International Scientific Journal of Engineering and Management
Volume: 02 Issue: 05 | May – 2023
DOI: 10.55041/ISJEM01101
ISSN: 2583-6129
www.isjem.com
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

A Survey on Credit-Card Fraudulence using Data Mining

Under the Guidance of
Mr. Niraj Kumar Sahu
Assistant Professor
(Department of IT)
n.sahu@ssipmt.com

SSIPMT, Raipur

N. Chitra
Information Technology
SSIPMT. Raipur
Raipur, India
n.chitra@ssipmt.com

Niharika Laheja
Information Techn010kv
SSIPMT. Raipur
Raipur, India
niharika.laheja@ssipmt.com

ABSTRACT-

Fraud detection is a growing concern in various industries, including banking, retail, financial services, and healthcare. It involves implementing measures to prevent the acquisition of money or property through deceptive means. Online fraud detection is particularly challenging due to the constantly evolving tactics employed by scammers. The objective of this study is to explore methods for identifying fraudulent credit card activity that negatively impacts financial institutions.

According to CPO Magazine, in 2017, an average of 4,444 US consumers lost an average of $429 each to credit card fraud. Shockingly, 79% of the affected consumers did not suffer any financial consequences. Consequently, financial institutions bear the burden of these losses. The Federal Trade Commission Report indicates a 44.6% increase in credit card thefts from 2019 to 2020, resulting in approximately $4.444 billion in credit card fraud losses in 2020. To mitigate the impact of credit card fraud, financial institutions should promptly implement technology protection measures and cybersecurity.

To evaluate the effectiveness of different machine learning algorithms in predicting fraudulent transactions, we conducted a comparative analysis and pattern recognition using the credit card fraud dataset. The study trained the algorithms on two resampling methods: under-sampling and oversampling. The goal was to identify the most suitable algorithm for fraud prediction by comparing their performance using metrics such as AUC scores.

Our research revealed that the proposed machine learning algorithms, including random forests, decision trees, XGBoost, K-Means, logistic regression, and neural networks, outperformed previous studies in credit card fraud prediction. Notably, ensemble tree algorithms like Random Forest, Decision Trees, and XGBoost exhibited the highest performance, achieving AUC scores of 1.00% and 0.99% respectively.

The best algorithm identified in this study demonstrated significant improvements, achieving an overall power of % AUC value when utilizing the oversampled dataset.

1. In summary, our research focused on addressing credit card fraud by evaluating various machine learning algorithms. Through comparative analysis, we identified the topperforming algorithms and found that ensemble tree algorithms, particularly Random Forest, Decision Trees, and XGBoost, were the most effective in predicting credit card fraud. These findings highlight the importance of implementing robust fraud detection systems in financial institutions to combat fraudulent activities and protect consumers and the institutions themselves.

1. INTRODUCTION

Cybercrime such as credit card and debit card fraud is on the rise as the world becomes more and more digital. "The Consumer Financial Protection Agency's 2019 report on the consumer credit card market stated, "fraud remains a costly reality in the credit card market." of individuals, public and private organizations. The issue is a little more difficult to handle. International credit card transactions or transactions exceeding certain limits are used to flag some transactions as fraudulent. However, it was also found that 70% of the flagged transactions were false positives, resulting in decrekLsed merchant sales and loss of trust. This study examines methods of detecting credit card fraud and how proposed solutions can help detect this fraud. [1]

CREDIT-CARD FRAUD:

Credit card fraud is one of the biggest threats facing financial institutions and businesses today. Credit card fraud is defined as "an unauthorized person using a credit card for personal use without the cardholder's consent or knowledge and the card issuer not knowing what the card is being used for. Can be defined as "if / Models, processes and precautions to help end credit card fraud and reduce financial risk. High volume transactions on credit card account are matched by financial institutions and businesses. A plastic card called is called a credit card and is issued to various users as one way to make transactions. It allows authorized card users to purchase goods and services based on the cardholder's promise to pay at a later date. Credit cards have become a popular means of personal finance in the last years. Praise and approval rates are Clearly reflected in the number of credit card holders. According to US credit card PUBLISHED ON THE STATIC website, approximately 1.1 trillion credit cards were issued between 2012 and 2018. As Of 2019, Visa is the largest credit card issuer, with more than of his 300 million credit cards issued to customers. Secure Credit Services Financial Institutions and E-Business Development Credit Card Use. Fraud detection 1%Lsed on analysis of existing cardholder purchase data is a promising way to reduce the rate of fraud on credit cards. When a fraudster breaks her anti-fraud rules to initiate fraudulent transactions, the fraud detection system concludes

