Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.13, No.2, 2022

Design of a Model in Machine Learning For Credit Card Fraud Detection

Abderrahim Bourchouq, Wang Peiliang
Engineering College of Huzhou Normal University
E-mail: abdobour97@gmail.com

Abstract
In e-commerce, fraud has become a major problem, and much effort is being invested into identifying and preventing it. Currently, fraud detection and prevention systems are only able to detect and prevent a small percentage of fraudulent transactions, resulting in billions of dollars in losses. Because online transactions are likely to grow dramatically in the coming year, better fraud detection and prevention is critical. We provide a data-driven strategy for estimating the likelihood of a fraudulent or legal transaction based on machine learning techniques applied to large data sets. To predict the likelihood of a customer's next transaction being fraudulent, the algorithm was trained using past e-commerce credit card transaction data. Random Forest, Support Vector Machine, Gradient Boost, and combinations of these are used to compare the performance of supervised machine learning approaches. The problem of class imbalance is considered before the model is trained on a classifier, and methods such as oversampling and data pre-processing are employed.

Keywords: machine learning, model, fraud, credit card

DOI: 10.7176/CEIS/13-2-06
Publication date: April 30th 2022

1.0 Introduction

Credit card fraud detection is a prominent problem in finance research, with considerable economic implications. While traditional data analysis methods have been employed in the past, the similarities between this and other difficulties, such as the development of recommendation systems and diagnostic/prognostic medical tools, suggest that a complex network approach could give considerable benefits. It is feasible to find unauthorized instances in a real card transaction data set. It's based on a recently developed network reconstruction method that enables the creation of representations of a single instance's divergence from a reference group. As a result, credit card companies must be able to detect fraudulent credit card transactions so that customers are not charged for items they did not purchase. This is a list of fraudulent transaction statistics in China. When compared to 2017, the number of prosecutions for funding fraud increased by about 8.17 percent in 2019. The number of financial fraud prosecutions accepted by Chinese courts decreased between 2016 and 2020. The term "financial fraud" refers to crimes committed with the use of credit cards, insurance, securities, or other financial instruments. As part of the synthesis project in Big Data and Statistical Learning, we did the following study on one of today's most pressing banking issues. As a result, every bank and financial services provider in China and abroad is working to protect credit card operations to assure a more stable and trusted transaction system. Because fraud can be exceedingly damaging to both clients and service providers, fraud detection is a critical topic to work on. Engineers interested in finance may find that constructing fraud detection models is one of our daily project duties, resulting in the subject's striking aforementioned desire. Assume you've been engaged to assist a credit card company in detecting possible fraud cases so that customers aren't charged for items they didn't purchase. You're given a dataset containing transactions between people, as well as information on whether or not they're fake, and you have to tell them apart. This is the case we're going to deal with. Our ultimate goal is to solve this problem by creating classification models that can classify and differentiate fraud transactions.

1.2 Motivation for the Paper

From bridges to satellites, all modern engineering products require some form of monitoring to ensure that they function as intended. This monitoring could be required for safety reasons, or it could be used to improve operating efficiency by reducing energy consumption or assisting with maintenance planning to optimize system usage. Traditionally, someone with particular training has performed this monitoring. This permits them to examine the system and make health assessments. This could be for a variety of reasons in a variety of situations. Even a seasoned professional may not be able to fully comprehend the system. The average time between failures may be so long that having a specialist on hand to keep track of it isn't cost-effective. There's a chance that no specialists

are available. The system may be too new for the expert to have gained sufficient experience with it. The risk of a diagnosing error cannot be accepted since the implications of failure are so severe. Complex monitoring and diagnostic systems are required because the business environment may be too hostile for human examination

