Analysis of Credit Card Fraud detection using Machine Learning models on balanced and imbalanced datasets

WARSE The World Academy of Research in Science and Engineering, Yeshwanth Z

International Journal of Emerging Trends in Engineering Research

ISSN 2347 - 3983

Volume 9. No. 7, July 2021

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter02972021.pdf

https://doi.org/10.30534/ijeter/2021/02972021

Analysis of Credit Card Fraud detection using Machine Learning models on balanced and imbalanced datasets

Manoj Kumar Reddy Mallidi[1], Yeshwanth Zagabathuni[2]

[1]Vignan's Foundation for Science, Technology and Research, India, manojkumarreddy.mallidi@gmail.com

[2]Vignan's Foundation for Science, Technology and Research, India, zyeshwanth@gmail.com

ABSTRACT

With the advent of modern transaction technology, many are using online transactions to transfer money from one person to another. Credit Card Fraud, a rising problem in the financial department goes unnoticed most of the time. A lot of research is going on in this area.The Credit Card Fraud Detection project is developed to spot whether a new transaction is fraudulent or not with the knowledge of previous data. We use various predictive models to ascertain how accurate they are in predicting whether a transaction is abnormal or regular. Techniques like Decision Tree, Logistic Regression, SVM and Naïve Bayes are the classification algorithms to detect non-fraud and fraud transactions.

In modern conditions where data may vary in a matter of minutes or even seconds, conventional classification techniques may not perform well. When dataset involves huge numbers of differences in data distribution and also changing data with high dimensionality and volume issues supervised learning comes up short. Hence we may resort to unsupervised learning, semi-supervised or any other means to cope with that.

The number of online transactions has grown enormously these days and credit card transactions hold an enormous share of these transactions. More numbers of people are using a credit card for shopping, e-commerce, e-wallets and even for education purposes. Therefore, banks and other stakeholders give fraud detection applications priority and value. Fraudulent transactions can be in different categories. They may be through Online or Offline. Our paper deals with the online category and one of many methods to handle them, which is the machine learning way.

Key words: Credit Card Fraud, Fraud Detection, Machine Learning, Supervised learning, Un-supervised learning

1. INTRODUCTION

Credit Card Fraud is an illegal activity which fraudsters do in order to gain profit in a less amount of time and this will be known to the users, a few days after the fraud has happened and they will respond later registering a complaint regarding the fraud which they have been through. Basically the fraudsters may use online or offline payment on using credit card. Through the online mode these days, the attackers only need to know the phone number or Aadhaar number or simply the mail-id of the user as one of them is already linked to their account. One of these details is quite easy to obtain as the user might have already given these details across multiple websites. If the mobile number or email-id is obtained, then all the attacker does is track it until a certain message containing a one-time password has arrived and attack. Finally, the victim is left gazing at messages stating that a certain amount of money drawn at a certain time. Having an easily decodable password is an added advantage is also a boon for the fraudsters which was explained by Omkar and Kinn[29].

The fraudsters using offline means on the other hand need to have the credit card and the corresponding 4-digit pin to get money from the account.

Credit card fraud detection is a difficult task as money may differ from one account to the other and so there is no particular pattern for identifying the fraud. These days, most of the businesses are through online means, so there will be more number of frauds and there will be a very huge data set as a result and this will be confidential and not released to the public without prior permission from the head of the department (finances). Even the customer with a minimum balance is using net banking to buy things online and they do shopping, e-commerce etc. through online means. Data Mining is mostly used to detect these kinds of fraud using several algorithms. In this project we used supervised algorithms like logistic regression, Naive Bayes, Decision Tree, KNN and un-supervised algorithms like K-Means and DBSCAN algorithms in order to figure out the best suited algorithm for these kinds of problems. For better results we used few of the advanced algorithms such as Multi-Layer Perceptron, ensemble techniques such as Gradient Boost, Random Forest and XGBoost.

