CREDIT CARD FRAUD DETECTION USING SUPERVISED AND UNSUPERVISED LEARNING

Vikas Thammanna Gowda

Department of Electrical Engineering and Computer Science,
Wichita State University, Kansas, USA

ABSTRACT

In the present monetary situation, credit card use has gotten normal. These cards allow the user to make payments online and even in person. Online payments are very convenient, but it comes with its own risk of fraud. With the expanding number of credit card users, frauds are also expanding at the same rate. Some machine learning algorithms can be applied to tackle this problem. In this paper an evaluation of supervised and unsupervised machine learning algorithms has been presented for credit card fraud detection.

KEYWORDS

Credit card fraud detection, Supervised learning, Unsupervised learning.

1. INTRODUCTION

With the increase in internet usage, online shopping has become a trend and growing rapidly as it has become a one stop place for shoppers' diverse purchase list. There are over 1.9 billion online shoppers worldwide and USA alone has over 240 million shoppers. According to US Census Bureau News, in the third quarter of 2020 there has been an increase of 36% in online sales when compared to that in 2019. Debit card or credit card is used as the main mode of payment for online sales which has led to a raise in frauds. According to a report by Shift CC Processing, credit card frauds has resulted in $24.26 billion in 2018 and US leads as the most credit card fraud prone country with 38.6% of reported credit card frauds. It is necessary to support the payment systems with an efficient fraud detection capability to minimize unwanted adversary activities.

Credit card fraud detection is based on analysis of a card's spending behaviours and identifying their transactions into fraudulent and legitimate transactions. Various difficulties are related with credit card fraud detection: (1) fraudulent behaviour profiles are dynamic in nature that is fraudulent transactions generally appear as though genuine ones; (2) credit card transaction datasets are rarely available due to privacy and security concerns and the accessible datasets are profoundly imbalanced; (3) optimal feature selection for the models; (4) suitable metric to evaluate performance of models on skewed credit card fraud data. Many techniques have been applied to credit card fraud detection such as artificial neural network [1], genetic algorithms [2,3], frequent item set mining [4], decision trees [5], migrating birds optimization algorithm [6], naïve Bayes [7].

The objective of this paper is to evaluate an imbalanced dataset based on few performance parameters using supervised machine learning (a. *Logistic Regression*, b. *Support Vector Machine* (SVM), c. *Random Forest*) and unsupervised machine learning (a. *Isolation Forest*, b.

David C. Wyld et al. (Eds): CMC, NCO, SOFT, CDKP, MLT, ICAITA - 2021
pp. 93-99, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111107

Local Outlier Factor, c. *k-Means*) algorithms and determine the best algorithm for credit card fraud detection. The rest of the paper is organised as following. A brief literature review is done in section 2. In section 3, we study the fundamentals of the algorithms used in this paper. Section 4 deals with the parameters used in this paper to determine the best algorithm. Experimental results are analysed in section 5 and concluded in section 6.

2. RELATED WORK

In recent days credit card fraud detection has drawn a lot of research interest and several techniques and strategies for detection. The work in [8] gives an exhaustive discussion on the difficulties and issues of fraud detection research. Mohammad et. al., [9] inspected the most well-known sorts of credit card fraud and the current nature-inspired detection strategies that are utilized in detection methods. A detailed comparison is made between decision tree and support vector machine by Sahin and Duman [10] in detecting credit card fraud. They divide the entire dataset into three groups which differ in ratio between fraudulent transactions and legitimate ones and develop a series of seven decision tree and SVM based models. The experimental results indicate that decision tree-based model is better than SVM model.

In 2019, Naik et. al., [11] have used naïve Bayes, logistic regression J48 and adaboost algorithms for credit card fraud detection and observed that the highest accuracy is obtained for both adaboost and logistic regression algorithms. Since both the algorithms had the same accuracy, time factor was taken into consideration to determine that adaboost algorithm works well to detect credit card fraud.

