# An Innovative Model-Based Approach for Credit Card Fraud Detection Using Predictive Analysis and Logical Regression

**S. Praveen Kumar, A. Sahithi Choudary**

***Abstract*: The development of information technology and advancements in communication channels has resulted in increasing fraud throughout the world and immense monetary losses. The objective of fraud detection frameworks is to check each exchange for the likelihood of being false and to recognize fraudulent ones as fast as possible after the fraudster has started to execute a fraudulent transaction, paying little mind to the prevention mechanisms. For this purpose, we utilize a steady foolproof 5 stage verification model with, predictive analysis, logistic regression, outlier model, custom rule management and global profiling. A predictive (LVQ) algorithm alongside logistic regression would improve credit card fraud detection. The benchmark Kaggle dataset is used. The outcomes portray a convincing decrease in credit card frauds.***

***Keywords: Credit card fraud detection, Logistic regression, Classification, Kaggle.***

## 1. INTRODUCTION

Credit card fraud is a form of theft where the credit card information of an individual is abused to make false purchases and withdraw funds in their name. This can be done via theft or by unlawfully acquiring the cardholder's personal and account information. All payment cards, including debit and credit cards, can be liable to this type of fraud. The type of threats that one can be exposed to can be categorised into internal and external threats. Types of fraud attacks[1] include accidental leak, espionage, misuse, lost or stolen cards, card not present fraud(CNP), identity fraud (masquerading), counterfeit card fraud (skimming), triangulation, and so on. One primary setback due to this is monetary and those most affected by it include banks and merchants. These losses incurred reached $21.84 billion in 2015 and is expected to surpass $12 billion by 2020. However, according to The Nilson Report[2], these numbers still remain lower than the peak years back in the 1970s.

To prevent this credit card fraud, an innovative model has been proposed recommending various enhancements in fraud detection techniques. Four methodologies, in particular, predictive analysis, fraud analysis, outlier models, custom rule management and global profiling have been incorporated. Predictive analytics is used to estimate activity, behaviour and trends by creating predictive models by using statistical analysis techniques, analytical queries and applying machine learning algorithms to data sets. A numerical value is then assigned for the possibility of a particular event occurring. Fraud-detection algorithms are implemented in order to detect fraudulent activities on credit cards. It uses real-time data analysis to determine if a transaction is genuine or not. Despite the fact that the framework hasn't been perfected yet, it has managed to reduce losses drastically by almost 70 per cent in the US since 1992. Outliers are observations that diverge from an overall pattern on a sample. An outlier model can be used to detect such anomalies and rectify them. We can also define custom rules which can typically be used for whitelisting or blacklisting transactions. This helps us recognise transactions that do not comply with the rules and are suspicious. Global profiling can be used to distinguish a transaction as fraudulent or legitimate depending on the user profile built from past data, or transaction behaviour explicit to the user. The approach is as follows. Refer fig 1.1 This strategy serves to characterise each credit card transaction as either valid or fraudulent.

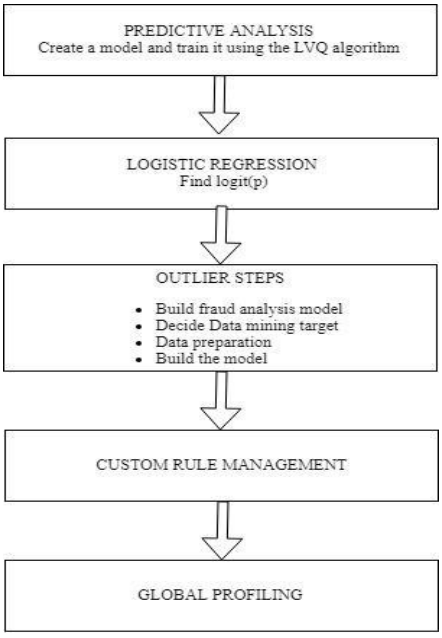



**Fig 1.1 Architectural model for credit card fraud detection**

**Revised Manuscript Received on May 07, 2019.**

**S. Praveen Kumar,** Assistant Professor, Department of Information Technology, GITAM Institute of Technology, GITAM (Deemed to be)University, Visakhapatnam, Andhra Pradesh, India

