Credit Card Fraud Detection Using Machine Learning

Shalini Avinashbhai Naik and Dr. Nitin Pise
Department of Computer Science & Technology, MIT-Wpu,
Survey No,124, Paud o Road, Kothrud, Pune, Maharashtra 411038
shailynaik24@gmail.com, nitin.pise@mitwpu.edu.in

Abstract

Due to the pandemic in this time economic scenario increased, credit cards or debit cards use has become extremely commonplace due to online payments. Credit cards enable individuals to make large payments without having to carry huge amounts of money. The number of users is increasing, so credit card fraud also increased as well. The credentials on a credit card can be obtained fraudulently and used to defraud. So we're going to use Machine Learning Algorithms to collect data and overcome this issue. This project compares supervised algorithms like Logistic regression, Support vector machine, KNN, Decision Tree, Xgboost, etc., and finds the best model through hyper parameter tuning, Grid search and applies resampling techniques.

Keywords
Fraud, Machine Learning, Machine Learning Models, Sampling techniques

1. Introduction
An individual commits credit fraud when they use another's identity and creditworthiness to obtain credit or purchase goods and services without the intention of repaying the debt.
Credit cards are scanned at the machine and are used with no permission from their owners. A duplicate card is created when a card is copied using a special swipe machine. Databases or email scams are sold, then they are used over the internet or the phone, a scam known as "card, not present" fraud. Obtain credit cards on someone else's behalf. The increase in fraud is directly due to the advancement of technology and everyday communication across all borders. Fraud has two sides one is prevention and the other is detection. so in this paper detection of fraud is going to learn. Prevention protects attackers as a defense. After prevention has failed, detection occurs. As a result, detection aids in the discovery and notification of fraudulent transactions as soon as they occur. Web payment gateways have recently become popular for card-not-present transactions in credit card operations. “The number of reports of identity theft climbed by 113% between 2019 and 2020, while the number of reports of identity theft using credit cards rose by 44.6%. Of the almost 1.4 million instances of identity theft in 2020, 393,207 involved credit card fraud.As a result, credit card fraud surpassed government documents and benefits fraud as the second most frequent identity theft recorded crime for the year.

According to the nilson report, Card-not-present fraud is on the rise and is to blame for 65% of all fraud-related losses. Credit cards were the most often stolen form of payment, accounting for almost 25% of fraud complaints including payment information. Card fraud losses in the US are anticipated to reach $12.5 billion by 2025. The data shows that 3,432 cases of credit and debit card fraud were reported from all over India in 2021, an increase of almost 20% from the previous year. Such scams surged by more than 70% in 2020. It demonstrated an almost two-fold increase in credit and debit card-related fraud in only two years. However, there has been a significant surge in fraudulent transactions, which has had a significant impact on the economy. [globalnews]”
Credit card fraud includes any type of fraud with a payment card, such as a credit or debit card. Fraud Detection applies to many industries like banking and financial sectors, insurances, government agencies, etc. There 2 types from which fraud can occur.

1. physical card-based purchase in that the user of the card gave his card physically to someone for making a payment of the product which he/she purchased. If he or she does not recognize that a card has been lost, it can result in a large financial loss for them as well as the credit card companies.
2. online payment mode, as we know now- a day we all purchased things online and payment also done by online payment mode. In this mode, attackers need only a little information for fraud or doing fraudulent transactions.

© IEOM Society International

Machine learning is the solution for detecting the issue on large databases which are impossible for humans. Supervised learning and unsupervised learning are the most common techniques of machine learning. A prior diagnosis of anomalies is necessary for supervised learning. Several supervised algorithms have been used to detect credit card fraud in past years.

1.1 Objectives

The objectives of the project is to implement machine learning algorithms to detect credit card fraud detection with respect to time and amount of transaction. The key objective of any credit card fraud detection system is to identify suspicious events and report them to an analyst. And helpful to the bank as well as The cardholder.

2. Literature Review

There are many different approaches are available for the fraud detection. Different authors use different type of approaches. We also refer to various existing systems. Here some of the used methods are listed below for skin cancer with different datasets and different approaches (Table 1).