The Transaction has an id which can be used to track whether it is fraudulent or not. Basically, these kinds of problems have two classes legit or fraud (0 or 1). The fraud transactions are the illegal activities which lead to loss of money without notice. Credit card datasets are rarely available as they are confidential, being related to the finance department, and are highly imbalanced. The data-set we used, has complete data of credit card transactions of European customers for two days. PCA was also applied to it to reduce the dimensionality. The data set available is imbalanced and has a smaller number of frauds. Hence, it is difficult to identify the patterns in them, so we search for the algorithm best suited to adapt, fit the data and predict the patterns in data.

The columns 'V1-V28' are the results of PCA Transformation. The attribute 'Time' is the time between the current transaction and first transaction and the attribute 'Amount' is the amount of money which was drawn. The attribute Class is the one which has the value as 1 if the case is fraud and 0 otherwise. In this data set we have 492 fraud cases out of 284,807 transactions.

Fraudulent transactions can be of any type online fraud or offline, but the loss may be a huge one for users, reducing the reputation of the banks too. While online transactions happen through technical gadgets and virtual money, offline transactions happen in banks through physical cash.

These days the people are engaged in using social media and online transaction Applications which made them easy to transfer money from their account to other users and vendors for their business. Hence, most of the business people use online transactions and likewise, the frauds the becoming much easier.

2. RELATED WORK

In this Project we have performed the data normalization before performing cluster analysis or classification. The purpose of data normalization is to bring the data into a single and scalable format. Normalization may involve various methods such as min-max, decimal scaling, Z-score etc. Credit Card Fraud Detection[6],[8] involves immense research to find out the fraud cases by using Machine Learning[5],[26] and Data Mining[7],[27] as major fields.

Analysis of Credit Card Fraud detection[2],[12] involves continuous monitoring of performance of a given model. Although the frauds in a very large database are minute or negligible they create a huge impact and leave many users vulnerable in the process. There are basically two approaches which are supervised and unsupervised approaches and the recently popularized semi-supervised approach.

Firstly, in a supervised approach the data is pre-labelled. The machine is then provided with a new set of examples and produces an outcome using the labelled data. The outcome may be 0 or 1 based on which is much more simplified and highly accurate on datasets that do not change continuously.

One of the most popular supervised approaches in detecting fraud is logistic regression. It works by classifying users into two classes fraudulent and non-fraudulent based on a kernel function and this is to be monitored regularly. However, in the modern times where data changes in minutes or seconds, class labels are much more difficult to predict and assign and supervised methods may not be that accurate. This created the need for unsupervised, semi-supervised and other strategies such as ensemble learning. The clustering algorithms have also become reliable tools in the field of fraud detection[18].

The unsupervised algorithms K-Means and DBSCAN were also used in the project. There were many classification algorithms we used, such as decision tree (J48), KNN, Naïve Bayes etc. There is no single perfect algorithm. Hence, we followed the approach of combining results of multiple tree-based algorithms in a step-by-step approach for better results, which is commonly referred to as ensemble methods. Being implemented in python offers much more flexibility compared to other programming environments as it already has all the necessary algorithms implemented in modules such as sklearn, sci-py, pandas, pytorch, imblearn etc. It has many functions to ensure that we can implement any supervised or unsupervised algorithm and evaluate it. Hence we can also calculate performance metrics of the algorithm using some of those functions.

This paper presents a case study involving credit card fraud detection. We demonstrate how a seemingly perfect transactional database may contain a few unnoticeable frauds. It draws its inspiration from many other related fields such as text mining, game theory[21], firewall breach, intrusion detection etc. These fields are all based on fraud detection in different approaches.

Many models have been suggested for accurate fraud detection. For example, we can consider the neural network proposed by Ghosh and Reilly which is trained on a large sample of labelled credit card transactions[9]. These transactions contain a variety of fraud cases such as lost cards, stolen cards, application fraud, e-mail fraud etc[9]. Training on a variety of data certainly makes the model invulnerable to almost any kind of fraud. Hence, the quality or variety of data matters more than the quantity or bulk.