2.0 Review of Literatures

2.1 Review Other Models

Detecting fraudulent transactions can be done in a variety of ways. It's exceedingly tough to detect fraud, and it's only possible after it's happened. Because fraudulent transactions are minor in comparison to total transactions, this is the case. In their paper [8, the authors compared seven methods for detecting such transactions. ANN had the best results across the board, with an accuracy of 99.71 percent, a detection rate of 99.68 percent, and a false alarm rate of 0.12%. ANN is the most efficient, even though it takes the most time and processing power to train. SVM has a maximum false alarm rate of 5.2% and a detection rate of 85.45%, which is not comparable to other superior techniques. Fuzzy logic has the lowest detection rate, at 77.8 percent. With an accuracy of 97.93 percent, a detection rate of 98.52 percent, and a false alarm rate of 2.19 percent, decision trees are weighted in favor of complexity during training. Random forest is a categorical and numerical data decision tree regression and classification tool [6]. The authors employed a random forest and an SVM classifier to detect fraudulent transactions in the dataset. Pre-processing was done to avoid missing data and scale feature values. The authors discovered that SVM did not perform well with imbalanced data when compared to random forest classifiers. Another benefit of using the random forest technique was that, because it used a subset of data with different decision trees, adding more data points did not affect the model. Because each tree has a very small chance of influencing others, bias and overfitting are reduced. Random forest classification was used by Mohankumar and Karuppasamy [15] to detect fraudulent credit card transactions. The PCA approach was used to mask the values in the dataset. Feature values were adjusted to reduce volatility across characteristics. The SMOTE algorithm was used to balance the data. In the balanced data, there are 175000 classes. A random forest classifier is used to categorize data points into binary categories. According to the paper's findings, the precision-recall curve has an analogous value of around 0.85. The random forest classifier has been one of the most widely used techniques in e-commerce to detect credit card frost due to its flexibility and scalability for large datasets. The computational resources required to train the random forest model are modest as compared to superior state-of-the-art techniques like ANN. ANN is not commonly employed in real-time e-commerce applications due to computational and time constraints. There is a significant problem with a class imbalance in the existing datasets that mislead research [17]. In their research, the authors looked at data balance for efficient analysis, regression, and classification challenges. The primary methodologies they looked into were random oversampling and undersampling, statistical oversampling and undersampling, SMOTE, Feature Selection, Hybrid Sampling, Cost-effective Learning, and Ensemble Learning. The SMOTE approach, as well as feature selection, were found to be commonly used in research papers [9, 14, 16]. These two tactics produce the best results when it comes to balancing obstacles in data analysis.

Razooqi et al. [19] advised using fuzzy logic to change weights to use a genetic algorithm with ANN, which led to even better results and a lower FN rate. Although the amount of time spent training increased drastically, the end result for ANN was far better. Maes et al. [12] created a Bayesian network for anticipating financial data labels. Even with small datasets, the model provided good results and utilizing ANN to alter the network's parameters cut training time in half. Shirgave et al. [20] introduced a supervised learning random forest to classify alerts as bogus or authentic, paving the way for the implementation of a semi-supervised machine learning method to classify alerts. According to Lakshmi and Kavila [10], a random forest classifier surpasses choice trees and logistic regression in terms of accuracy, with 90.0, 94.3, and 95.5 for logistic regression, decision tree, and random forest classifier, respectively. The random forest classifier outperforms logistic regression and decision trees in a comparison of the three approaches. Furthermore, as Havarapu Bhanusri et al. [2] point out, we can't discern the difference between fraudulent and lawful transactions using machine learning approaches for the current dataset, necessitating additional research. The choice of algorithms for a credit card fraud detection system, according to Sorournejad et al. [21], should minimize False Positive and False Negative rates while maximizing True Positive and True Negative rates and assuring a respectable detection rate. Integrating Genetic Algorithms (GAs) with ANNs for credit card fraud detection can improve engine performance, according to Carsten [3].

2.3 Payments Fraud

In the United States and around the world, payment fraud is a critical and growing problem. In the United States, fraud on non-cash payment systems totaled more than $8 billion in 2015, up 37% from 2012 [18]. Financial institutions and payment operators are increasingly relying on machine learning algorithms to build efficient and effective fraud detection systems [24]. I implement and analyze the performance of various machine learning

models, such as logistic regression, random forests, and neural networks, using a big dataset from Kaggle. Around 300,000 credit card transactions were recorded across Europe over two days.

99.8% of transactions in the Kaggle dataset are regarded as genuine, whereas 0.2 percent are declared fraudulent. As a result of class imbalance, standard models may struggle to distinguish between the dominant and minority classes [3]. As part of this work, I analyze and assess various approaches to dealing with this problem, including sampling techniques such as under-sampling the majority class and over-sampling the minority class. Due to privacy concerns, it is impossible to construct meaningful correlations between most of the variables in the dataset. As a result, rather than inference, my research focuses on performance prediction.

Machine learning has a long history of being used to detect fraud in the payments industry. While banks and payment companies regularly use fraud algorithms, Bhattacharyya et al. point out that there are few research on the use of machine learning techniques for payment fraud detection [4], partially due to the sensitive nature of the data. According to their findings, random forests, despite being underutilized, may outperform more known approaches. The most critical aspect in a neural network's fraud classification ability, according to Roy et al. [5, is network size. According to Chaudhary et al., no single method is optimum across all performance parameters; each has its own set of strengths and limitations [6]. There is a lot of machine learning research on class imbalance. Some of the tactics used to address this problem include oversampling the minority class, undersampling the majority class, creating synthetic minority samples, and changing the relative cost of misclassifying the minority and majority classes. According to Japkowicz and Stephen, the effects of class imbalance vary on the degree of imbalance, sample size, and classifier used, with sampling procedures hurting the performance of particular classifiers [7]. Oversampling can lead to overfitting, and synthetic data creation adds noise that can reduce prediction performance, according to Cui et al. [8].

