

As shown in the figure below, how do people determine which Category a green circle should belong to, whether it is a red triangle or a blue square? If K=3, the green process will be judged to belong to the red triangle class because the proportion of red triangles is 2/3, and if K =5, the green circle will be considered to belong to the blue square class because the ratio of blue squares is 3/5(Gaikwad et al. 2014).



The k-nearest neighbor sample

IV. RESULT

The dataset of the bank credit card is from kaggle.com. Also, we are pre-processing and feature engineering scales and selects features and uses the smote algorithm (undersampling and downsampling) todeal with theunbalance of thedata set. Then we build an anti-fraud prediction model based on the five algorithms: Logistic regression, KNN, Support vector machine (SVM), Decision tree(DT), Category&boosting(Catboost). The model can predict whether a user has made fraudulent purchases. Then we used a confusion matrix to compare theresults of the two sampling methods. The best solution is logistic regression (undersampling) which is more in line with our expectations. It also achieves an accuracy of 97.00%. Then although credit card spoofing detection, most of the current research is still using decision tree and logistic regression test. But in this project, I think two points where we added SVM and universal algorithm catboost, to make training comparison together. I also believe meaningful results emerged. catboost did not perform poorly, and also we dealt with the sample imbalance problem to get significant marks. Finally, while KNN and catboost perform well, it is also possible to get a better notation if they are trained later on for integration. Secondly, the training of SVM algorithms usually takes a long time, and if we are still increasing the amount of data, we may process the results differently.

V. CONCLUSION

This research is all about studying credit card fraud-detection models based on different machine learning classification algorithms. The goal is to be in this training and testing. To find out the best way to process the dataset and the best machine learning classification algorithm for the dataset of this credit card transaction. So, to achieve this, we chose five different classifiers, respectively. Between them, ten different combinations of algorithms and sampling methods were used to evaluate their predicted performance as a way to get better results for credit card fraud detection. Finally, we cross-validated the technique applied to all the individual classifiers to obtain more accurate results. Logistic regression, as one of the simpler few algorithms, still has their advantages in targeting differential data processing, followed by the SVM algorithm. There is also the catboost algorithm which both perform well.

REFERENCES

1. Aihua, S. et al. 2007. Application of Classification Models on Credit Card Fraud Detection. IEEE.

2. Al Daoud, E. J. I. J. o. C. and Engineering, I. 2019. Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. 13(1), pp. 6-10.

3. Alghamdi, M. et al. 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. 12(7), p. e0179805.

4. Awoyemi, J. O. et al. eds. 2017. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI). IEEE.

5. Bahnsen, A. C. et al. eds. 2014. Improving credit card fraud detection with calibrated probabilities. Proceedings of the 2014 SIAM international conference on data mining. SIAM.

6. Barandela, R. et al. eds. 2004. The imbalanced training sample problem: Under or over sampling? Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR). Springer.

7. Bhatla, T. et al. 2003. Understanding credit card frauds. Cards Business Review# 2003–1.

8. Bhattacharyya, S. et al. 2011. Data mining for credit card fraud: A comparative study. 50(3), pp. 602- 613.

9. Dal Pozzolo, A. et al. eds. 2015. Calibrating probability with undersampling for unbalanced classification. 2015 IEEE Symposium Series on Computational Intelligence. IEEE.

10. Dornadula, V. N. and Geetha, S. J. P. C. S. 2019. Credit Card Fraud Detection using Machine Learning Algorithms. 165, pp. 631-641.

11. Dorogush, A. V. et al. 2018. CatBoost: gradient boosting with categorical features support.

12. Duman, E. et al. eds. 2013. A novel and successful credit card fraud detection system implemented in a turkish bank. 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE.

13. Foulsham, M. 2019. Living with the new general data protection regulation (GDPR).Financial Compliance. Springer, pp. 113-136.

14. Gaikwad, J. R. et al. 2014. Credit Card Fraud Detection using Decision Tree Induction Algorithm. 4(6),

15. Han, H. et al. eds. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International conference on intelligent computing. Springer.

16. Hancock, J. and Khoshgoftaar, T. M. 2020. CatBoost for Big Data: An Interdisciplinary Review.