There were many other approaches in the past and there are going to be many in the future and there is still a lot of scope for this field as the frauds continue to be inevitable. Some of interesting approaches in the past were meta-classifier based fraud detection[23], fraud detection based on behavior[16], machine learning models[1],[22],[25], data mining approaches[4],[25], neural classification[9], game-theory approach[21] web services based detection[24], Hidden Markov Model[11],[14] , Predictive Analysis[19] and so on. In addition, applying some of the most robust classification algorithms[10] such as SVM[27] and ensemble algorithms[20] such as Random Forest[3],[15],[17] and Adaboost[13] was also preferred by a lot of researchers. Especially with voluminous data generated these days almost any algorithm which failed in the past could shine and the frauds of course shine accordingly. And hence, this cycle of new frauds-to-solution appears to be never ending. We just cannot aim for perfection in an uncertain field as this as no matter how secure we assume our systems to be, they are only as secure as we think they are.

3. PROPOSED METHODOLOGY

3.1. Logistic Regression

Logistic Regression is a popular means of supervised learning which is used to estimate outcomes such as win/loss, positive/negative etc. It makes use of a sigmoid function whose value lies in between 0 and 1.
The basic logistic regression model is as follows,

$$p = \frac{e^{\alpha+\beta_n X}}{1+e^{\alpha+\beta_n X}} \qquad (1)$$

As shown above it makes use of the outcome of linear regression, but it might not be restricted to one variable.

3.2. Decision Tree

Decision Tree is a tree structured algorithm that uses a series of decisions to predict a class. It has the attribute selection measures such as Information Gain, Gini Index and Gain Ratio based on which it is called ID3, C4.5 and CART respectively. Our project uses the CART algorithm which is known to handle outliers most effectively.
The following formulae are used in the algorithm:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^{\,2} \qquad (1)$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$
(2)

3.3. Random Forest

Random Forest is an ensemble method which relies on averaging a lot of decision trees and is used for classification and regression. Unlike decision trees this method is less prone to overfitting.
Its goal is to reduce variance. Although there is a small increase in bias and some loss of interpretability the overall performance is boosted.
The visualization of how the algorithm works is in the Figure 1 given below:

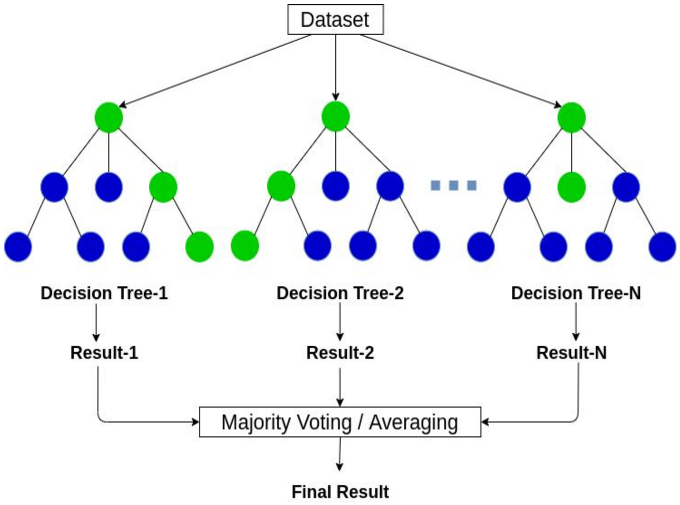

Figure 1: Random Forest Visualization

*Figure 1 was taken from https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

3.4. KNN-Classifier

K-Nearest Neighbors is an algorithm for classification and regression. A data object is classified using the majority vote of its nearest neighbors.
For example, if K=3 then the object is assigned to class of its three nearest neighbors. In order to find out how near a data object is to a neighbor and finally assign it a class, there are several distance measures:

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (1)$$

$$Manhattan\ Distance = \sum_{i=1}^{k}|x_i - y_i| \qquad (2)$$

$$Minkowski\ Distance = \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \qquad (3)$$

3.5. K-Means

K-Means uses the initial cluster centers to group similar objects to any one of them and thus form arbitrary shapes called clusters.
The parameters that are required are the value of K and the initial choice of cluster centers for the K clusters. The shapes of the clusters highly depend on the initial choice of cluster centers.
The proximity measures used are many. Some of the popular ones are Manhattan distance, Minkowski distance, Euclidean distance etc and our project uses Euclidean distance.
It is applicable in case of large data and is computationally efficient. Its easy implementation and adaptability to new examples makes it much better than other hierarchical clustering methods.