Table 1. Key Features

Year	Title	Dataset	Model	Accuracy
2020	Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison	http://www.ulb.ac.be/di/map/adalpozz/imbalancedatasets.z	Decision Tree,KNN,Random forest, Logistic Regression,Naïve bayes	They take sensitivity and precision. Didn't mention accuracy
2019	Credit Card Fraud Detection using Machine Learning and Data Science	Kaggle	Local outlier factor, Isolation forest algorithm	99.67%, 99.77%
2021	Credit Card Fraud Detection Using Machine Learning	Kaggle	Decision tree, Random forest, logitstic regression, naïve bayes	91.12%, 96.77%, 95.16%
2019	Real-time Credit Card Fraud Detection Using Machine Learning	Real time data	Svm, Knn, Logistic regression, naïve bayes	91%, 72%, 74%, 83%
2020	Credit Card Fraud Detection Using Machine Learning	Kaggle	Random forest, Adaboost algorithm	Random forest has highest than adaboost algorithm.
2019	Credit Card Fraud Detection - Machine Learning methods	Kaggle	Logistic regression, Naïve bayes, Ranodm forest, Multilayer percepton	97.46%, 99.23%, 99.96%, 99.93%
2021	Credit Card Fraud Detection using Machine Learning	kaggle	Adaboost algorithm,Random Forest,LightGBM	96.13%, 95.95%, 95.78%
2021	Prediction of credit card defaults through data analysis and machine learning techniques	Uci library	KNN, random forest, Logistic regression,Svm,Naïve bayes	79%, 80%, 81%, 82%,

© IEOM Society International

				76%

Samidha Khatri et al. (2020) Credit card information about a specific person might be fraudulently acquired and used for fraudulent transactions. To solve this problem, certain Machine Learning Algorithms can be used to collect data. This study compares three well-known supervised learning techniques for distinguishing between legitimate and fraudulent transactions. They provide precision and sensitivity.

S P Maniraj et al. (2019) On the PCA converted Credit Card Transaction data, we focused on evaluating and pre-processing data sets, as well as applying different anomaly detection techniques such as the Local Outlier Factor and Isolation Forest algorithm.

V.Sellam et al. (2021) For handling the highly imbalanced dataset, this research provides various machine learning-based classification techniques such as logistic regression, random forest, and Naive Bayes. Finally, the accuracy, precision, recall, f1 score, confusion matrix, and Roc-AUC score will be evaluated in this study.

Anuruddha Thennakoon et al. (2019) In this paper, Authors are using real time data and predictive analytics which is performed by ML models and an API module to detect the transaction is fraud or not. We also look at a new technique for dealing with data distribution that is skewed. According to a private disclosure agreement, the data used in our research came from a financial institution.

Ruttala Sailusha et al. (2020) In this research they focused on the random forest algorithm and the Adaboost algorithm are the algorithms employed. The accuracy, precision, recall, and F1-score of the two algorithms are used to compare their result. The confusion matrix is used to plot the ROC curve.

Dejan Varmedja et al. (2019) The Credit Card Fraud Detection dataset was used in this study. Because the dataset was highly imbalanced, the SMOTE technique was used to oversample it. The dataset was divided into two sections: training data and test data. The authors used Logistic Regression, Random Forest, Naive Bayes, and Multilayer Perceptron in this research. The findings show that each algorithm is capable of accurately detecting credit card fraud.

D. Tanouz et al. (2021) They make a graph, often known as a plot, and they analyze it. The model's recall, precision, and accuracy are then determined using three machine learning algorithms: light GBM, Adaboost, and random forest classifier. There's also a function for calculating the time it takes to run various algorithms. Finally, the value produced by these three algorithms is compared to determine which one produces the best result.

Saurabh Arora et al. (2021) They assess the dataset in this study, then do feature selection and apply various machine learning methods.

Hasan I and Rizvi S (2022) In this paper, the authors reviewed some Artificial intelligence and machine learning techniques to reduce fraud detection. They analyzed some techniques for the research challenge and provide the advantages and disadvantages of the techniques. From that, they provide the best techniques for credit card fraud detection.

Hussein, Ameer Saleh et al. (2021) In this paper, the authors used the fuzzy-rough nearest neighbor and sequential minimal optimization as base classifiers. They represent a combination of multiple classifiers through ensemble classifiers. They consider logistic classifiers as an outcome of the predictive model.

Kumar S et al. (2022) In this research, they tried support vector machine to overcome the drawbacks and gave result to detect the fraud using SVM.

3. Data Collection

The dataset is collected from the Kaggle. The dataset provides credit card transactions done by European cardholders in September 2013 and this transaction was done in two days. In this dataset, we have found that there were only 492 frauds cases out of 284,807 transactions that occurred in last two days. The dataset is heavily skewed, with positive class that representing only 0.172 percent of all transactions (Figure 1).