For a variety of reasons, detecting credit card theft is difficult. One of these is the likelihood of changing buying patterns over time. Because the approaches to effectively learn on the training sample may not correlate to the distribution of the testing set, the monitoring system may need adaptation as a result of this dataset shift or concept drift. Concept drift and dataset shift have previously been discussed in the literature. Others attempted to adapt to concept drift, while others attempted to define it. Abdallah & al. [9] define idea drift as a phenomenon in which the underlying model (or concept) evolves: buying behavior may change over time, and fraudsters' techniques may alter. Given the features of the testing set transactions, the decision function learned on the training set may become out-of-date and no longer represent the conditional distribution of the target variable (y) (X). They propose two approaches to dealing with this issue: the developing approach involves learners who keep up with the data stream, and the regulated approach involves detecting concept drift and intervening to change the hybrid technique when it occurs (usually retraining the learner).

2.3 Neural networks

Neural networks are based on the human brain, and their ability to learn has aided scientists and engineers in solving a range of problems. For identifying credit card fraud, C. Wang et al. [8] proposed a whale method based on neural networks. [10] Presented a cost-sensitive neural network-based approach for detecting credit card fraud. [16] utilized the Support Vector Machine to detect credit card fraud. To divide data into many categories or clusters, classification and clustering procedures are used. These techniques can be used to a wide range of problems. [5] identified credit card fraud using a meta classifier on a big dataset. To detect credit card fraud, [13] employed a system based on partitioning and clustering methods. [14] employed skewed data and a variety of classification algorithms to detect credit card fraud. [17] proposed a credit card fraud detection algorithm based on a classification algorithm.

Machine learning algorithms are artificial intelligence (AI) technologies that are used to solve problems involving massive amounts of data in a range of industries. Several research employed machine learning [4][6][7][22][15] and deep learning [23][15] techniques to detect credit card fraud. However, further research is needed, as is the application of machine learning algorithms to detect fraud in credit card transactions. Here are a few examples of how machine learning algorithms can be used.

3.0 Model Development

3.1 Approaches: *Explanatory Data Analysis*

Logistic Regression is a supervised classification approach that is used to model the likelihood that Y belongs to a specific category. The method of logistic regression is used to estimate discrete values from a set of independent variables. It assists you in predicting the likelihood of an event occurring by fitting data to a valid function. Its output value is between 0 and 1 because it forecasts probability.

Linear Discriminant Analysis (LDA) is a dimensional reduction technique for supervised classification problems. It's used to represent group differences, such as separating two or more classes.

K Nearest Neighbors (KNN) is a simple classification technique that identifies the query's closest neighbors and uses those neighbors to determine the query's class.

Support A discriminate classifier explicitly described by a separating hyperactive plane is known as a vector classifier. In other words, the algorithm produces an ideal hyperplane that categorizes fresh samples given labeled training data (supervised learning).

The Random Forest Classifier is a classification approach that employs the following five steps: (1) Choose "k" features at random from a total of "m" features (where k m). (2) Using the optimal split point, calculate the node "d" among the "k" characteristics. (3) Using the best split, split the node into daughter nodes. (4) Repeat steps 1–3 until the "l" number of nodes is attained. (5) Build a forest by repeating steps 1 through 4 "n" times to create "n" trees.

3.2 Fraud Detection Framework

The architecture of the proposed methodology is depicted in Figure 1. As the first step in the Normalize Inputs block, the training dataset is normalized using the min-max scaling strategy in Equation (4) [31]. The scaling ensures that all of the input values fall inside a particular range. In the GA Feature Selection block, the normalized data from the Normalize Inputs block is used to implement the GA method. For each iteration of the GA Feature Selection block, the GA offers a candidate attribute vector vn, which is utilized to train the models in the Training block, which is represented by the Training data and Train the model's blocks.