3.6. DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a widely used clustering algorithm in detecting fraud. The key concept is that for each cluster the neighborhood in a given epsilon radius has to contain at least a specified number of data points.
This method can also find clusters inside clusters and clusters of any shape. It is not restricted to only certain arbitrary shapes of clusters and is extremely effective in filtering outliers. The process involved in DBSCAN can be understood from Figure 2 given below:

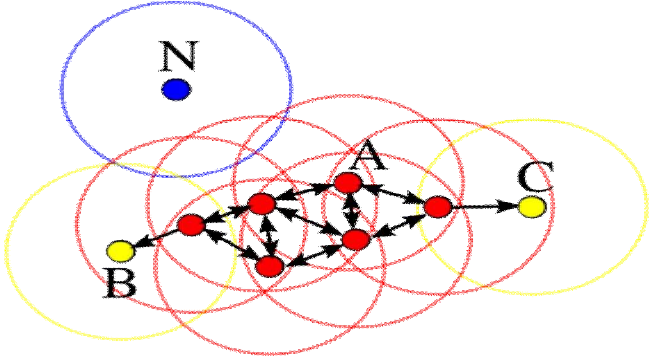

Figure 2: DBSCAN Visualization

*Figure 2 was taken from https://en.wikipedia.org/wiki/DBSCAN.

3.7. Multi-Layer Perceptron

Multi-Layer Perceptron works on data which cannot be linearly separated. Considering the dynamic environment these days its additional advantage comes in handy.

Multi-Layer Perceptron is a feed-forward ANN, which works by forward propagating the weights carried by each unit (neuron) to the next layer. The layer may be any one of: input layer, hidden layer or output layer.

In addition there is a summation function which calculates the weighted sum and adds an overall bias in the end and an activation function to map weighted inputs to the outputs.

3.8. Gradient Boost

It is one of the boosting techniques and can be used for classification and regression.

Gradient boosting follows a feedback like mechanism from different weak learners called decision trees and finally produces an output minimizing sum of the squares of errors. It operates in stages and in each stage there is a decision tree which gives a decision by selecting different set of features. Taking into account the decision and the errors made by the previous learner (decision tree), the current learner looks to improve by rectifying the errors. And thus the error is minimized in in a step by step manner.

3.9. Naïve Bayes

This Naïve Bayes classifier is based on simplest Bayesian network models. This classifier is highly scalable requiring a number of parameters in a problem. It is based on Bayes theorem on conditional probability and the attributes however are assumed to be independent of each other. The formulation is as shown in Figure 3 below:

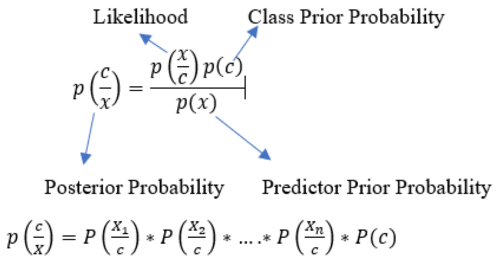
Likelihood
Class Prior Probability
$p\left(\frac{c}{x}\right) = \frac{p\left(\frac{x}{c}\right)p(c)}{p(x)}$
Posterior Probability
Predictor Prior Probability
$p\left(\frac{c}{x}\right) = P\left(\frac{x_1}{c}\right) * P\left(\frac{x_2}{c}\right) * \ldots .* P\left(\frac{x_n}{c}\right) * P(c)$

Figure 3: Naïve Bayes

3.10. XGBoost

XGBoost (Extreme Gradient Boosting) is another boosting technique which was introduced as an improvement over Gradient Boosting. The risk of overfitting the dataset as in Gradient Boosting is reduced in XGBoost. It also offers the additional leverage of handling missing values on its own. The working procedure however, remains the same as Gradient Boosting. The technique offers much more advanced regularization and accurate approximations. Its training is also very fast compared to GBT (Gradient Boosting Technique).

4. EXPERIMENTATION RESULTS

For our experimental study we have used a dataset available on kaggle[30]-[38], an online repository containing thousands of datasets.