SIGNIFICANT OF THE STUDY:

The benefit of this research paper is to help financial institutions by improving existing machine learning algorithms that can predict fraudulent actions with very high accuracy. This makes it easier to prevent fraudsters from performing unauthorized transactions or authorized by the rightful owner of the various accounts. Despite the wide scope of the problem, the costs of fraud, the root causes, how and why it occurs. and productive ways to detect, deter, and prevent it. A relatively large amount Of scientific research has been carried out [l l. They. The need for fraud prevention expertise is even more pressing as fraudsters are not reported to public authorities. Organizations must invest significant resources to protect his itself from fraud and reputational damage. This issue is difficult to manage in smaller organizations as they do not have sufficient resources to set up anti-fraud departments. Small businesses wishing to take advantage of specialized expertise in troubleshooting fraud issues should turn to a private research firm, which can be very costly. One Of the Current solutions helping banks and financial institutions move forward is a machine learning approach.

2. LITERATURE REVIEW

Most card users are very aware of the threat of fraud. With this, Card Thief evolves its mode of operation to break the constantly updated Security Wall. Therefore. this aspect briefly describes common patterns of credit card fraud.

(a) Stolen/Misplaced Card: This is the most Common method. This has to do with stealing someone's credit card and using it as their own. In fact, getting information from the front and back Of the card Without stealing it is the same as stealing the cᵛard. Banks typically instruct customers to call and notify them if their card is stolen or misplaced. Thieves can use this information to purchase goods online, and banks may not notify the owner until the end of the month.

(b) Synthetic Fraud: Synthetic fraud is the act of a fraudster applying for her credit card on behalf of another person. The scammer obtains vital information from the victim, such as her social security number (SSN), date of birth, and address, and the applies for credit cards on behalf of the victim. This method is also known as "Unauthorized Application Method".

(c) Data Breach: People make some transactions through her Internet, which makes your data vulnerable to hackers. Hackers can take several routes to get the victim's data. It can even take over someone'S phone or computer completely after visiting some websites. One of the recommended ways to resolve this situation is to avoid storing sensitive information on any device. Alternatively, we recommend wiping your data frequently beforeit falls into the wrong hands.

(d)_E-mail Interception: Fraudsters can also intercept email addressed to users. Perhaps after applying for a new card, the scammer can perform as many as operations before the card reaches the owner. The money would have been gone before the card finally reached its owner.

(e) Skimming: This type of fraud is usually hidden because there is not much money involved. It Can be a few cents. But if that applies to his 444 million customers, that's a lot. Fraudsters can capture and activate card details, such as numbers, so that thieves can receive a fee for each transaction whenever the cardholder makes a transaction through the card. Dealer Collusion: This is a type Of fraud usually run by the organization. A business owner or its employees may use a customer'S credit her card or pass it to a fraudster. Card information may be stored at trusted merchants to facilitate customer purchases, so company owners or employees may not extract some card iniörmation and use it for destructive purposes. may be used in

(f) Triangulation: This is another form of fraud used by scammers to scrape hard-earned money from her. Some products may be listed on the website at a price of only to attract customers. Site owner is for the sole purpose of obtaining customer's card information. In some cases, the scammer may not have the item, but tricks the victim into providing credit card information so that the credit card can be used. The only way around this is to check the authenticity of all her websites and read reviews about [21.

DATASET:

The dataset for this study was provided by Kaggle and generated using Sparkov Data Generation, a GitHub tool authored by Brandon Harris. A record is a simulated credit card transaction that includes both legitimate and fraudulent transactions. This covers the credit cards Of his 1000 customers trading with his pool of 800 merchants. The transactions presented by this dataset totaled 1048575 transactions, and the number of fraudulent transactions was recorded as 6006 of the total number of transactions. The data set is severely imbalanced. The positive class (fraud) accounted for a small percentage ofabout 0.5727 of the total transactions. The record contains 22 characteristics with various data types such as "amount", "category", " is fraud". It also contains both numeric and categorical features. Each transaction recorded by transaction date and time is included in the "trans_date_trans_time" attribute column. The Amount column of function contains the amount of the transaction made. The final feature of this record, the label "is a fraud", is a response variable that indicates whether the transaction is his fraud.

