International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 05 | May -2020 ISSN: 2582-3930

A Survey on Credit Card Fraud Detection Using Machine Learning Algorithm

Ms. Hemlata S. Dahake[1]
Department of Computer Science and Engineering
Priyadarshini Bhagwati College Of Engineering, Nagpur
Nagpur, Maharastra

Mrs. Archana A. Nikose[2]
Department of Computer Science and Engineering
Priyadarshini Bhagwati College Of Engineering, Nagpur
Nagpur, Maharastra

Abstract---- Credit card payment has become very popular today. Credit card is an easiest way to pay directly through your bank account. But we all know that everything have some pros as well as some cons. In the case of credit card, fraudsters are the main intruder. These intruders can access some unauthorised transactions. It is very important to prevent your account transaction from these intruders. In this paper we used three different classification algorithms (Isolation forest, LOF, SVM) for fraud detection. In this regard, implementation of efficient fraud detection algorithms using machine-learning techniques, and to assist fraud investigators. we use SMOTE sampling method. The problem of ever-changing fraud patterns is considered with employing incremental learning of selected ML algorithms in experiments In this paper isolation forest, based machine learning approach is utilized to detect credit card fraud. The results show isolation forest based approaches outperforms with and it the highest accuracy can be effectively used for fraud investigators.

Keywords- Machine learning; Isolation forest, Local Outlier Factor (LOF) Algorithm, SVM.

----------***----------

1. INTRODUCTION

Credit-card fraud is a general term for the unauthorized use of funds in a transaction typically by means of a credit or debit card. Incidents of fraud have increased significantly in recent years with the rising popularity of online shopping and e-commerce. Credit-card fraud can be classified into two different types, card-not-present fraud and card-present fraud. Card-not-present fraud takes place when a customer's card details including card number, expiration date, and card verification- code (CVC) are compromised and then used without physically presenting a credit card to a vendor, such as online transactions. Card-present fraud happen when credit card information is stolen directly from a physical credit card. Since 2015, credit card companies have issued chip-payment (EMV) cards to combat card-present fraud. Although this measure has been effective at reducing point-of-sale fraud by 28% within the last three years, card-not-present fraud has risen by 106%, increasing the need for online security to prevent data breaches. Although less than 0.1% of all credit card transactions are fraudulent, analysts predict that credit card fraud losses incurred by banks and credit-card companies can surpass $12 billion in the United States in 2020. Evidently, there is a direful need for robust detection of card-present and card-not-present fraudulent transactions to minimize monetary losses.

Currently, credit-card companies attempt to predict the legitimacy of purchase through analyzing anomalies in various fields such as purchase location, transaction amount, and user purchase history. However, with the recent increases in cases of credit card fraud it is crucial for credit card companies to optimize their algorithmic solutions. This paper compares various machine learning algoritham and regression algorithmic models to reseaech which algorithm and combination of factors provide the most accurate method of classifying a credit-card transaction as fraudulent or non-fraudulent (normal).

2. LITERATURE SURVEY

- In this paper, S.P. Manirajan describes Random forest algorithm applicable on Find fraud detection. Random forest has two types. They describe in detail and their accuracy of 91.96% and 96.77% respectively. These paper summaries the second type is better than the first type.
- Suman Arora, In this paper, many supervised machine learning algorithms apply on 70% training and 30% testing dataset. Random forest, stacking classifier, XGB classifier, SVM, Decision tree and KNN algorithms compare each other i.e. 94.59%, 95.27%, 94.59%, 93.24%, 90.87%, 90.54% and 94.25% respectively. Summaries of this paper, SVM has the highest ranking with 0.5360 FPR, and the stacking classifier has the lowest ranking with 0.0335.
- Kosemani Temitayo Hafiz, In this paper, they describe flow chart of fraud detection process. i.e. data Acquisition, data pre-processing, Explorative data analysis, and methods or algorithms are in detail. Algorithms are K- nearest neighbor (KNN), random tree and Logistic regression accuracy are 96.91%, 94.32%, 57.73%, and 98.24% respectively

2. PROPOSED SYSTEM

- The proposed model is introduced to overcome all the disadvantages that arises in the existing system.
- This system will increase the accuracy of the classification results by classifying the data based

on the attacks and others using naive-bayes classification algorithm.

- It enhances the performance of the overall classification results.