The dataset contains a total of 284,807 transactions out of which 492 were fraudulent cases and the remaining are the non-fraudulent cases[30]-[38]. Table 1 is the summary for a train test split samples.

Table 1: Dataset Summary

Source	Samples	Train_size	Test_size	Split
Kaggle Repository	284,807	190,820	93,987	2:1

This resembles that the dataset which we have is imbalanced so in order to handle such kind of datasets we have implemented SMOTE over-sampling over the train data so that the dataset is re-sampled[28] and balanced and as a result the number of samples increase to over 380,000.

Table 2: Imbalanced dataset results

Algorithm	Performance Analysis on Imbalanced Dataset			
Modern Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	99.961	94.444	79.865	86.545
XG Boost	99.962	93.181	82.550	87.544
Gradient Boost	99.911	72.916	70.469	71.672
MLP	99.954	87.323	83.221	85.223
Conventional Classification Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	99.927	87.155	63.758	73.643
Naïve Bayes	97.787	5.733	83.892	10.734
J48	99.914	69.942	81.208	75.155
KNN	99.944	87.022	76.510	81.428
Unsupervised Algorithms	**NMI**	**Adjusted Rand Index**	**Fowlkes Mallows Score**	**MSE**
K-Means	2.77e-05	-0.0012	0.7084	0.4579
DBSCAN	0.004	-0.0029	0.915	459.30

Table 3: Results after balancing

Algorithm	**Performance Analysis on Balanced Dataset**			
Modern Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	99.955	85.906	85.906	85.906
XG Boost	98.951	12.274	91.275	21.638
Gradient Boost	98.016	6.974	93.288	12.978
MLP	98.036	6.867	90.604	12.765
Conventional Classification Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	97.463	5.460	91.946	10.308
Naïve Bayes	97.618	5.534	87.248	10.408
J48	99.743	35.714	77.181	48.832
KNN	99.802	43.894	89.261	58.849
Unsupervised Algorithms	**NMI**	**Adjusted Rand Index**	**Fowlkes Mallows Score**	**MSE**
K-Means	0.436	0.546	0.998	0.00107
DBSCAN	0.00044	-0.0028	0.819	3180.4

The supervised learning algorithms were evaluated with metrics obtained from the confusion matrix such as precision, recall and accuracy. The harmonic mean of precision and recall (F1-Score), gives a better idea about the model performance.

The clustering algorithms are evaluated with different metrics specified in Tables 2 and 3.

From Table 2, it can be identified that almost all supervised algorithms have high accuracy. However, the unsupervised algorithms have underperformed. This can be clearly observed considering the very high MSE and very low NMI. So the dataset should be balanced and to balance the imbalanced dataset we have used the *imblearn.over_sampling* module to oversample the data.

In the next table we have the same metrics on the balanced dataset. In this case the results vary significantly from the first. Certain algorithms which may have underperformed previously perform much better now. For instance, the K-means algorithm has shown a huge change in NMI and Adjusted Rand Index. At the same time certain algorithms such as MLP and XGB which performed really well previously have now underperformed.

Although XGB showed the highest F1-score on the imbalanced dataset, it failed to do so, on the balanced dataset and its score dropped significantly. Although Random Forest turned out to be the second best in Table 2, it turned to be the best in Table 3.

5. CONCLUSION

A total of 10 different algorithms were trained first on the balanced dataset and then on the imbalanced dataset in this project. Many algorithms ended up underperforming as the size of the dataset increased dramatically after balancing.

However, Random Forest yielded the most satisfactory results with high accuracy and consistent F1-Score in both cases. Hence, it appears to be the most suitable for classifying large-volumes of data samples.

ACKNOWLEDGMENT

The first author thanks fellow colleague Manoj Kumar for his sincere dedication and valuable contribution towards this project. He also thanks Mrs. B. Jyostna Devi, Assistant Professor, Vignan's Foundation for Science, Technology and Research and Dr. S.K. Satpathy, Associate Professor, Vignan's Foundation for Science, Technology and Research for their enormous support throughout the completion of the paper.