© IEOM Society International

It has only numerical input variables most of these are PCA transformed. So V1; V2; V3; to V28 is the main component obtained using PCA; the only characteristics not changed through PCA are 'Time', 'class' and 'Amount.' The another variable is 'Class,' and it has a value of 1 which indicates frauds and o indicates genuine one. So we have a training dataset for performing the task.

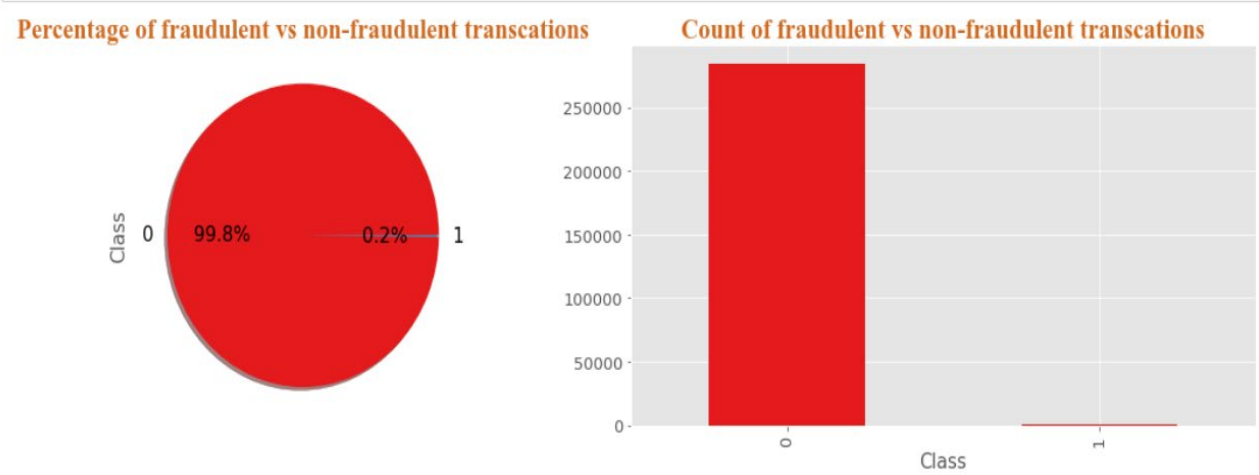

Figure 1. Understanding the dataset

4. Methodology

We are planning to build classification models like Logistic Regression, Random Forest, KNN, Decision Tree, XGboost, etc. And measure their performance with hyperparameter tuning. Building models with an unbalanced dataset as well as after balancing the dataset also build the same model with grid search and choose the best model.
Here, we are going to use 3 resampling techniques for balancing the dataset. The techniques are

1. Random Oversampling
2. SMOTE
3. ADASYN

Here, the dataset is highly imbalanced so we can't rely on the accuracy so we can focus on the sensitivity, precision, ROC - AUC curve, etc.
So firstly validate the dataset into train and test split. After train your model on imbalanced dataset as well as on balancing the dataset. We are going to use resampling techniques to balancing the dataset like undersampling, oversampling, SMOTE and ADASYN. Comparing the models on each and every perspective so that it will be beneficial to the bank as well as that cardholder.

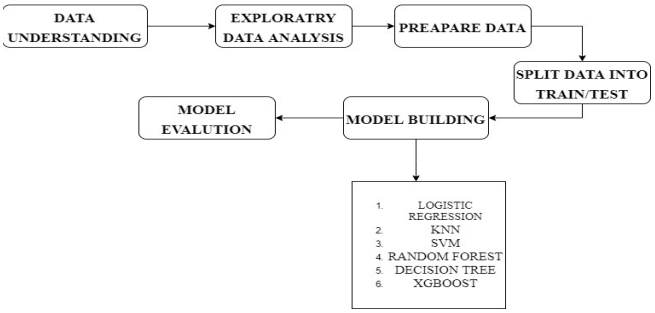

Figure 2. Proposed Workflow

From the above Figure 2, These are the steps e are going to follow.
Data Understanding: - In this, we have to read data properly which features are there and which features we have to take for the project.

© IEOM Society International

Exploratory Data Analysis: - In this section we need to perform some analysis of the data, also performed feature transformation if necessary. Check if there is any skewness in the data.
Prepare Data Prepare the data for the model building. See the variants in data. Check the data is imbalanced or balnced.
Split Data: - Split data into train and test.eg. 80% 20% ratio. Scaling the data using normalization.
Model Building: - Train the model with various machine learning algorithms like, logistic regression, decision tree, Random forest etc.
Model evaluation: - We have seen that data is highly imbalanced so we can't rely on only accuracy, so we have to see balance between recall and precision. We also validate the result with roc auc score.