**A. Sahithi Choudary,** Department of Information Technology, GITAM Institute of Technology, GITAM (Deemed to be)University, Visakhapatnam, Andhra Pradesh, India

## 2. LITERATURE SURVEY

A credit card is an indirect method of obtaining merchandise or services without the use of cash, by the means of offering credit. The number and type of such transactions that can be made with these third parties are interminable. In any case, this simplicity in making such transactions accompanies numerous potential downfalls, the fundamental one being credit card fraud. Most of the time, these fraudsters are associated with the affected party. The type of fraudsters can be internal or external. External fraudsters can be additionally classified as the average offender, criminal offender, and the organised crime offender [3].

One early methodology devised under machine learning is the Gradient descent method. It is an optimization algorithm under which we find the values of parameters (coefficients) of a function that minimizes the cost function. It is most appropriate for search by an optimisation algorithm. However, it runs slowly on very large datasets. This is because each instance in the training dataset requires a prediction for every single iteration, which can be very tedious. [21]

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \quad \text{----(1)}$$

Another approach is the radial basis function kernel, or RBF kernel, which is a well-known kernel function utilised in different kernelized learning algorithms. It is most commonly used in support vector machine classification. In the field of mathematical modelling, a radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control.[22]

$$f(x) = \sum_{i}^{N} \alpha_i y_i \exp(-||x - x_i||^2/2\sigma^2) + b \quad \text{----(2)}$$

One approach devised to detect fraud is the use of neural networks[4]. Neural networks empower a computer to think for itself along the lines of how a human brain functions. In a human brain, the information and knowledge gained from past encounters are used in decision making, a similar strategy is applied to detect credit card fraud and characterise it as false or genuine. The customers have a fixed pattern of credit card use, which is learnt from past experience. The neural network is then trained using this information and additionally about the various types of credit card fraud that the particular bank is susceptible to. Based on these patterns, a prediction algorithm is used to classify that specific transaction as fake or genuine. The system checks the usage pattern used by the fraudster and matches it with existing patterns of the original card holder. If the event that the pattern matches, then the transaction is declared authentic. Else, the concerned authorities are alerted that the credit card is being used by an unapproved individual.

Another approach is the use of logistic regression [5]. It is a statistical method used to anticipate binomial or multinomial results. Multinomial Logistic Regression algorithm is used to create models when the target field is a set field with at least two possible values. Binomial Logistic Regression algorithm is restricted to models where the target field is a flag or binary field. [10]Logistic regression methods built in IBM SPSS Clementine 12 are stepwise, enter, forwards and backwards. The stepwise method can be utilised in multinomial LR. The other strategies can be used in both binomial logistic regression and multinomial logistic regression.

User profiling is another method for identifying fraud.[6] Though this method was implemented in the scenario of mobile networks, the process of modelling characteristic aspects of client behaviour can likewise be implemented in credit card fraud detection. Users are characterised into particular groups. It is based on this that a customer usage profile is built. There are several models that can be utilized for profiling, for example, the Hidden Markov model (HMM) and the Hierarchical regime-switching model (HRSM), and the Finite mixture model (FMM). In this specific study, FMM was used.

## 3. METHODOLOGY

### 3.1 Predictive Analysis Model

Neural networks are inspired by the human brain. They comprise a network of neurons that are interconnected. That is, it is a set of computational units, which take a set of inputs and transfer the result to a predefined output. Every one of these units is ordered and then organised in layers in a way that the features of an input vector can be associated with the features of an output vector. Neural networks are thus coached to model the relationships within the given data sets. Specifically, for predictive analysis, the methodology utilised is as follows

1. Make a model dependent on rules set up by the LVQ algorithm during the training stage.
2. Test the model on the verification data set – the data is fed to the model and the anticipated values are compared to the actual values. The model is thus tested for accuracy.
3. The model is utilised to classify new incoming data and makes a move depending on the output of the model. Weights are updated if necessary.