REFERENCES

[1] Navanshu Khare and Saad Yunus Sait. **Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models**, International Journal of Pure and Applied Mathematics (IJPAM), Volume 118 No.20 2018, pp. 825-838, ISSN: 1314-3395, www.ijpam.eu.

[2] Dushyant Singh, Saubhagya Vardhan and Dr.Neha Agrawal. **Credit Card Fraud Detection Analysis**, International Research Journal of Engineering and Technology (IRJET), Volume 5 Issue 11, Nov 2018,ISSN:2395-0056,www.irjet.net.

[3] Devi Meenakshi B, Janani B, Gayatri S and Indira N. **Credit Card Fraud Detection Using Random Forest**, International Research Journal of Engineering and Technology (IRJET),Volume 6 Issue 3, March 2019,ISSN:2395-0056, www.irjet.net.

[4] Priyanka Yadav, Pavan Wangade, Manish Thakur, Mohammed Fakih and Gayatri Hegde. **Proposed Distributed Data Mining In Credit Card Fraud Detection**, International Research Journal of Engineering and Technology (IRJET),Volume 3 Issue: 4,April 2016,ISSN:2395-0056, www.irjet.net.

[5] Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga and Nuwan Kuruwitaarachchi. **Real Time Credit Card Fraud Detection Using Machine Learning**, IEEE, July 2019. **INSPEC Accession Number:** 18868933.

[6] Dinesh L. Talekar and K.P. Adhiya. **Credit Card Fraud Detection System**, International Journal of Modern Engineering Research (IJMER), Volume 4 Issue 9, Sept 2014, ISSN: 2249-6645, www.ijmer.com.

[7] Franscisca, Nonyelum and Ogwueleka. **Data Mining Application in Credit Card Fraud Detection System**, Journal of Engineering Science and Technology, Volume 6 Issue 3, June 2011.

[8] Nutan Suman. **Review Paper On Credit Card Fraud Detection**, International Journal of Computer Trends and Technology (IJCTT), Volume 4 Issue 7, July 2013, ISSN:2231-2803,www.ijcttjournal.org.

[9] Sushmito Ghosh and Douglas L. Reilly. **Credit Card Fraud Detection with a Neural Network**, International Conference on System Sciences 1994.

[10] Aihua Shen, Rencheng Tong and Yaochen Deng. **Application of Classification Models on Credit Card Fraud Detection**, IEEE 2007.

[11] Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun K.Majumdar. **Credit Card Fraud Detection Using Hidden Markov Model**, IEEE Transactions On Dependable and Secure Computing, Volume 5 Issue 1, Jan 2008.

[12] S.Benson Edwin Raj and A.Annie Portia. **Analysis on Credit Card Fraud Detection Methods**, International Conference on Computer, Communication and Electrical Technology (ICCCET), 18th March 2011.

[13] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim and Asoke K.Nandi. **Credit Card Fraud Detection Using AdaBoost and Majority Voting**, IEEE 2018.

[14] Shaiesh S.Dhok. **Credit Card Fraud Detection using Hidden Markov Model**, International Journal of Soft Computing and Engineering (IJSCE), Volume 2 Issue 1, March 2012, ISSN: 2231-2307.

[15] M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika and E. Aswini. **Credit Card Fraud Detection Using Random Forest Algorithm**, 2019 3rd International Conference on Computing and Communications Technologies (ICCT), 2019

[16] Yongbin Zhang, Fucheng You and Huaqun Liu. **Behavior-Based Credit Card Fraud Detecting Model**, 2009 Fifth International Joint Conference on INC, IMS and IDC, 2009

[17] Debashree Devi, Saroj. K. Biswas and Biswajit Purkayastha. **A Cost-sensitive weighted Random Forest Technique for Credit Card Fraud Detection**, 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCNT), 2019

[18] Kononenko, Igor and Matjaz Kukar. **Cluster Analysis**, Machine Learning and data mining, 2007.

[19] Kosemani Temitayo Hafiz, Shaun Aghili and Pavol Zavarsky. **The use of predictive analysis technology to detect credit card fraud in Canada**, 2016 11th Iberain Conference on Information Systems and Technologies (CISTI) 2016.