4.1 Machine Learning Models

4.1.1 Logistic Regression:

Logistic regression is a supervised ML model and widely used in classification problem. This model is best suited for the binary classification problem. It works well with discrete classes. The odds ratio is one concept that can be used to define the logit function. It's the probability of anything occurring.

It takes input in the [0,1] range and converts it to values in the real-number range. The logit function can be defined as follows:

Logit (P) = log

In this model, we can also use a sigmoid function.

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

4.1.2 Decision Tree:

This is one of the most used algorithm for classify the problem and for predicting the model. The model is designed in the form of a tree structure, as the name suggests. In the case of a multi-dimensional study with several classes, this approach can be applied. The past data is used to create a model that predicts the output value based on the input. The tree comes at a leaf node, each of which represents a possible outcome or result.

4.1.3 Random Forest:

This model is an ensemble classifier, because it integrates the results of several decision tree classifiers. The main agenda for using a lot of trees is that if we train lots of trees then each tree's contribution is showing in the form of a model. It makes use of a number of decision trees, which is based on a single dataset with a similar distribution across the tree. This approach is capable of balancing imbalanced data sets in a class population efficiently. It can be used to solve classification as well as regression also

4.1.4 Naïve Bayes:

This algorithm used as probabilistic classifier, which means it can generate predictions for numerous classes at the same time. Classifiers that predict many classes are known as probabilistic classifiers. Conditional probability is used to make the decision. Instead of a single algorithm, this paradigm employs a collection of algorithms, all of which share a similar premise. In this paradigm, each feature is assumed to contribute an equal and distinct amount to the result. Because it simply requires a little quantity of training data, this model has an edge over others.

4.1.5 KNN:

One of the simplest but most effective models is the k-Nearest Neighbor model. The test datasets class label is defined by the class label of adjacent training data components in this model. Euclidean Distance is used to determine how similar the two elements are. It's also known as a lazy model. The value of 'k,' which is the number of closest neighbors that must be considered. For 'k,' a suitable value should be chosen. It's also necessary to use the right distance metric. Sometimes 'Murkowski' distance is used. The Manhattan distance is a generalization of the Euclidean distance and mostly used distance formula is Euclidean. It can be expressed mathematically as:

$$d\left(x^{(i)}, x^{(j)}\right) = \sqrt[p]{\sum_k |x_k^{(i)} - x_k^{(j)}|^p}$$

© IEOM Society International

5. Future Enhancement & Conclusion

In the future enhancement of this project, We will give a solution after the prediction of fraud transaction classifies what can we do further. We also go with real-time data or make real-time fraud detection systems as well. We can also use the deep learning method for better results. We can build an app as well as a website also who detects fraud transactions in real-time.

From the literature, they used an imbalanced dataset to check the accuracy, precision and recall of different machine learning algorithms to predict the fraudulent transaction. But we will use sampling techniques to the balanced dataset. For an imbalanced dataset we can't rely on the accuracy, we have to see precision, recall, F1-score, and roc-AUC curve, etc. From this evaluation, we can easily see which model works best on the imbalanced dataset as well as a balanced dataset.

References

Credit Card Fraud Available: https://en.wikipedia.org/wiki/Credit_card_fraud, Accessed: 6 October, 2021.

About Performance Matrix Available:https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b, Accessed: 6 October, 2021.

Imbalance Technique Available: https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/,Accessed : 7 October, 2021.

About credit card fraud available : https://mint.intuit.com/blog/planning/credit-card-fraud-statistics/, Accessed: 2 september, 2022.

About Performance Matrix Avalible:https://www.tutorialspoint.com/machine_learning_with_python, Accessed: 7 October, 2021

News of Credit Card Fraud Available: https://globalnews.ca/tag/credit-card-fraud/, Accessed: 7 October, 2021.

Samidha Khatri, Aishwarya Arora, Arun Prakash Agrawal, Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison, *10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 978-l-7281-2791-0/20/$31.00* © 2020 IEEE, 2020.

Anuruddha Thennakoon, Chee Bhagyani , Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi, " Real-time Credit Card Fraud Detection Using Machine Learning" , 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 978-1-5386-5933-5/19/$31.00 © 2019 IEEE, 2019.

D. Tanouz ,G V Parameswara ,R Raja Subramanian ,A. Ranjith kumar ,D. Eswar ,CH V N M praneeth, "Credit Card Fraud Detection Using Machine Learning", Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021) IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2, 2021.

