CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Gopinath A R [1], Sukruth D N [2], Sri ajay S [3], Varunraj P K [4] , Manohar S R [5]

[1] *Associate Professor, Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Bangalore, India*

[2, 3, 4] *B.E. Students, Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Bangalore, India*

Abstract: The use of credit cards to make large-scale online purchases online and systematically increases the number of fraudulent transactions that occur every day. Fraudulent credit card transactions occur intermittently and then cause huge financial difficulties. Criminals can use various methods such as phishing to steal credit card information from strangers, which requires an inappropriate method to detect fraudulent transactions. method is to use historical data transactions, including repeated transactions and fraudulent transactions. This research paper uses machine learning here, which includes supervised learning algorithms such as random forests, decision trees, and logistic regression to identify fraudulent transactions. This data set is collected by Kaggle and contains a total of 2,84,808 transactions. Treat fraudulent transactions as positive and real transactions as negative. Oversampling this to balance the dataset.

Keywords: Random forest, Naïve bayes, Logistic regression.

I. INTRODUCTION

A credit card is a convenient thin plastic card that contains recorded information, including a seal or picture, and allows the person on it to force purchases or assist in personal accounts. Now a days details on the credit card read by automatic teller machine. that as unique agenda cardinal number interrogative of the essential. the having past chronicle for digout allegiant purchaser. money lender dispute of credit mortage agency, credit card issuer, department store and benefit corparation can analysis consumer credit report and history to realize how constant and accountable purchaser in recompense tag end individual debts.

The rapid growth in the number of Mastercard transactions has led to a significant increase in fraud. Credit card fraud is a broad term for theft and fraud, using the primary card as the source of fraudulent funds in certain transactions. Generally, statistical methods and many data processing algorithms cannot solve this fraud detection problem. Most Mastercard's fraud detection systems support artificial intelligence, meta-learning, and pattern matching. Genetic algorithm is an evolutionary algorithm designed to provide the best solution to eliminate fraud. Attaches great importance to the development of efficient and secure electronic payment systems to determine whether transactions are fraudulent. In this investigation, he specializes in MasterCard fraud and its identification measures. Credit card fraud occurs when one person uses another person's card for personal purposes without the owner's knowledge. When scammers start such silent transactions, they will be used until their available limit is exhausted. Therefore, we want to receive a response that minimizes the total available limit of the main card, which is even more important for fraud. Over time, ordinary algorithms will produce better solutions. The development of effective and secure electronic payment systems to detect fraudsters has played a leading role. Precervation of buying credit card it also provide customer supplementary preservation if the being solution setoff past, deface both the client credit card declaration and agency can verify the customer as purchace if earliest receiving is stolen. In today's world, terminal companies are expanding the availability of financial resources by using modern services such as credit cards, ATMs, the Internet, and mobile banking. In addition to quick access to e-commerce, using a credit card has become an essential part of your financial life. Credit card is a settlement card provided to the customers as a system of settlement.

II. LITERATURE REVIEW

1) Previous research has proposed a variety of methods to provide fraud detection solutions through controlled and unsupervised hybrid methods. This requires checking the technologies related to credit card fraud detection and clearly understanding the types of credit card fraud. Over time, fraud models have evolved and new forms of fraud have been introduced, making them an area of great interest to researchers. The rest of this section describes the unique machine learning algorithms, machine learning models, and fraud detection systems used for

fraud detection. The issues raised in the review were analyzed for further use in the implementation of effective machine learning models. When analyzing different detection models, previous researchers found many problems related to fraud detection.

2) They think the lack of real data is a serious problem. Real existence information are missing due to the information sensitivity and private ness issues. They also found it difficult to process classified data. When viewing credit card transaction data, most of the characteristics are classified. In this case, almost all machine learning algorithms do not support rank values.
3) The document [7] emphasizes that the cost and the inability to adapt to fraud detection can cause problems in the fraud detection process. When considering the system, you need to consider the cost of fraud and the cost of prevention. When exposed to new types of fraud and traditional transactions, the algorithm lacks adaptability. Effectiveness may vary depending on the definition of the problem and its characteristics, so it is necessary to have a good understanding of effectiveness measures [4].
4) Various types of models have been implemented to detect credit card fraud. These models use different algorithms. dusting the fraud detection system to deal with emerging fraud may be problematic, regardless of whether retraining the machine learning model will be costly and risky due to the huge changes in the fraud model. For example, Taylor et al. The framework proposed in [12].

III. PROPOSED SYETEM

In the proposed system, we use four algorithms: random forest, logistic regression, naive bayes, and decision tree. Compare between different algorithms based on the accuracy algorithm that selects the best algorithm for the data set.

Decision Tree

It is one of the most widely used predictive modeling methods. As the name suggests, the structure of the model is a tree structure. When there are multiple classes, the model can be used for multivariate analysis. The transferred data, also known as the transfer vector, is used to build a model that can be used to predict the output value based on the input provided. There are several nodes in the tree, and each node corresponds to a specific vector. The tree ends with leaf nodes, and each node represents a possible result or output. Decision tree is the simplest and most popular classification algorithm. In order to build the model, the decision tree algorithm considers all the data features provided and presents important features. Because of this advantage, decision tree algorithms are also used to determine the importance of feature metrics used to select features. After identifying the key features, use the training data to train the model to create a set of rules for predicting future observations or test sets.