Sailusha et. al., [12] compares random forest and adaboost algorithms as machine learning techniques for credit card fraud detection. Both the algorithms have same accuracy but when precision, recall and F1 scores are considered, the random forest algorithm has the highest value than adaboost algorithm. Lorenzo et. al., [13] have used isolation forest and local outlier factor for anomaly detection. It works better on unlabelled dataset. The algorithm allows avoiding the subtask of detection.

3. ALGORITHMS

Machine learning is an art of programming computer, so they can learn from data. Machine learning systems can be classified according to the amount and type of supervision they get during training process. There are four major categories: Supervised Learning, Unsupervised Learning, Semi Supervised Learning, Reinforcement Learning. In Supervised Learning [14], the training data carries a label (desired solution) that is fed to the algorithm. The training data in Unsupervised Learning is unlabelled, and the system tries to learn by itself without a teacher. Semi Supervised Learning deals with partially labelled training data, usually a lot of unlabelled data and a little bit of labelled data. In Reinforcement Learning, the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.

3.1. Supervised Machine Learning Algorithms

The learning process in a simple supervised learning model is divided into two steps: training and testing. During the training process, the training data is taken as input in which features are extracted and learned by the learning algorithm to build the learning model [15]. In testing process, the predictions are made on the test data using the model that was built in the training

process. Supervised learning is the most common technique used in the classification problems. Let us see the three supervised learning algorithms used in this paper.

Logistic Regression: It is basically a probabilistic model which makes use of a logistic function to model a binary dependent variable. A logistic model has a dependent variable with two possibilities such as pass/fail, true/false, 0/1. The output of this function will be one of the possibilities with a probability value. The logit function is the logarithm of the odds ratio (probability of an event occurring). The function maps the input in the range [0,1] to a real-number range.

$$odds\ ratio = \frac{p}{1-p}$$

Where p = probability of the positive event

$$logit = \log(odds\ ratio)$$
$$logit = \log(\frac{p}{1-p})$$

Support Vector Machine: It is a classifier that maps feature from the non-linear input space to a higher dimensional feature space. The objective of the support vector machine algorithm is to find a hyper plane in an N-dimensional space that distinctly classifies the data points. This converts complex classification problems to linear in a higher dimensional space. For any two classes of data points there are many possible hyperplane that separates the data points and the goal is to find one such hyperplane whose distance between the data points are at a maximum distance. Maximizing the margin distance provides some reinforcement so that future datapoints will be classified with more confidence.

Random Forest: This is basically an ensemble classifier (ensemble method is about combining models to an ensemble such that the ensemble has a better performance than the individual model on an average). It combines through a majority decision tree classifier and the output is combined through a majority. Random Forest can be understood as bagging (bagging is similar to majority voting but uses some learning algorithm to fit models on different subsets of the training data) with decision trees, but instead of growing the decision trees by basing the splitting criterion on the complete feature set, we use random feature subsets. To summarize, in random forests, the decision tree is fit on different bootstrap samples, and for each decision tree, a random subset of features is selected at each node upon optimal split.

3.2. Unsupervised Machine Learning Algorithms

It refers to the utilization of Artificial intelligence algorithms to recognize patterns in datasets containing datapoints that are neither classified nor labelled. Unlike supervised learning, data is not split into training and testing datasets. The algorithms are thus allowed to classify labels and/or group the datapoints in the datasets without having any external guidance in performing that task. It allows the system to identify pattens within datasets on its own. Unsupervised learning system will group unsorted information according to similarities and differences even though there are no categories provided. Unsupervised Learning is the most common technique in the clustering problems. Let us see the three unsupervised learning algorithms used in this paper.

Local Outlier Factor: Outliers are patterns in the datasets that do not conform to the expected behaviour. There are mainly two types of outliers: Global Outliers and Local Outliers. In global

outliers the datapoints are significantly different from the rest of the dataset. In local outliers, the datapoints are significantly different from their neighbours in the dataset. Local outlier factor is a score that tells how likely a certain data is an outlier. It is a calculation that looks at the neighbours of a certain point to find out its density and compare this to the density of other points later. It performs well when the density of the data is not the same throughout the dataset.