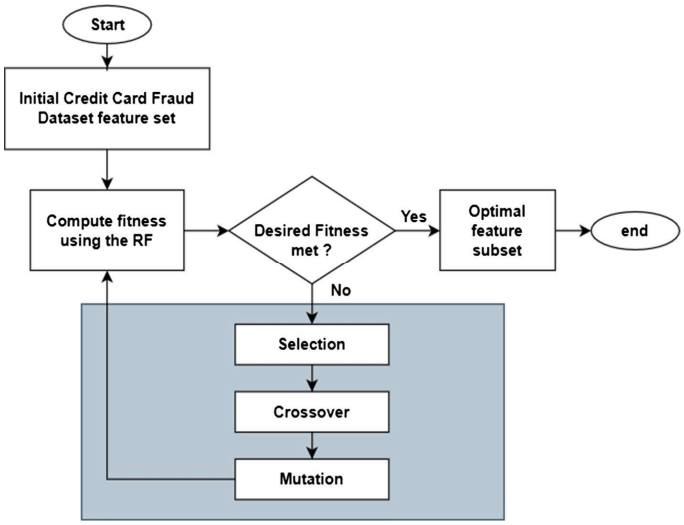
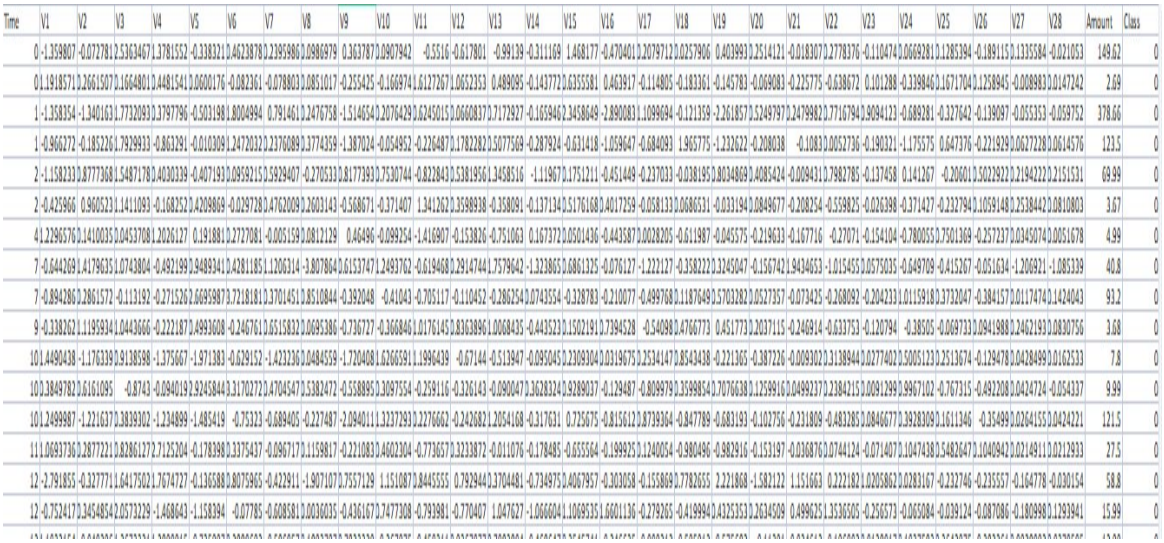


Figure 1 Flow chat of the detections System

3.3 Dataset

The credit card fraud transaction data set we're using comes from Online, and it contains 28315 transaction details, 0.5 fractions of them were found to be fake. The goal is to develop a classifier model that can be used to detect anomalies. The data collection contains only numerical input after PCA transformation. The major parts are features V1, V2, and V28, except for 'Time' and 'Amount,' which have not been adjusted using PCA.' 'Class,' the response variable, has a value of 1 when there is a scam and 0 if there's not.



Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	-1.359807	-0.072781	2.536346	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	-0.5516	-0.617801	-0.99139	-0.311169	1.468177	-0.470401	0.207971	0.025796	0.403993	0.251412	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.08236	-0.078803	0.085102	-0.25542	-0.16697	1.612727	1.065235	0.489095	-0.143772	0.635558	0.463917	-0.11481	-0.18336	-0.14578	-0.06908	-0.22577	-0.63867	0.101288	-0.33984	0.167170	0.125895	-0.00898	0.014724	2.69	0
1	-1.35835	-1.34016	1.773209	0.379779	-0.50319	0.800499	0.791461	0.247676	-1.51465	0.207643	0.624501	0.066084	0.717293	-0.16595	2.345865	-2.89008	1.109968	-0.12136	-2.26186	0.524979	0.247998	0.771679	0.909412	-0.68928	-0.32764	-0.13909	-0.05535	-0.05975	378.66	0
1	-0.96627	-0.18522	1.792993	-0.86329	-0.01031	1.247203	0.237608	0.377436	-1.38702	-0.05495	-0.22647	0.178228	2.507757	-0.28792	0.631418	-1.05965	-0.68409	1.965775	-1.23262	-0.20804	-0.1083	0.005274	-0.19032	-1.17558	0.647376	-0.22193	0.062723	0.061458	123.5	0
2	-1.15823	0.877736	1.548718	0.403034	-0.40719	0.095921	0.592941	-0.27053	0.817739	0.753074	-0.82284	0.538196	1.345852	-1.11967	0.175121	-0.45145	-0.23703	-0.03819	0.803486	0.408542	-0.00943	0.798278	-0.13746	0.141267	-0.20601	0.502292	0.219422	0.215153	69.99	0
2	-0.42597	0.960523	1.141109	-0.16825	0.420987	-0.02973	0.476201	0.260314	-0.56867	-0.37141	1.341262	0.359894	-0.35809	-0.13713	0.517617	0.401736	-0.05813	0.068653	-0.03319	0.084967	-0.20825	-0.55983	-0.02639	-0.37143	-0.23279	0.105915	0.253844	0.081080	3.67	0
4	1.229657	0.141003	0.045370	1.202612	0.191881	0.272708	-0.00515	0.081213	0.46496	-0.09935	-1.41690	-0.15382	-0.75106	0.167372	0.050143	-0.44358	0.002820	-0.61198	-0.04557	-0.21963	-0.16771	-0.27071	-0.15410	-0.78005	0.750136	-0.25723	0.034507	0.005167	4.99	0
7	-0.64426	1.417963	1.074380	-0.49219	0.948934	0.428118	1.120631	-3.80786	0.615374	1.249376	-0.61946	2.291474	1.757964	-1.32386	0.686132	-0.07613	-1.22212	-0.35822	0.324504	-0.15674	1.943465	-1.01545	0.057503	-0.64970	-0.41526	-0.05163	-1.20692	-1.08533	40.8	0
7	-0.89428	0.286157	2.113192	-0.27152	0.668598	0.721818	1.370145	1.851084	-0.39204	-0.41043	-0.70511	-0.11045	-0.28625	0.074355	0.328783	-0.21007	-0.49976	0.118764	0.570328	0.052735	-0.07342	-0.26809	-0.20423	0.011159	0.373204	-0.38415	0.011747	0.142404	93.2	0
9	-0.33826	1.119593	0.044366	-0.22218	0.499360	-0.24676	0.651583	0.069538	-0.73672	-0.36684	1.017614	0.836398	1.006843	-0.44352	0.150219	0.739452	-0.54098	0.476677	0.451773	0.203711	-0.24691	-0.63375	-0.12078	-0.38505	-0.06973	0.094198	0.246219	0.083075	3.68	0
10	1.449043	-1.17633	0.913859	-1.37566	-1.97138	-0.62915	-1.42323	0.048455	-1.72040	1.626659	1.199643	-0.67144	-0.51394	-0.09504	0.230930	0.031967	0.253414	0.854343	-0.22136	-0.38722	-0.00893	0.313894	0.027740	0.500512	0.251367	-0.12947	0.042849	0.016253	7.8	0
10	0.384978	2.616109	5	-0.8743	-0.09401	2.924584	3.170272	0.704547	0.538247	2	-0.55889	0.309755	-0.25911	6	-0.32614	3	-0.09004	7	0.362832	4	0.928903	7	-0.12948	7	-0.80997	0.359985	0.707063	8	0.125991	0.049823	7	0.238421	5	0.009129	0.996710	2	-0.76731	5	-0.48220	8	0.042472	4	-0.05433	7	9.99	0
10	1.249998	-1.22163	0.383930	2	-1.23489	9	-1.48541	9	-0.75323	-0.68940	5	-0.22748	7	-2.09401	1.323729	3	2.227666	2	1.242682	1.205416	8	-0.31763	1	0.725675	-0.81561	2	0.873936	4	-0.84778	9	-0.68319	3	-0.10275	6	-0.23180	9	-0.48328	5	0.084667	7	0.392830	9	0.161134	6	-0.35498	9	0.026415	5	0.042422	1	121.5	0
11	1.069373	0.287722	1	0.828612	7	2.712320	4	-0.17839	8	0.337543	7	-0.09671	7	0.115981	7	-0.22108	3	0.460230	4	-0.77365	7	0.323387	2	-0.01107	6	-0.17848	5	-0.65556	4	-0.19992	5	0.124005	4	-0.98046	-0.98291	6	-0.15319	7	-0.03687	6	0.074412	4	-0.07140	7	0.104743	8	0.548264	7	0.104094	2	0.021491	1	0.021293	3	27.5	0
12	-2.79185	-0.32777	1	1.641752	1	2.767472	7	-0.13638	8	0.807596	5	-0.42291	1	-1.90710	7	0.755712	9	1.151087	2	0.844555	5	0.792944	3	0.370448	1	-0.73497	5	0.406795	7	-0.30305	8	-0.15586	9	0.778265	5	2.221868	-1.58212	2	1.151663	0.222182	1	0.020586	2	0.028316	7	-0.23274	6	-0.23555	7	-0.16477	8	-0.03015	4	58.8	0
12	-0.75241	7	0.345485	4	2.057322	9	-1.46864	3	-1.15839	4	-0.07785	-0.60858	1	0.003603	5	-0.43616	7	0.747730	8	-0.79398	1	-0.77004	7	1.047627	-1.06660	4	1.108935	1	0.660113	6	-0.27926	5	-0.41999	4	0.432535	3	0.263450	9	0.489625	1	0.353650	5	-0.25657	3	-0.06508	4	-0.03912	4	-0.08708	6	-0.18099	8	0.129394	1	15.99	0