Random Forest

In this research work, random forest [20] is regarded as a classifier. The popularity of decision tree models [23] in data mining can be explained by simplified algorithms and flexibility to handle different types of data attributes. However, a single tree model is likely to respond to specific training data and is easy to use. Become the recruitment method can solve these problems by combining multiple point solutions in a certain way, and is more accurate than a single classifier [21] Random forest is one of the ensemble methods. It is a combination of several tree predictors, so each tree depends on a random and independent data set, and all trees in the forest have the same distribution [20]. The capacity of a random forest depends not only on the strength of a tree, but also on the correlation between different trees. The stronger the tree and the lower the correlation between different trees, the better the performance of the random forest. The diversity of trees is due to their randomness, including random selection of original samples and subsets of data attributes. Although our data set may contain some mislabeled instances, random forests are still robust to noise and outliers.

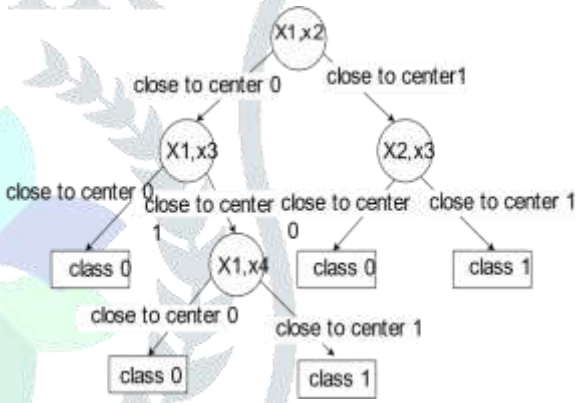

Fig 1: Single tree model

Logistic Regression

Logistic function, also called sigmoid function, was developed by statisticians to describe the characteristics of population growth in terms of ecology, rapid growth, and maximum environmental carrying capacity. Basically, it is a statistical model that uses logistic functions to model binary dependent variables. This model is mainly used in situations where binary classification problems may arise. It is suitable for linearly partitioned classes. Odds ratio is a concept, and we can also use it to define the function of logit. This is the probability of the event occurring. Take input data in the range of [0,1] and convert them to values in the range of real numbers.

Naïve Bayes

Naive Bayes is based on the well-known Bayesian method and has a simple, clear and fast classifier. The naive Bayes classifier is a simple probabilistic classifier based on the application of Bayes' theorem and the assumption of strong (naive) independence. A more descriptive term for basic probabilistic models is "independent feature model". The naive Bayes classifier

assumes that the existence (or non-existence) of a particular function of the class is independent of the existence (or non-existence) of another function of the class variable

IV. SYSTEM DESIGN

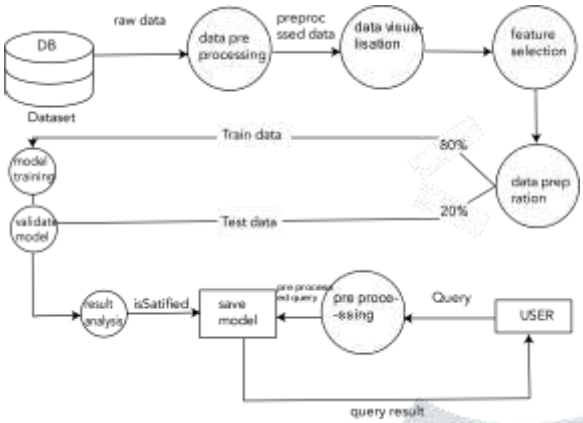

Fig 2: System Block Diagram

1.Data Selection

Data is the most important part of using a forecasting system. It plays a very important role in your entire project, which means your system depends on this data. Therefore, data selection is the first and most important step that needs to be completed correctly. For our project, we received data from the government website. These data sets are available to everyone. There are a large number of other websites that provide this data.

2.Data Cleaning and Data Transformation

After selecting the record. The next step is to clean up the data and convert it to the format you want, because the data set we use may be in a different format. We can also use multiple data sets from different sources, which can be in different file formats. Another reason for deleting data is that records can also contain null and junk values. So the solution to this problem is to replace the garbage value when converting the data. There are many ways to do this.

3.Data Processing and Algorithm Implementation

After the data is cleaned and transformed, it can be processed further. After the data has been deleted and we have accepted the necessary restrictions. We divide the entire data set into two parts, which can be 70-30 or 80-20. Most of the data has been processed. The algorithm is applied to the data element. This helps the algorithm learn on its own and predict future or unknown dates. The algorithm is executed, in which we only obtain the necessary constraints from the original data. The result of the algorithm is "yes" and "no". Provide error rate and success rate.

4.Output and User Side Experience

Once the forecast system is ready for use. The website is designed for users. All the user has to do is fill out a form with various options. They are similar to weather types, car types, etc. After the user submits the form, the algorithm is activated and the data entered by the user is transmitted to the prediction system. The user will be informed of the danger level of the road.