[20] Inderpreet Kaur and Mala Kalra. **Ensemble Classification Method for Credit Card Fraud Detection**, International Journal of Recent Technology and Engineering, Volume: 8 Issue 3, September 2019.

[21] Vishal Vatsa, Shamik Sural and A.K. Majumdar. **A Game-Theoretic Approach to Credit Card Fraud Detection**, International Conference on Information Systems Security, pp. 263-276, 2005.

[22] Khadija Abdul Sattar and Mustafa Hammad. **Fraudulent Transaction Detection in FinTech using Machine Learning Algorithms**, International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT), 2020.

[23] J Pun and Y Lawryshyn. **Improving credit card fraud detection using a meta-classification strategy**, International Journal of Computer Applications, 2012.

[24] Chuang-Cheng Chiu and Chieh-Yuan Tsai. **A web services-based collaborative scheme for credit card fraud detection**, IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004. EEE'04. 2004, pp. 177-181, 2004.

[25] Xiangji Huang, Yan Huang, Miao Wen, Aijun An, Yang Liu and Josiah Poon. **Applying Data Mining to Pseudo-Relevance Feedback for High Performance Text Retrieval**, Sixth International Conference on Data Mining (ICDM'06), 2006.

[26] Ruttala Sailusha, V. Gnaneswar, R. Ramesh and G. Ramakoteswara Rao. **Credit Card Fraud Detection Using Machine Learning**, 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020.

[27] Naoufal Rtayli and Nourddine Enneya. **Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization**, Journal of Information Security and Applications, 2020.

[28] Nur Farhana Hordri, Siti Sophiayati, Nurulhuda Firdaus, Siti Mariyam. **Handling Class Imbalance in Credit Card Fraud using Resampling Methods**, International Journal of Advanced Computer Science and Applications, 2018.

[29] Omkar Dastane and Kinn Abass Bakon. **The Effect of Bad Password Habits on Personal Data Breach**, International Journal of Emerging Trends in Engineering Research, Volume 7 No.10, October 2020.

[30] Andrea Dal Pozzolo, Oliver 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.

[31] Andrea Dal Pozzolo, Oliver Caelen, Yann-Aël Le Borgne, Serge Waterschoot and Gianluca Bontempi. **Learned lessons in credit card fraud detection from a practitioner perspective**, Expert systems with applications, Volume 41 Issue 10, pp. 4915-4928, 2014, Pergamon.

[32] Andrea Dal Pozzolo, Giacomo Boracchi, Oliver Caelen, Cesare Alippi and Gianluca Bontempi. **Credit card fraud detection: a realistic modelling and a novel learning strategy**, IEEE transactions on neural networks and learning systems, Volume 29 Issue 8, pp. 3784-3797, 2018, IEEE.

[33] Andrea Dal Pozzolo. **Adaptive Machine Learning for credit card fraud detection**, ULB MLG PhD thesis (supervised by G. Bontempi).

[34] Fabrizio Carcillo, Andrea Dal Pozzolo, Yann-Aël Le Borgne, Oliver Caelen, Yannis Mazzer and Gianluca Bontempi. **Scarff: a scalable framework for streaming credit card fraud detection with Spark**, Information fusion, Volume 41, pp. 182-194, 2018, Elsevier.

[35] Fabrizio Carcillo, Yann-Aël Le Borgne, Oliver Caelen and Gianluca Bontempi. **Streaming active learning strategies for real-time credit card fraud detection: assessment and visualization**, International Journal of Data Science and Analytics, Volume 5 Issue 4, pp. 285-300, 2018, Springer International Publishing.

[36] Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oble and Gianluca Bontempi. **Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection**, INNSBDDL 2019: Recent Advances in Big Data Deep Learning, pp. 78-88, 2019.

[37] Fabrizio Carcillo, Yann-Aël Le Borgne, Oliver Caelen, Frederic Oble and Gianluca Bontempi. **Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection**, Information Sciences, 2019.

[38] Yann-Ael Le Borgne and Gianluca Bontempi. **Machine Learning for Credit Card Fraud Detection – Practical Handbook**.