Figure 2 Data set Extraction from sources

Machine learning and data mining techniques have been widely used to detect credit card fraud. On the other hand, purchase behavior and fraudulent tactics might change over time. This phenomenon is known as dataset shift [11] or concept drift [12] in the realm of fraud detection. In this study, we present a method for measuring the dataset shift at our face-to-face credit card transactions dataset (cardholder in the shop) day by day. In practice, we compare and contrast the days and assess the classification's usefulness. The more precise the classification, the more varied the buying behavior across time, and vice versa. As a consequence, a distance matrix describing the dataset shift is obtained. After agglomerative clustering of the distance matrix, we notice that the dataset shift pattern matches the calendar events for this time period (holidays, weekends, etc). The credit card fraud detection challenge then incorporates this dataset shift information as a new feature. As a consequence, detection has improved slightly.

Detecting fraudulent transactions in bank accounts is the main issue that our fraud detection model may help us with. The banking sector is being beset by a significant fraud problem. This is a prevalent problem. Internet fraudsters have mastered the skill of hijacking online sessions by obtaining client credentials and employing malware to steal cash from unwary account holders, and no country is fully secure from them. To train the system, you require well-labeled data in supervised learning. It means that some data has already been labeled with the proper response. It's similar to learning that takes place in the presence of a teacher or supervisor. Unexpected data outputs can be predicted using a supervised learning system, which learns from labeled training data. To be more specific, classification algorithms. However, I will conduct data analysis by: processing, solving classification issues using XGBoost, Random forest, KNN, Logistic regression, SVM, and Decision tree, and selecting the most efficient model for the job at hand. Six distinct classification models may be built: Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Although there are many more models that may be employed, these are the most widely used models for classification problems. All of these models may be built using the approaches provided by the packages. For the XGBoost model, we'll merely utilize the xgboost package.

Credit card fraud detection is the most prevalent, and therefore the most expensive, problem in the modern world. The rise in both the number of online transactions and the number of e-commerce platforms is to blame. Credit card fraud happens when a credit card is stolen and used for an illegal purpose, or when a fraudster uses credit card information for his gain. Credit card problems are becoming increasingly prevalent in our culture. In order to detect fraudulent activities, a credit card fraud detection system has to be implemented. The focus of this study will be on machine learning techniques and their applications. The algorithms used were the random forest algorithm and the Adaboost algorithm. The performance of the two algorithms is assessed using their accuracy, precision, recall, and F1-score. As a starting point, the confusion matrix is used to create the ROC curve. Several algorithms are examined, including the Random Forest and the Adaboost, and the algorithm with the best accuracy, precision, recall, and F1-score is chosen as the best approach for identifying fraud.

3.4 Data Processing Stages

Data Cleaning - One of the most significant tasks in the data cleaning process is to fill in the missing information. There are various techniques to address this problem, such as disregarding the complete tuple, but most of them are likely to skew the data. Filling them was no longer a problem since the source file, which included legitimate transactions, did not contain any entries with missing data. Tuples with no meaning were eliminated from the files since they do not contribute to the production of useful data and do not skew the data. In addition, adjustments such as deleting redundant columns and splitting the date-time column into two were made.

Data Integration - Because the fraudulent and authentic record files were in two different files, the two data sources were combined before they were exposed to further alteration.

Data Transformation - Here, all of the category data was condensed into a numerical representation that could be understood. The transactional dataset includes a variety of data kinds and ranges. As a result, data transformation includes data normalization. Data normalization reduces the numeric range of attribute data to a manageable size.

Data Reduction - Dimension reduction is the approach utilized for this. We must avoid the possibility of learning incorrect data patterns, and the chosen characteristics must remove the fraud domain's irrelevant elements and attributes [10]. PCA stands for principal component analysis, and it is a well-known transform technique. From the standpoint of numerical analysis, this strategy addresses the feature selection problem. By determining the appropriate number of principal components, PCA was able to effectively execute feature selection.

4.0 Model Evaluation

In terms of effectiveness, some algorithms exceed others. However, to create a more appropriate and acceptable model, it is deemed vital, to begin with, a data pre-analysis method. To note, if we had the option to work with the anonymized data, our model would surely have more full knowledge of the credit card fraud issue. Financial parties can't do their analyses using their unique private information for the customers' privacy, so better knowledge of the data factors would be much appreciated and will assist in the process of classifying the various input variables. Exploratory Data Analysis (EDA) is a vital element in the analytics process as it helps us to comprehend our data. It is vital to study data collection before doing formal analysis. Without a competent EDA, no initial models should be produced. This will help us to better evaluate the patterns in the data, reveal outliers or odd events, and discover intriguing links between variables, among other things. In our instance, the EDA will aid us in better defining our data and preparing it for statistical learning approaches such as those outlined in the preview poster.