Ruttala Sailusha, V. Gnaneswar, R. Ramesh G., Ramakoteswara Rao, "Credit Card Fraud Detection Using Machine Learning", International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2, 2020.

Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, "Credit Card Fraud Detection - Machine Learning methods ", 18th InternationalSymposium INFOTEH-JAHORINA, 20-22 March 2019, 978-1-5386-7073-6/19/$31.00 ©2019 IEEE, 2019.

Eyad Btoush, Xujuan Zhou, Rai Gururaian,KC Chan,XiaoHui Tao, "A Survey on Credit Card Fraud Detection Techniques in Banking Industry for Cyber Security", 2021 8th International Conference on Behavioral and Social Computing (BESC),2021, IEEE,ISBN: 978-1-6654-0023-7, 2021.

Hussein, Ameer Saleh; Khairy, Rihab Salah; Najeeb, Shaima Miqdad Mohamed; ALRikabi, Haider Th. Salim, "Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression", International Journal of Interactive Mobile Technologies. 2021, Vol. 15 Issue 5, p24-42. 19p, 2021.

Kumar S., Gunjan V.K., Ansari M.D., Pathak R., "Credit Card Fraud Detection Using Support Vector Machine",2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications, vol 237. Springer, Singapore, https://doi.org/10.1007/978-981-16-6407-6_3, 2020.

Srinidhi S., Sowmya K., Karthika S., "Automatic Credit Fraud Detection Using Ensemble Model", ICT Analysis and Applications. Lecture Notes in Networks and Systems, vol 314. Springer, Singapore. https://doi.org/10.1007/978-981-16-5655-2_20, 2018.

Hasan I., Rizvi S. "AI-Driven Fraud Detection and Mitigation in e-Commerce Transactions", Data Analytics and Management. Lecture Notes on Data Engineering and Communications Technologies, vol 90. Springer, Singapore. https://doi.org/10.1007/978-981-16-6289-8_34, 2018.

© IEOM Society International

S P Maniraj, Aditya Saini, Swarna Deep Sarkar, Shadab Ahmed, "Credit Card Fraud Detection using Machine Learning and Data Science", International Journal of Engineering Research & Technology (IJERT),ISSN: 2278-0181 IJERTV8IS090031,Vol. 8 Issue 09, September, 2019.

V.Sellam, P.Tushar, G.Rohit, S.Sanyam, "Credit Card Fraud Detection using Machine Learning ", Indian Journal of Computer Graphics and Multimedia (IJCGM) ISSN: 2582-8592 (Online), Volume-1, Issue-1 February, 2021.

Ruttala Sailusha, V. Gnaneswar, R. Ramesh G., Ramakoteswara Rao, "Credit Card Fraud Detection Using Machine Learning", International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2, 2020.

Saurabh Arora, Sushant Bindra, Survesh Singh, Vinay Kumar Nassa, "Prediction of credit card defaults through data analysis and machine learning techniques", scientific committee of the 1st International Conference on Computations in Materials and Applied Engineering – 2021.

Hussein, Ameer Saleh; Khairy, Rihab Salah; Najeeb, Shaima Miqdad Mohamed; ALRikabi, Haider Th. Salim, "Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression", International Journal of Interactive Mobile Technologies. 2021, Vol. 15 Issue 5, p24-42. 19p, 2021.

Kumar S., Gunjan V.K., Ansari M.D., Pathak R., "Credit Card Fraud Detection Using Support Vector Machine",2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications, vol 237. Springer, Singapore, https://doi.org/10.1007/978-981-16-6407-6_3, 2021.

Biographies

Shalini Naik is now a second year M.Tech. Data Science and Analysis Student at Dr. Vishwanath Karad MIT World Peace University, Pune and graduated with BE. In Computer science & engineering from R.N.G.Patel institute of technology,Bardoli, Gujarat, India. She is doing internship as a data analyst.

Dr. Nitin N. Pise is currently working as Professor in School of Computer Engineering and Technology at Dr. Vishwanath Karad MIT World Peace University, Pune. He has done B.E. (Computer Engineering) and M.E. (Computer Science and Engineering) from Walchand College of Engineering, Sangli. He has received his Ph.D. in Computer Engineering from College of Engineering, Pune, Savitribai Phule Pune University in 2016. Currently, he is guiding three Ph.D. scholars and has published sixty-two research publication in national, international conferences and journals. He has total twenty-five years of teaching experience and two years industry experience. His areas of research as Artificial Intelligence, Machine Learning, Data Science and Cyber Security.

© IEOM Society International