Isolation Forest: It is similar to random forest and is built on the basis of decision trees. It explicitly identifies anomalies or outliers rather than profiling normal datapoints. It isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of that selected feature. The split depends on how long it takes to separate the points. In principle, outliers are less frequent than regular observation and are different from them in terms of values as they lie further away from the regular observations in the feature space. When a forest of random trees collectively produces shorter path lengths for samples, they are highly likely to be anomalies.

K-means: It is an iterative method that tries to partition the dataset into 'K' pre-defined distinct non-overlapping clusters where each datapoint belongs to only one cluster. It tries to make the intra-cluster datapoint as similar as possible while keeping the cluster as far as possible. It assigns datapoints to a cluster such that the sum of squared distance between the datapoints and the cluster centroid is at the minimum. The less variations we have within the clusters, the more homogeneous the datapoints are within the same cluster. In K-means algorithm, we first specify the number of clusters K and initialize centroids by shuffling the dataset and then randomly selecting K data points for centroids without replacement. Continue iterating until there is no change to the centroids or until the iteration process has been completed.

4. PERFORMANCE MEASURE

To evaluate the performance of a particular model, we make use of various parameters. Confusion matrix is a summary table showing how good the model is at prediction by plotting the number of correct predictions against the number of incorrect predictions. It has four categories: *True Positive* (TP), here the predicted value matches the actual value. Actual value was positive, and the model predicted a positive value. *True Negative* (TN), here the predicted value matches the actual value. The actual value was negative, and the model predicted a negative value. *False Positive* (FP), here the predicted value was falsely predicted. Actual value was negative, but the model predicted a positive value. *False Negative* (FN), here the predicted value was falsely predicted, the actual value was positive, but the model predicted a negative value.

Accuracy is a measure of how many correct predictions your model made. It is a good basic metric to measure the performance of a model, but the downside of a simple accuracy is that it works well in balanced datasets and becomes poorer metric in unbalanced datasets.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall is a measure of how many true positives get predicted out of all the positives in the dataset. It is also called as sensitivity. The recall value can often be turned by tuning several parameters of the machine learning model. A high recall means that most of the positive cases was labelled as positive. A low recall means that there is a high number of false negative.

$$recall = \frac{TP}{TP + FN}$$

Precision is a measure for the correctness of a positive prediction. It means that if a result is predicted as positive, how sure can you be that the result is actual positive.

$$precision = \frac{TP}{TP + FP}$$

As with recall, precision can be turned by tuning the parameters of the model. A higher precision typically leads to a lower recall and higher recall leads to a lower precision. So, there is a trade-off between precision and recall.

F1 is a combination of precision and recall, namely their harmonic mean. It is needed when the balance between precision and recall must be maintained.

$$F1 = 2\,\frac{precision * recall}{precision + recall}$$

5. EXPERIMENTAL RESULTS

The dataset for the experiment was taken from Kaggle [18] website. It contains transactions made by credit cards in 2013. The dataset is labeled and contains fraudulent transactions with 492 out of total transactions of 284,807. Therefore, the data is considered to be unbalanced since the fraudulent cases are 0.173%. Figure 1 shows the distribution of dataset. It consists of 30 columns without the column labels. In order to conserve privacy, a PCA projection was applied to all columns excluding: time and amount features. Therefore, all columns are numerical variables. The labels columns contain a breakdown of the two classes where 0 and 1 correspond to a valid transaction and fraud, respectively.