#### *3.1.1 Learning Vector Quantization algorithm (or LVQ for short):*

An artificial neural network algorithm enables you to pick the number of training instances to use and learn exactly what those instances should resemble.

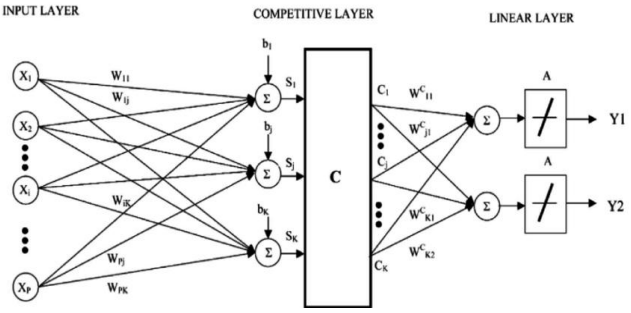



**fig: 3.1 Learning Vector Quantization**

A neural network with a competitive layer is used to implement the LVQ algorithm. [11] Each layer comprises of competitive neurons which are equal to the number of clusters. Every competitive neuron i corresponds to a cluster and its weight vector given in equation 1 corresponds to the centroid of the clusteri.

$$w_i = (w_{i1}, ..., w_{id})^T, \; i = 1, ...., M \quad \text{----(3)}$$

The algorithm is repetitive and requires the initialization of the networks weight vectors w_i . In every iteration or epoch, a vector ⟦x ⟦^n⟧ ⟧ ^is given as input to the network, the distances from each centroid w_i are determined and finally, the winner neuron m with the minimum value of Euclidean distance given in equation 2 is chosen.

$$d^2(\mathbf{x}^n, \mathbf{w}_m) = \sum_{j=1}^{d} (x_j - w_{ij})^2. \quad \text{-----(4)}$$

The last step is to update the weight vectors by "moving" the winner's neuron centroid $w_m$ "closer" to the input vector $x^n$. The amount of "moving" depends on a η parameter (learning rate).

LVQ clustering algorithm

1. Define the number of clusters.
2. Initialize the M centroids $w_i(0)$, $i = 1, ...,M$.
3. Initialize learning rate η, epochs counter k=0 and repetitions counter k=0.
4. For every epoch k, do the accompanying steps for
   - Set vector $x^n$ as the Neural Network's input.
   - Select the winner neuron *m*.
   - Update the weight vector for the winner neuron

$$w_{ij}(k+1) = \begin{cases} w_{ij}(k) & i \neq m, \; i=1,...,M \\ w_{ij}(k)+\eta(x_{nj}-w_{ij}(k)) & i=m, j=1,...,M \end{cases} \quad \text{---(5)}$$

   - k=k+1.
5. Check for termination. If not set k=k+1 and return to step 4.

[12] The LVQ is represented by a collection of codebook vectors. Initially, they are chosen arbitrarily and later adjusted to best summarize the training dataset over a number of iterations of the learning algorithm. After learned, the codebook vectors can be used to make predictions similar to the manner of K-Nearest Neighbours. The most similar neighbour (best matching codebook vector) is found by finding the distance between each codebook vector and the new data instance. The class value or (real value in the case of regression) for the best matching unit is then returned as the prediction. Best outcomes can be obtained if you rescale your data to have a similar range, such as between 0 and 1.

### 3.2 Logistic Regression

Several statistical models have been applied for applications in data mining tasks, such as regression analysis, multiple discriminant analysis, logistic regression and probit method, etc.[7][8] A set of predictive variables are utilised to predict the result and the absence or presence of a specific characteristic in logistic regression. Though alike to a linear regression model, it is more apt for models which comprise dichotomous dependent variables. Logistic regression coefficients can be used to assess odds ratios for every one of the independent variables in the model and are applicable to a wider range of research situations than discriminant analysis.