3. METHODOLOGY

DATA PREPROCESSING:

Data preprocessing is an essential step before implementing a machine learning algorithm. It involves cleaning and preparing the data by addressing biases, handling missing values, and resolving discrepancies. The dataset used in this study contains both numerical and categorical data. Therefore, categorical data needs to be encoded into numeric values before being used for modeling. Outlier detection and removal are performed, and feature scaling is applied to ensure that independent variables are in the same space. Additionally, a Box-Cox transformation is performed to reduce feature distortion. To address the imbalance in the dataset, resampling techniques such as undersampling and oversampling are applied. Python libraries, specifically Pandas and scikit-learn, are used for data manipulation and machine learning tasks.

DATA CLEANING:

The credit card records are imported using the Python import command, and data sanitization is carried out during the data cleansing process. This process involves removing nulls and missing values and handling outliers. The dataset used in the study contains a total of 1,048,575 transactions, and no null values or missing values are present. Outliers are identified using a box plot technique, and the interquartile range (IQR) method is used to remove them. The IQR method identifies outliers as data points outside the whiskers of the box plot. By discarding these outliers, the machine learning model becomes more robust and accurate.

ENCODING CATEGORICAL VARIABLES:

After cleaning the dataset, categorical features are converted to numeric values. Most machine learning algorithms perform better with numeric inputs. The One Hot Encoder is used in this study to convert categorical variables to numeric values. For features with two categories, a numerical value of 1 or 0 is assigned.

FEATURE SCALING:

Feature scaling is a preprocessing step used to normalize the ranges of independent variables in the dataset. It ensures that variables with large values do not dominate or distort the machine learning algorithms. The Robust Scaler method, also known as robust standardization, is used for feature scaling. This method calculates the average, 50th percentile, 25th percentile, and 75th percentile of each variable. Then, the median is subtracted from the values for each variable and divided by the interquartile range (IQR), representing the percentile difference between the 75th and 25th percentiles.

DATA RESAMPLING:

Due to the significant imbalance in the dataset, data resampling techniques such as undersampling and oversampling are applied. These techniques help address biases and overfitting in the training model. Undersampling involves randomly removing samples from the majority class to balance the dataset, while oversampling involves replicating samples from the minority class. Both techniques aim to create a more balanced representation of fraudulent and non-fraudulent transactions.

FEATURE CORRELATION AND SELECTION:

Feature correlation analysis is performed to identify relevant features that contribute to the prediction accuracy of the machine learning models. Highly correlated features may be linearly dependent and have a similar impact on the dependent variable. Therefore, if two features are highly correlated, one of them can be omitted. A heat map is generated to visualize the correlation between the original dataset and the resampled datasets (undersampled and oversampled). Feature selection is performed using the LASSO method, which helps minimize the cost function. LASSO regression automatically selects useful features for the model and discards redundant features. In this study, four significant variables were selected for modeling, while the remaining 25 variables were eliminated.

MACHINE LEARNING MODELS:

In this study, both supervised and unsupervised machine learning models were experimented with to classify fraudulent transactions. The specific machine learning models used are described in the subsequent subsections. The process of model creation and selecting hyperparameter values for the best model is also discussed.

DECISION TREE:

The decision tree model is a commonly used algorithm in machine learning. It works by quickly and intelligently analyzing large amounts of data. The decision tree model uses basic root questions and branching to direct transactions based on data characteristics. It recursively partitions the data based on specific parameters until all elements are assigned to a particular class. Decision trees are nonparametric supervised learning methods that can be used for classification and regression tasks. They are robust to outliers, handle missing values automatically, and require less time for training. However, a single decision tree can become complex and prone to overfitting as the dataset size increases.

XGBOOST CLASSIFIER:

XGBoost (eXtreme Gradient Boosting) is an ensemble learning algorithm that implements gradient-enhanced decision trees. It is highly scalable and widely used for regression, classification, and ranking problems. XGBoost provides optimization and processing constraints, and it can handle large datasets efficiently. It supports various interfaces and offers robustness through hyperparameter tuning. XGBoost reduces computation time and improves parallelism for tree building, block structuring, and handling missing values.