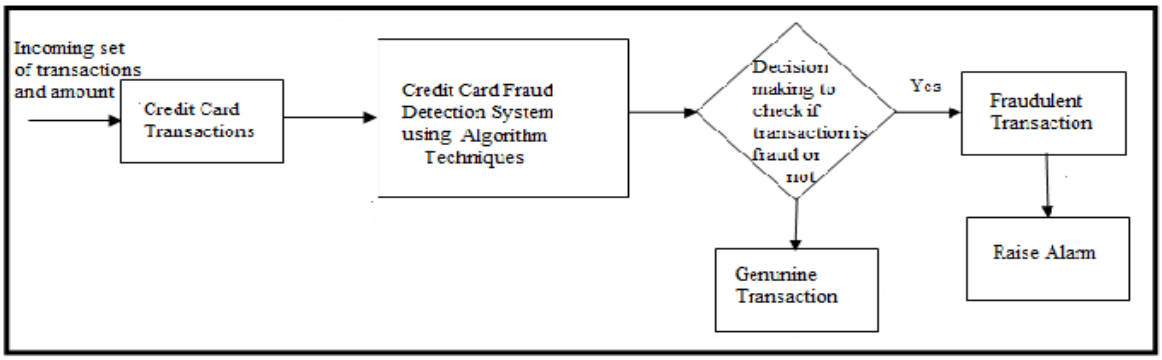


Fig 3: Block Diagram of Credit Card Fraud Detection System

3. METHODOLOGY

1. This work focusing on an application which is use to detect the fraudulent credit card activities on internet transaction. In this peculiar type, the pattern of current fraudulent usage of the credit card has been analysed with the previous transaction. By using the BNN in algorithm of machine learning algorithm.
2. In credit card fraud detection train an auto encoder neural network (BNN) (implemented in keras) in unsupervised or semi-supervised machine learning for anomaly detection .
3. The train model will be evaluated on pre label an anomalymized data set.
    i. Will be using:
    ii. Tenser flow
    iii. Keras

4. IMPLEMENTATION

About dataset

The dataset contains transactions made by credit card in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 490 frauds out of 284,805 transactions. The dataset is highly unbalanced, positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the output of a PCA transformation. Unfortunately, due to confidentiality issues, we can't provide the perfect features and more background information about the dataset. Features V1, V2, V3 ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between all of transaction and the first transaction in the dataset. The feature 'Amount' is the Amount of the transaction, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the reaction variable and it takes value 1 in case of fraud and 0 otherwise.

Machine Learning-Based Approaches

Below is a brief overview of popular machine learning-based techniques for anomaly detection.

a) **Density-Based Anomaly Detection**

Density-based anomaly detection is based on the k-nearest neighbours algorithm. Assumption: Normal data points occur around a dense neighbourhood and abnormalities are far away.

The nearest set of data points are evaluated using a score, which could be Euclidian distance or a similar measure dependent on the type of the data (categorical or numerical). They could be broadly classified into two algorithms:

K-nearest neighbor: k-NN is a simple, non-parametric lazy learning technique used to classify data based on similarities in distance metrics such as Euclidian, Manhattan, Minkowski, or Hamming distance.
Relative density of data: This is better known as local outlier factor (LOF). This concept is based on a distance metric called reachability distance.

b) **Clustering-Based Anomaly Detection**

Clustering is one of the most popular concepts in the domain of unsupervised learning.

Assumption: Data points that are similar tend to belong to similar groups or clusters, as determined by their distance from local centroids.
K-means is a widely used clustering algorithm. It creates 'k' similar clusters of data points. Data instances that fall outside of these groups could potentially be marked as anomalies.

c) **Support Vector Machine-Based Anomaly Detection**

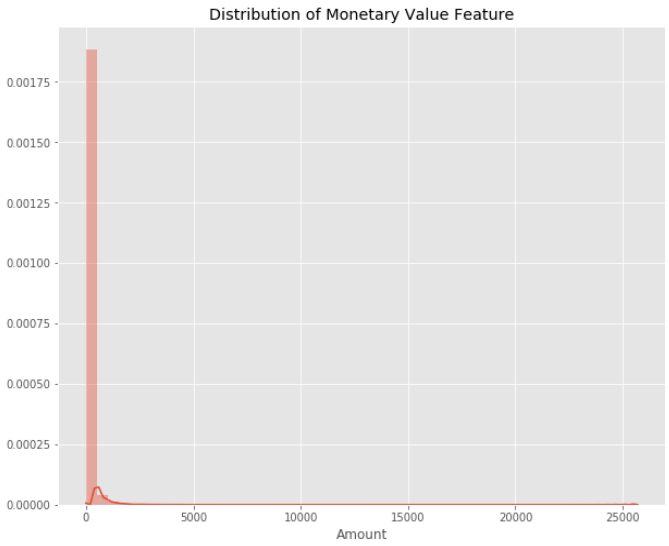
- A support vector machine is another effective technique for detecting anomalies.
- A SVM is typically associated with supervised learning, but there are extensions (One Class CVM, for instance) that can be used to identify anomalies as an unsupervised problem (in which training data are not labeled).
- The algorithm learns a soft boundary in order to cluster the normal data instances using the training set, and then, using the testing instance, it tunes itself to identify the abnormalities that fall outside the learned region.
- Depending on the use case, the output of an anomaly detector could be numeric scalar values for filtering on domain-specific thresholds or textual labels (such as binary/multi labels).