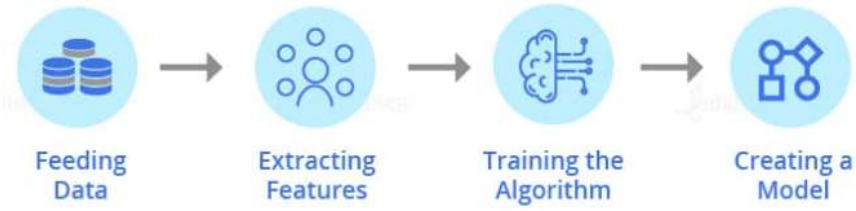


Figure 3 Component of the System and machine Learning Training Model

This paper presents an approach for identifying fraud at the interface. For fraud detection, the recommended solution utilizes unequal, substantially skewed transactions and a modeling methodology. The dataset for credit card fraud detection utilized here is the machine learning Kaggle dataset, which features highly skewed data. The qualities that are assessed are 1 for fraud and 0 for non-fraud. In the banking business, fraud detection analysis was a vital device. Artificial neural networks are presently the least effective technology for identifying credit card fraud. Problems in a way and excessive false positives plague the present technique for identifying fraud. In such circumstances, this research study leverages the collaboration of fully convolutional units to develop a model for spotting credit card fraud that is exceptionally accurate.

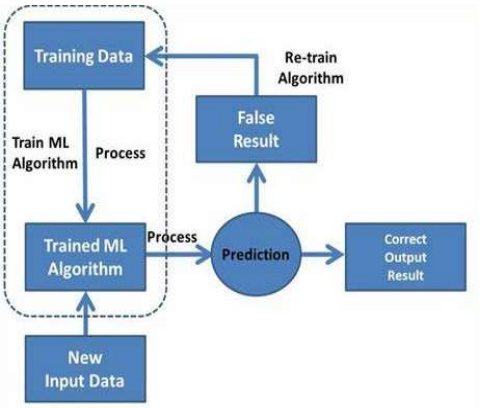


Figure 4: Flow chart of Training Model

Furthermore, for conclusions to be generalizable to the target population, the evaluation prior fraud probability must mirror the population's prior fraud probability (the naturally occurring prior fraud probability).

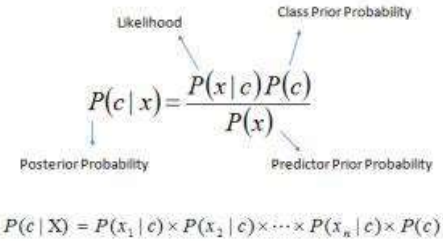


Figure 5 Predictions equations

Credit card fraud is common and causes substantial financial losses [1]. The number of online transactions has exploded, with online credit card transactions accounting for a significant share of them. As a result, credit card fraud detection software is highly valued by banks and financial organizations. Fraudulent transactions can take many different shapes and fall into several different categories. This research looks at four forms of fraud that occur in real-world transactions. Each fraud is treated using a series of machine learning models, with the best answer picked after a thorough examination. This evaluation provides a step-by-step guide to selecting a successful algorithm depending on the kind of fraud, as well as a relevant performance measure. Another crucial feature of our project is real-time credit card fraud detection. To do so, we use machine learning models and an API module to do predictive analytics to assess if a transaction is legitimate or not. We also look at a novel method for coping with skewed data distribution. The data we used in our study came from a financial institution, according to a confidential disclosure agreement.

Credit cards are frequently used for online banking. In recent years, there have been several reports of credit card fraud. Fraud done using a credit card is notoriously difficult to spot and prosecute. Machine Learning (ML) is a type of Artificial Intelligence (AI) that is used in research and engineering to solve several problems. In this research, machine learning algorithms are applied to a data set of credit card frauds, and the performance of three machine learning approaches for detecting credit card thefts is compared. The Random Forest machine learning method has the best accuracy when compared to the Decision Tree and XGBOOST approaches.

5.0 Conclusion

To summarize, the work on this project has been a challenging step forward in my grasp of large data and statistical learning. I got the excellent chance to learn about a variety of data approaches, including Exploratory Data Analysis and Principal Components Analysis, as well as create seven classification algorithms from which we identified which one was the most reliable in terms of classification forecast. Furthermore, tree-based algorithms appear to be better suited to our Credit Card Fraud detection problem. Furthermore, more insightful and enhanced datasets will help our research in a deeper investigation of the problem and more thorough characterization of the target demographic. However, due to the wide range of data that may be employed, resolving the class disparity is a difficult problem. One such example is fraud. There is no one-size-fits-all solution to class disparity; instead,

several alternatives should be tested to find the optimal one. Furthermore, because certain algorithms take a long time to execute, it may be essential to make concessions and give up some data, just as people do. Finally, while we concentrated on Supervised Learning, other areas such as Unsupervised Learning or even Deep Learning may be effective in tackling these challenges and identifying fraud.