Logistic regression uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression (or logit regression) is used in determining the parameters of a logistic model, which is a type of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are designated by an indicator variable, where the two values are marked "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of at least one independent variable. ("Predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labelled "1" can fluctuate between 0 (certainly the value "0") and 1 (certainly the value "1"), thus the labelling; the function that changes log-odds to probability is the logistic function, and thus the name. The binary logistic regression model has extensions to multiple levels of the reliant variable: categorical outputs with multiple values are modelled by multinomial logistic regression, and if the multiple categories are ordered- by ordinal logistic regression.

The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), however it tends to be used to make a classifier, for example by picking a cutoff value and classifying inputs with probability greater than the cutoff as one class and below the cut off as an alternate; this is a typical way to make a binary classifier. Major suppositions for binary logistic regression:[9]

1. The dependent variable should be dichotomous in nature (e.g., presence vs. absence).
2. There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores, and removing values below -3.29 or greater than 3.29.
3. There should be no high correlations (multi Collin earity) among the predictors. This can be assessed by a correlation matrix among the predictors and suggest that as long correlation coefficients among independent variables are less than 0.90 the assumption is met.

At the centre of the logistic regression analysis is the task of estimating the log odds of an event. Mathematically, logistic regression estimates multiple linear regression functions defined as in equation 6.

logit(p)

$$= \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$

for i = 1…n -----(6)

### 3.3 Outlier Model

An outlier model involves building a credit card fraud analysis model. This involves business analysis, data research, data adjustment, modelling, model verifications etc., and furthermore gives us an idea about the implementation of the model. There are several steps involved which are given in figure 3.3

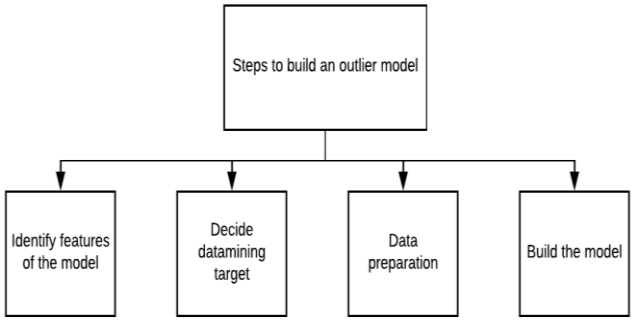



**Fig 3.3 Steps to build an outlier model**

### *3.3.1 Build fraud analysis model*

Merchant fraud features analysis: Main characteristics of merchant fraud include

*a. Merchant size:*

Merchant size is primarily of two types- nature of the merchant company and operational floor area. In a general scenario, large merchants with standard administration are relatively protected, while the probabilities of fraud in smaller, recently established merchants are greater.

*b. If the merchant has the intention of fraud:*

In most cases, small, private merchants who sell antiques, ginseng & medicine and so forth, have a greater intention of committing fraud.

*c. Nature of operations:*

Luxury consumption sites have a higher or bigger likelihood of fraud.

*d. Transaction time:*

The odds of fraud are a lot bigger amid the non-business hours.

*e. Transaction amount:*

The probabilities of fraud in large, integral transaction amounts are bigger.

*f. Transaction goods:*

Small sized goods which are high in value and can be easily turned into cash are increasingly advantageous for committing fraud.

*g. Transaction frequency:*

Frequent transactions of large dollar value within short intervals is also profoundly favourable for fraud.

*h. Successive transactions on cards from different countries:*

Successive high dollar-valued exchanges on cards from various countries in a short timeframe also make the likelihood of risk bigger.

*i. Has failed transaction record:*

When a certain transaction on a card fails, the probability of fraud occurring on that same card increases Multifood.

### *3.3.2 Decide data mining target*

The primary goal of data mining is to anticipate the probability of a certain credit card transaction made, as genuine or fraudulent.

### *3.3.3 Data preparation*

All transactions in the transaction records are considered as target variables. During the data preparation stage, n number of variables is obtained from the original data. These variables are then standardised-this implies that the influence of the dimensions is eliminated by transforming them mathematically. This is done by normal standardisation (0,.1) of the indexes. Next, data level regression and IV conversion are performed on discrete data. The advantage of the latter is that absent and zero values are treated as a major category without influencing the other normal values and also find the correlation between this variable and target variable from the information value. To assess the model in an objective manner, we divide the modelling data into a training set and a measuring set.