- In this jupyter notebook we are going to take the credit card fraud detection as the case study for understanding this concept in detail using the following Anomaly Detection Techniques namely

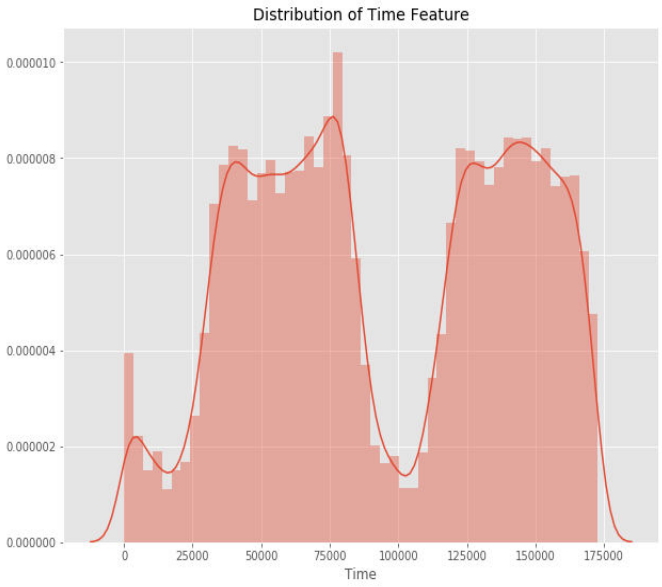
- **Isolation Forest Anomaly Detection Algorithm.**
- **Density-Based Anomaly Detection (Local Outlier Factor) Algorithm.**
- **Support Vector Machine Anomaly Detection Algorithm**

Exploratory Data Analysis (EDA)

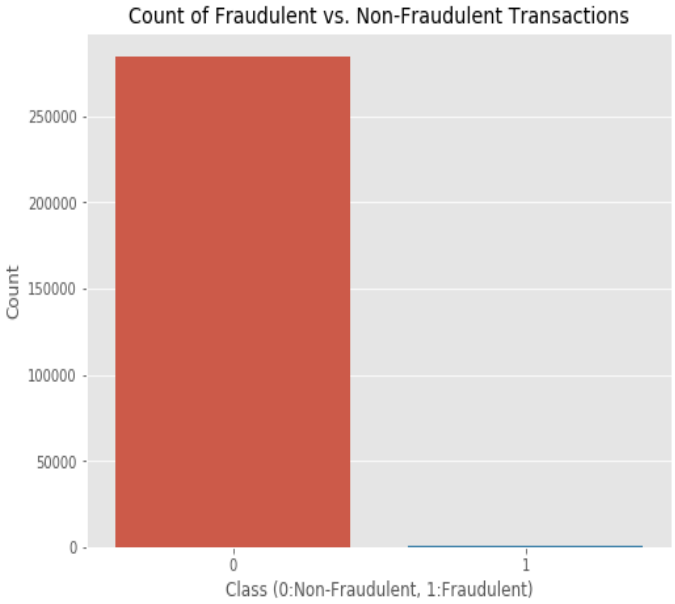
Since nearly all predictors have been anonymized, I decided to focus on the non-anonymized predictors time and amount of the transaction during my EDA. The data set contains 284,807 transactions. The mean value of all transactions is $88.35 while the largest transaction recorded in this data set amounts to $25,691.16. However, as you might be guessing right now based on the mean and maximum, the distribution of the monetary value of all transactions is heavily right-skewed. The vast majority of transactions are relatively small and only a tiny fraction of transactions comes even close to the maximum.



The time is recorded in the number of seconds since the first transaction in the data set. Therefore, we can conclude that this data set includes all transactions recorded over the course of two days. As opposed to the distribution of the monetary value of the transactions, it is bimodal. This indicates that approximately 28 hours after the first transaction there was a significant drop in the volume of transactions. While the time of the first transaction is not provided, it would be reasonable to assume that the drop-in volume occurred during the night.