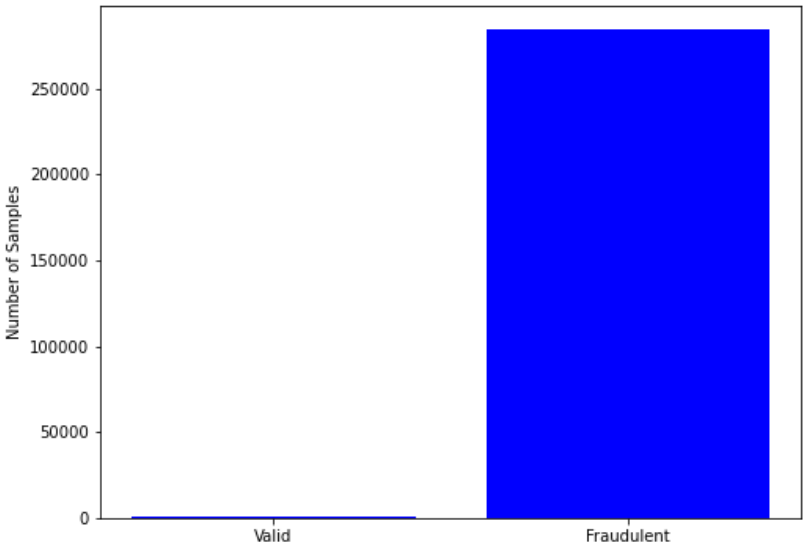

Figure 1. Distribution of dataset

The main purpose of this study is to demonstrate how the various algorithms perform on the dataset. Figure 2 shows the accuracy scores for all the algorithms. The highest accuracy scores are averaged about 99% but these accuracy scores are misleading since, accuracy metric is only

well suited for balanced datasets. Table 1 shows the calculations of accuracy, precision, recall and F1 scores which will help in determining the best algorithm.

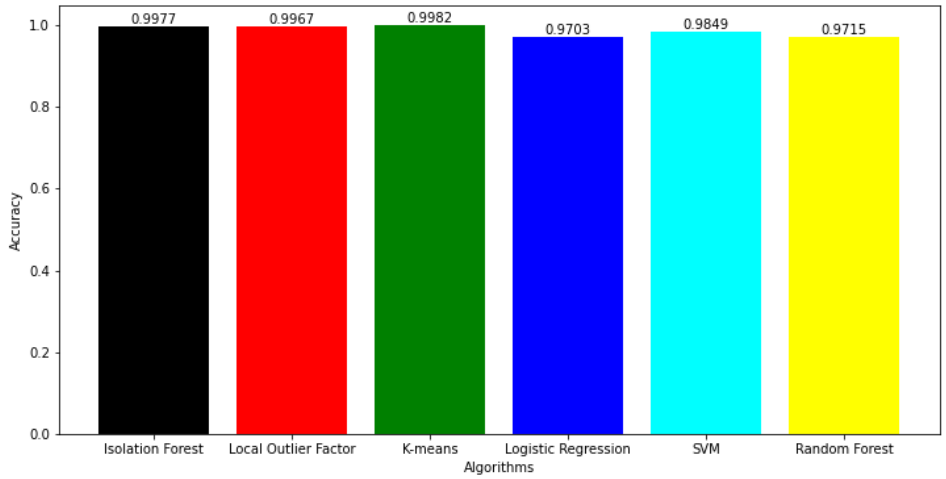

Figure 2. Accuracy scores for each algorithm.

Table 1. Performance measure for supervised and unsupervised learning algorithms.

Model	F1 score	Accuracy	Recall	Precision
Logistic Regression	**0.9326**	**0.9703**	**0.909**	**0.9574**
SVM	**0.9479**	**0.9849**	**0.9191**	**0.9785**
Random Forest	**0.9435**	**0.9715**	**0.9293**	**0.9583**
Isolation Forest	**0.67**	**0.9977**	**0.67**	**0.67**
Local Outlier Factor	**0.25**	**0.9967**	**0.25**	**0.25**
K-means	**0.926**	**0.9982**	**0.879**	**0.9798**

Precision gives us an idea of how many times the algorithm has detected the fraud correctly, recall gives an idea of how much it detects and F1 score helps in maintaining the precision recall trade off. Based on these ideas, it is observed that precision value for K-means algorithm is highest, recall value is highest for random forest algorithm and F1 score is highest for support vector machine algorithm. Precision value for support vector machine is very close to the precision score for K-means and has a very good recall value of 0.9191 which is next to recall value of random forest. Support vector machine algorithms performs very well for credit card fraud detection with an accuracy of 98.49% and a high precision/recall value.