References

[1] Artificial Neural Networks- Encyclopedia of Physical Science and Technology https://www.academia.edu/15726358/Artificial_ Neural_Networks, Last Visited, 2021.

[2] Bhanusri A., Valli K., Jyothi P., Sai G., Rohith R., and Subash S., "Credit Card Fraud Detection Using Machine Learning Algorithms," *Journal of Research in Humanities and Social Science*, vol. 8, no. 2, pp. 04-11, 2020.

[3] Classification Accuracy is Not Enough: More Performance Measures You Can Use, Machine Learning Mastery. https://machinelearningmastery.com/classificatio n-accuracy-is-not-enough-more-performance measures-you-can-use, Last Visited, 2021.

[4] Credit Card Fraud Detection: Anonymized credit card transactions labeled as fraudulent or genuine, https://www.kaggle.com/mlgulb/creditcardfraud, Last Visited, 2021.

[5] Devi M., Janani B., Gayathri S., and Indira N., "Credit Card Fraud Detection using Random Forest Technique," *International Research Journal of Engineering and Technology*, vol. 06, no. 3, pp. 6662-6666, 2019.

[6] Entropy: How Decision Trees Make Decisions, https://towardsdatascience.com/entropy-howdecision-trees-make-decisions-2946b9c18c8, Last Visited, 2021.

[7] Jain N., Tiwari N., Dubey S., and Jain S., "A Comparative Analysis of Various Credit Card Fraud Detection Techniques," *International Journal of Recent Technology and Engineering*, vol. 7, no. 5S2, pp. 402-407, 2019.

[8] Kalra M. and Patni J., "Playing Doom with Deep Reinforcement Learning," *International Journal of Computer Applications*, vol. 1, pp.14-20, 2019.

[9] Lakshmi S. and Kavila S., "Machine Learning for Credit Card Fraud Detection System," *International Journal of Applied Engineering Research*, vol. 13, no. 24, pp. 16819-16824, 2018.

[10] Logistic Regression, https://en.wikipedia.org/wiki/Logistic_regression, Last Visited, 2021.

[11] Mishra P., Patel V., Mittal P., and Patni J., "Algorithm Analysis Tool Based on Execution Time Input Instance-based Runtime Performance Benchmarking," *International Journal of Computer Applications*, pp. 27-30, 2018.

[12] Mohankumar B. and Karuppasamy K., "Credit Card Fraud Detection Using Random Forest Technique," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 8, no. 4, pp. 4128-4135, 2019.

[13] Patni J., Billus S., Billus S., and Singh R., "Feature-Based Opinion Mining and Managed Machine Learning with Sentimental Classification Models," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 2, pp. 3992-3998, 2020.

[14] Pavithra P. and Babu S., "Data Mining Techniques for Handling Imbalanced Datasets: A Review," *International Journal of Scientific Research and Engineering Development*, vol. 2, no. 3, 2018.

[15] Racz A., Bajusz D., and Heberger K., "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification," *Molecules*, vol. 26, no. 4, 2021.

[16] Razoogi T., Khurana P., Raahemifar K., and Abhari A., "Credit Card Fraud Detection Using Fuzzy Logic and Neural Networks," *in Proceedings of the 19th Communications and Networking Symposium*, San Diego, pp. 1-5, 2016.

[17] Shirgave S., Awati C., More R., and Patil S., "A Review on Credit Card Fraud Detection Using Machine Learning," *International Journal of Scientific and Technology Research*, vol. 8, no. 10, pp. 1217-1220, 2019.

[18] Sorournejad S., Atani Z., and Monadjemi A., "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective," https://arxiv.org/abs/1611.06439, Cornell University, Last Visited, 2021.

[19] Support-Vector Machine, https://en.wikipedia.org/wiki/Supportvector_mac hine, Last Visited, 2021.

[20] Uqaili I. and Ahsan S., "Machine Learning-Based Prediction of Complex Bugs in Source Code," *The International Arab Journal of Information Technology*, vol. 17, no. 1, pp. 26-37, 2020.

[21] Board of Governors of the Federal Reserve System. *Federal Reserve Payments Study finds U.S. payments fraud a small but growing fraction of overall payments*. 2018. Web. 11 June 2019.

[22] Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., and Beling, P. Deep learning detecting fraud in credit card transactions. Systems and Information Engineering Design Symposium (SIEDS) (2018), pp. 129–134.

[23] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss is based on the effective number of samples. CoRR, abs/1901.05555, 2019.

[24] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015.