VII. RESULTS

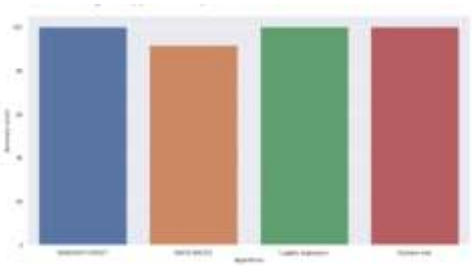

Fig 3: Bar plots of algorithm

Fig 4: Accuracy of algorithms

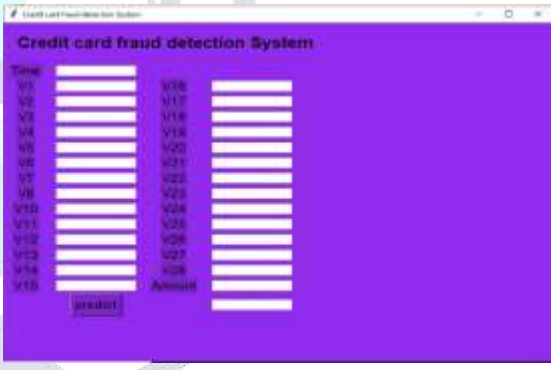

Fig 5: Gui to predict the transaction

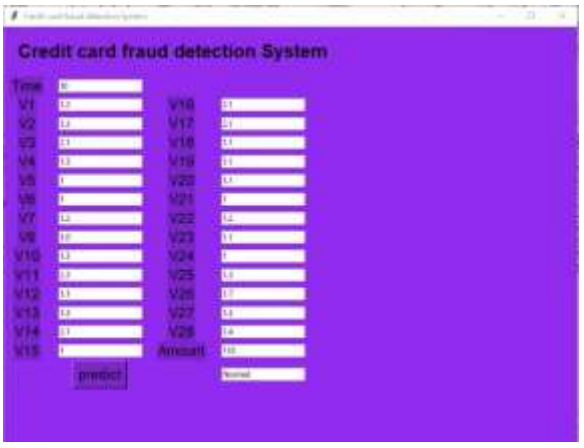

Fig 6: Data entry to predict transaction

VI. CONCLUSION

Credit card fraud is undoubtedly a crime of dishonesty. This article lists the most common scam methods,

detection methods, and outlines the latest findings in the field. This research paper also explains in detail how to apply machine learning to obtain better results in fraud detection, as well as the algorithm, pseudo-code, explanation of its implementation, and experimental results.

Since the entire data set contains only two days of transaction logs, if the project is used for commercial purposes, this is only a small part of the available data. Based on machine learning algorithms, if you input more data, the program will only become more efficient over time.

REFERENCES

[1] NituKumari, S. Kannan, and A. Mutukumaravel, "Credit Card Fraud Detection and Gene A Research", veröffentlicht vom Middle East Journal of Scientific Research, IDOSI Publications, 2014.

[2] Satwik Watts, Surya Kant Dubey, Navin Kumar Pandey, "A Tool for Effectively Detecting Card Fraud", International Communication Security Journal ISSN: 2231-1882, Volume 2, Issue 1, 2013.

[3] Rinky D. Patel and Diraj Kumar Singh, "Genetic Algorithm Detection and Prevention of Credit Card Fraud", herausgegeben vom International Journal of Soft Computing and Engineering (IAOE) ISSN: 2231-2307, Band 2, Ausgabe 6. Januar 2013.

[4] M. Hamdi Ozcelik, EkremDuman, Mine Isik, TugbaCevik, "Using Genetic Algorithms to Enhance Credit Card Fraud Detection", veröffentlicht von der International Conference on Network and Information Technology, 2010.

[5] Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to.

[6] Soltani, N., Akbari, M., Sargolzai Javan, M., "A New User Model for Credit Card Fraud Detection Based on Artificial Immune System", Artificial Intelligence and Signal Processing (AISP), the 16th International Symposium in 2012 CSI, IEEE, page 029-033, 2012.

[7] Wen-Fang Yu, Na Wang, "Investigating a Remote Amount-Based Credit Card Fraud Detection Model", veröffentlicht von der IEEE International Cooperation Conference on Artificial Intelligence, 2009.

[8] Fraud Detection"; Department of Computer Science- Columbia University;2000. 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.

[9] S. Ghosh and D. Reilly, "Using Neural Networks to Detect Card Fraud", Proceedings of the 27th Annual Systems Science Conference, Level 3: Information Systems: DSS/System Knowledge, Seiten 621-630, 1994 Herausgegeben von Presse IEEE Computer Society.

[10] MasoumehZareapoor,Seeja.R,M.Afshar.Alam, Analysis of Credit Card Fraud Detection Methods: Based on Design Standards, International Journal of Computer Applications (0975-8887), Vol. 52–#3, 2012.

[11] Universal protection: a safe anti-fraud strategy,Fair Isaac, http://www.fairsaac.com , 2007.

[12] Fraud Brief – AVS and CVM, Clear CommerceCorporation, 2003, http://www.clearcommerce.com.