K-MEANS CLUSTERING:

K-means clustering is an unsupervised algorithm used for grouping data into clusters. It aims to find k clusters in an unlabeled dataset. The algorithm iteratively assigns data points to the nearest cluster center and recalculates the mean of the clusters. It uses the Euclidean distance to measure the proximity between data points and cluster centers. K-means clustering is widely used for unsupervised machine learning and data mining tasks. However, the standard k-means algorithm may have limitations in terms of clustering efficiency, especially with large datasets.

4. COMPARATIVE ANALYSIS:

In this section, a comparative analysis of our model was made based on the types of datasets and the result of the metrics unsed to measure how each algorithm has performed. Based on the performance Of our model with a different dataset that we

have explored for this study using the AUC score to evaluate the performances and pick the best overall model, we observed that with the original dataset.
under sampling, and oversampling dataset; the ensemble treemodel performed very well rather than other model using the AUC score. the accuracy, precision, recall, and Fl -score to compare between them. shows the results of different model used for this study based on different dataset such as under sampling and oversampling are shown. Comparison was made. In this section, a comparative analysis of our model was made based on the types of datasets and the result of the metrics unsed to measure how each algorithm has performed. Based on the performance Of our model with a different dataset that we have explored for this study using the AUC score to evaluate the performances and pick the best overall model, we observed that with the original dataset. Under sampling, and oversampling dataset; the ensemble tree model performed very well rather than other model using the AUC score. the accuracy, precision, recall, and Fl -score to compare between them. The results of different model used for this study based on different dataset such as under sampling and oversampling are shown. Comparison was made to choose the best predictive model using the AUC Score as metric and comparing the metric With other metrics to further established how good each model has performed. The area under the curve known as (AUC) is the same as the probability that a model will ran randomly chosen positive instance higher than a randomly Chosen negative example. The higher the AUC score, the better the model can predict fraudulent and mm-fraudulenttransactions. When to identify the strength of a model to distinguish between two outcomes of AUC is a useful metric for such identification because it creates a sharp boundary between the positive and negative classes. The results for each classifier are shown below. Confusion matrix for, random forest, Xgboost, and ensemble tree models that are decision trees. From the confusion matrix output of the random forest, we can see that there are 387089 true positive results. This means that random forests could predict him 387089 out of his total number of 386427 transactions used for testing. Transactions can be flagged as fraudulent transactions. A false negative gives a value of O. This meansthat Random Forest did not mistake the fraudulent transaction for his genuine transaction. In this case, the algorithm did not flag any transactions as fraudulent or genuine. 10 Real Transactions Falsely Identified as Fraud

REFERENCES

[11 Aditya Mishra. (2018). Metrics to Evaluate your Machine Learning Algorithm, _https://towardsdatascience.com/metrics-to-evaluateyour- machine learning- algorithm-fl Oba6e38234

[21 Arden Dertat. (2017). Applied Deep Learning— autoencoder. https://towardsdatascience.com/applied-deep learning- part-3- autoencoders lc083af4d798

131 Azhan. Mohd. (2020). Credit Card Fraud Detection using Machine Learning and Deep LearningTechniques. 10.1109/1cßS49785.2020.9316002. [4] Consumer Financial prctection Bureau. (2019). https://www.consumerfinance.govidata research/the-consumer-credit-market 2019/

[5] Dornadula, Vaishnavi & S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. Procedia Computer Science. 165. 631641. 10,1016/j.procs.2020.01.057. Guo G., wang H., Bell D., Bi Y., Greer K. (2003). KNN ModelBased Approach in Classification. In: Meersman R., Tari Z, Schmidt DC. (eds) On the Move to Meaningful Internet Systems 2003: coop1S, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540 39964-3 62

[6] J. O. Awoyemi, A.O. Adetunmbi, and S.A.01uwadare. (2017). "Credit card fraud detection using machine learning techniques: A comparative analysis," 60 International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 1-9, Doil 10.1109/1CCN1.2017.8123782.

[7] Joshi, Aruna & Shirol, Vikram & Jogar, Shrikanth & Naik, Pavankumar & Yaligar,Annapoorna. (2020). Credit Card Fraud Detection Using Machine Learning Techniques. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology. 436442 10.32628/CSEIT2063114.