### *3.3.4 Build the model*

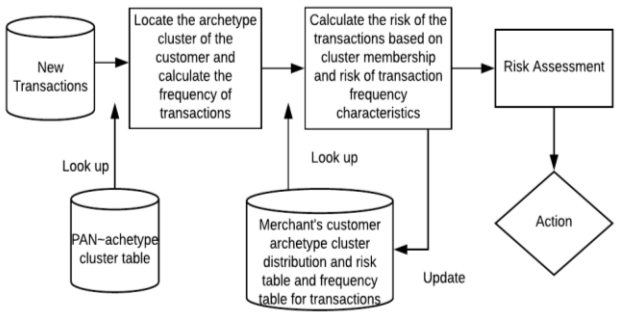
The classification model includes algorithms such as logistic regression, neural networks, and decision tree. A credit card fraud situation comes under this model. To assess such a model objectively, we first build it utilising transaction data of the previous year and then evaluate it using the transaction data of the consecutive year.

### *3.3.5 Advice on model application*

For precise predicting results, the model must be verified with test data. These outcomes can then be used as a source of reference for evaluating the risk of certain transactions made and make sensible decisions depending on that. In scenarios where the transaction conforms to some present rule of thumb and is also included in the fraudulent transaction list- the odds that the transaction is fake is very high and the risk control staff is required to step in. Building an effective data mining model is an on-going task. The passage of time, changes in the exterior environment, the technique, mode of fraud, is few of several factors that can vary and decrease the applicability of the model. It is thus necessary that the model is consistently tracked and adjusted based on its feedback.

## 3.4 Custom Rule Management

[18]When a device that comes to the online site is not recognised, it is evaluated for unique behavioural patterns associated with risk and gives recommendations based on the rules we specify. When a certain rule is triggered, that specific rule's weight contributes to the overall transaction score. Devices with higher scores are considered to be more reliable and might be permitted to proceed. However, those with lower scores may be automatically denied, flagged for manual review, or presented with authentication challenges, depending on the defined preferences. Thus, having a wide range of custom rules at your disposal is exceptionally helpful and valuable in preventing online fraud.



### 3.5 Global Profiling

The nature of the environment in which fraud happens, how it impacts the fraudsters capabilities, the kind of setting in which it occurs are considered alongside storing different fraudster profiles and also details of the more typical types of fraud that occur. All this helps us comprehend what a typical fraudster is like. This is called global profiling.

The defining characteristics, features, or behaviours assist us in identifying individuals within the organization and even outside who are more likely to perpetrate fraud. Contexts in which they commit fraud and with whom are also observed. This process provides us with information that can be used to battle fraud.

## 4. KAGGLE DATASET

The dataset used contains transactions made via credit cards by European Cardholders in September 2013 over a span of two days. Out of 284,807 transactions, 492 are fraudulent accounting for 0.172% of all transactions. The dataset is thus highly unbalanced.

Input variables are numerical and are a result of PCA transformation. Features V1, V2, V28 are the principal components obtained with PCA. The only features that haven't been transformed are 'Time'-which contains the seconds elapsed between each transaction and the first transaction and 'Amount'-which is the transaction amount. The latter can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

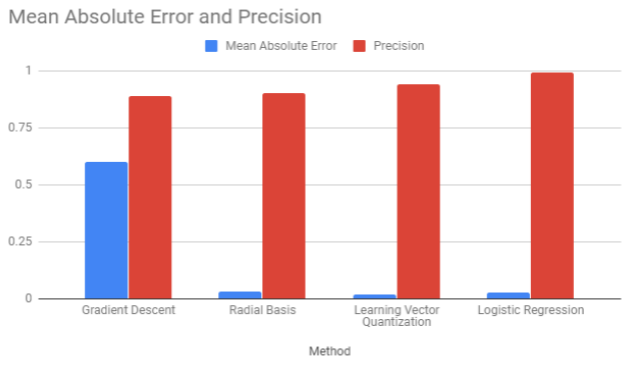
## 5. RESULTS

**Table 1: Output derived by comparing with that of the Existing Methods**

| Method | Identification of Credit card Fraud | Mean Absolute Error | Precision |
|---|---|---|---|
| **Gradient Descent** | 68.4 | 0.60 | 0.89 |
| **Radial Basis** | 80.2 | 0.033 | 0.90 |
| **Learning Vector Quantization** | 87.72 | 0.020 | 0.94 |
| **Logistic Regression** | 89.2 | 0.028 | 0.99 |

The table above compares the various methods used and how effective they are in terms of Precision and Mean Absolute Error.