6. CONCLUSION

We have developed supervised and unsupervised models with the goal to detect fraudulent transactions from a large unbalanced dataset. Comparative results in terms of the comparison metric is the percentage of correctly identifying fraudulent transactions and precision, recall, accuracy and F1 score have been presented. In fact, accuracy can be misleading where it could misrepresent a machine learning technique. For example, local outlier factor has an accuracy of 99.67% but performs poorly based on precision and recall values. So precision, recall and F1 score values plays a significant role in deciding the best algorithm for fraud detection. K-means algorithms is the best among the unsupervised learning algorithms and support vector machine performs well among all the algorithms used.

REFERENCES

[1] Ogwueleka F N, (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology Vol 6, No 3, pp 311-322..

[2] Rama Kalyani K and Uma Devi D, (2012). Fraud Detection of Credit Card payment system by Genetic Algorithm, International Journal of Scientific and Engineering Research, Vol 3 Issue 7,pp 1-6 ISSN 2229-5518.

[3] Meshram P L and Bhanarkar P, (2012). Credit and ATM card fraud detection using Genetic approach, International Journal of Engineering Research and Technology, Vol 1 Issue 10, pp- 1-5 ISSN 2278-0181.

[4] Seeja K R and Zareapoor M, (2014). Fraud Miner: A Novel credit card fraud detection model based on Frequent Itemset Mining, The Scientific World Journal Hindawi Publishing Corporation, Volume 2014 Article ID 252797, pp 1-10.

[5] Patil S, Somavanshi H, Gaikwad J, Deshmane A and Badgujar R (2015). Credit card fraud detection using Decision Tree induction algorithm, International Journal of Computer Science and Mobile Computing, Vol 4 Issue 4 pp 92-95 ISSN: 2320-088x.

[6] Duman E, Buvukkava A and Elikucuk I (2013). A novel and successful credit card fraud detection implemented in a Turkish bank. In Data Mining Workshops 2013. 13th International Conference on IEEE pp 162-171.

[7] Bhnsen A C, Stojanovic A, Aovada D and Ottersten B (2014). Improved credit card fraud detection with calibrated probabilities. SIAM International Conference on Data Mining pp 677-685. Society for industrial and applied mathematics.

[8] Bolton R J and Hand D J (2002). Statistical fraud detection: a review. Statistical Science 17(3), 235-249

[9] Behdad M, Barone L, Bennamoun M and French T (2012). Nature inspired techniques in the context of fraud detection. IEEE transaction on System Management and Cybernetics Part C, 42(6) 1273-1290.

[10] Sahin Y and Duman E (2011), Detecting credit card fraud by Decision Tree and Support Vector Machine. Lecture notes in Engineering and Computer Science, 2188(1).

[11] Heta Naik, PrashastiKanikar (2019). Credit card fraud detection based on Machine Learning Algorithms, International Journal of Computer Applications (0975-8887) Volume 182 No 44 March 2019.

[12] R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1264-1270, doi: 10.1109ICICCS48265.2020.9121114

[13] L. Meneghetti, M. Terzi, S. Del Favero, G. A Susto and C. Cobelli, "Data-Driven Anomaly Recognition for Unsupervised Model-Free Fault Detection in Artificial Pancreas", IEEE Transactions On Control Systems Technology, pp. 1-15, 2018

[14] Nasteski, Vladimir. (2017). An overview of the supervised machine learning methods. HORIZONS.B. 4. 51-62. 10.20544/HORIZONS.B.04.1.17.P05.

[15] Sandhya N. dhage, Charanjeet Kaur Raina. (2016) A review on Machine Learning Techniques. In International Journal on Recent and Innovation Trends in Computing and Communication, Volume 4 Issue 3.

[16] Sperandei, Sandro. (2014). Understanding logistic regression analysis. Biochemia medica. 24. 12-8. 10.11613/BM.2014.003.

[17] Powers, David &Ailab,. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981. 10.9735/2229-3981.

[18] https://www.kaggle.com/mlg-ulb/creditcardfraud

[19] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. Pages 116-125, 2007.