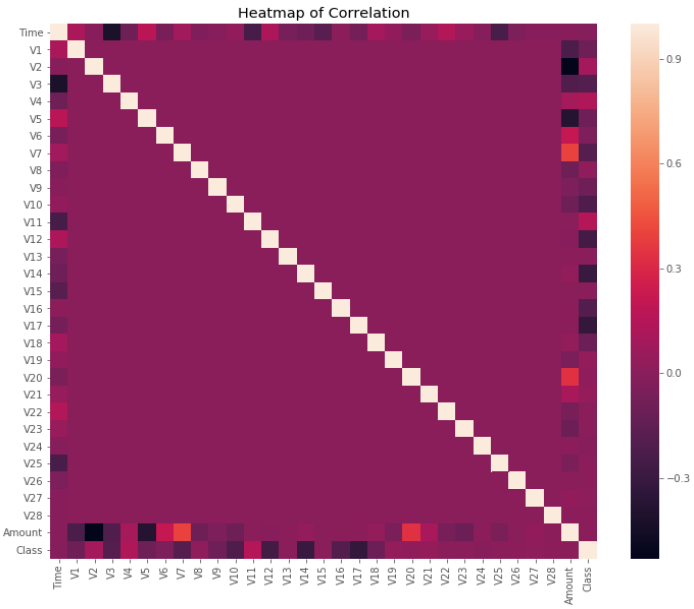


What about the class distributions? How many transactions are fraudulent and how many are not? Well, as can be expected, most transactions are non-fraudulent. In fact, 99.83% of the transactions in this data set were not fraudulent while only 0.17% were fraudulent. The following visualization underlines this significant contrast.



Finally, it would be interesting to know if there are any significant correlations between our predictors, especially with regards to our class variable. One of the most visually appealing ways to determine that is by using a heat map.

As you can see, some of our predictors do seem to be correlated with the class variable. Nonetheless, there seem to be relatively little significant correlations for such a big number of variables. This can probably be attributed to two factors.



1. The data was prepared using a PCA, therefore our predictors are principal components.

2. The huge class imbalance might distort the importance of certain correlations with regards to our class variable.

4. Conclusions

Fraud detection is a complex issue that requires a substantial amount of planning before throwing machine learning algorithms at it. Nonetheless, it is also an application of data science and machine learning for the good, which makes sure that the customer's money is safe and not easily tampered with.

Future work will include a comprehensive tuning of the Random Forest algorithm I talked about earlier. Having a data set with non-anonymized features would make this particularly interesting as outputting the feature importance would enable one to see what specific factors are most important for detecting fraudulent transactions. As always, if you have any questions or found mistakes, please do not hesitate to reach out to me. A link to the notebook with my code is provided at the beginning of this article.

5.References

1. R. M. jamail esmaily, "Intrusion detection system based on multilayer perceptron neural networks and decision tree," in International conference on Information and Knowledge Technology, 2015.
2. J. K. T. J. C. W. Siddhatha Bhattacharya, "Data Mining for credit card fraud: A comparative study," Elsevire, vol. 50, no. 3, pp. 602-613, 2011.
3. Raghavendra Patidar and Lokesh Sharma International Journal of soft computing and engineering, vol. 1, no. NCAI2011, 2011.
4. S.P. Tanmay kumar behera, "credit card fraud detection: a hybrid approach using fuzzy clustering and neural network," in international conference on advances in computing and communication Engineering, 2015.
5. N. W. Wen -Fang Yu, "Research on credit card fraud detection model based on distance sum," in International joint conference on artificial intelligence, Hainan Island, China, 2009.
6. S. k. A. K. M. Ayushi agarwal, "Credit card fraud detection: A case study," in IEEE, New Delhi, India, 2015.
7. K. T. B. V. Sam Maes, "Credit cards fraud detection using bayesian and neural networks," p. 7, August 2002.
8. P. K. D. K. R. D. A. A. Thuraya Razoogi, Credit card fraud detection using fuzzy logic and neural networks, Society for modelling and simulation International(SCS), 2016.
9. L.J.P. van der Maaten and G.E. Hinton, Visualizing High-Dimensional Data Using t-SNE (2014), Journal of Machine Learning Research
10. Machine Learning Group — ULB, Credit Card Fraud Detection (2018), Kaggle
11. Nathalie Japkowicz, Learning from Imbalanced Data Sets: A Comparison of Various Strategies (2000), AAAI Technical Report WS-00–05