The graph obtained for the Mean Absolute Error and Precision is as follows. From the graph it can be observed that Precision is highest for Logistic Regression followed by the Learning Vector Quantization Method.




## 6. CONCLUSION

As the usage of credit cards turn out to be increasingly more typical in each field of day to day life, credit card fraud has turned out to become much more rampant. To enhance the security of the financial transaction systems in an automatic and effective way, building an accurate and efficient credit card fraud detection system is highly necessary for the financial establishments.

This work demonstrates a 5 step model which is highly beneficiary in reducing the fraud risks that banks are subject to. The outcomes demonstrate that the proposed model involving predictive analysis, logistic regression, outlier model, custom rule management and global profiling processes is highly effective and efficient in solving the issue.

## REFERENCES

1. Credit card frauds and measures to detect and prevent them, International Journal of Marketing, Financial Services & Management Research ISSN 2277- 3622 Vol.2, No. 3, March (2013).
2. Bruno Buonaguidi. (n.d.). Credit card fraud: What you need to know. Author. Retrieved From Https://nilsonreport.com/upload/pdf/Credit_card_fraud_what_you_need_to_know.pdf
3. Phua, C., Lee, V., Smith, K. and Gayler, R., 2005. A Comprehensive Survey of Data Mining-based Fraud Detection Research., Artificial Intelligence Review.
4. Raghavendra Patidar, Lokesh Sharma, June 2011, Credit Card Fraud Detection Using a Neural Network.
5. Y. Sahin, E. Duman, Detecting Credit Card Fraud by ANN and Logistic Regression.
6. Ogwueleka, F.N.; and Enyeama H.C. (2009). Credit card fraud detection using artificial neural networks with a rule-based component. The IUP Journal of Science and Technology, 5(1), 40-47.
7. Altman, E. I., Marco, G., & Varetto, F. Corporate distress diagnosis comparisons using linear discriminant analysis and neural networks. Journal of Banking and Finance, 18(3), 505–529, 1994.
8. Flitman A.M. Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis. Computers & Operations Research, Volume 24, Issue 4, 367-377, 1997.
9. Logistic regression [Web log post]. (n.d.). Retrieved from https://www.statisticssolutions.com/what-is-logistic-regression/
10. Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. 2011 International Symposium on Innovations in Intelligent Systems and Applications.
11. Salastas, J. (2011, August 24). Implementation of Competitive Learning Networks for WEKA [Web log post]. Retrieved from https://jsalatas.ictpro.gr/implementation-of-competitive-learning-networks-for-weka/
12. Le, J. (n.d.). A Tour of The Top 10 Algorithms for Machine Learning Newbies [Web log post]. Retrieved from https://www.kdnuggets.com/2018/02/tour-top-10-algorithms-machine-learning-newbies.html/2
13. 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
14. Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon
15. Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modelling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

16. Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier
17. Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018, Springer International Publishing
18. Glenn, E. (2016, February 5). Fine-tuning Fraud Prevention with Granular Business Rules [Web log post]. Retrieved from https://www.iovation.com/blog/customizing-online-fraud-prevention-with-business-rules
19. Data Mining [Web log post]. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Portal:Data_mining
20. Logistic regression [Web log post]. (n.d.). Retrieved from http://sheepshoot.com/Logistic_regression
21. Brownlee, J. (2016, March 23). Gradient Descent For Machine Learning [Web log post]. Retrieved from https://machinelearningmastery.com/gradient-descent-for-machine-learning/
22. Radial basis function [Web log post]. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Radial_basis_function_network