Anomaly Detection using Unsupervised Methods: Credit Card Fraud Case Study

Article *in* International Journal of Advanced Computer Science and Applications · January 2019

DOI: 10.14569/IJACSA.2019.0101101

Mohammad Mahdi Rezapour Mashhadi

94 PUBLICATIONS **1,077** CITATIONS

Anomaly Detection using Unsupervised Methods: Credit Card Fraud Case Study

Mahdi Rezapour
University of Wyoming
United States

***Abstract*—The usage of credit card has increased dramatically due to a rapid development of credit cards. Consequently, credit card fraud and the loss to the credit card owners and credit cards companies have been increased dramatically. Credit card Supervised learning has been widely used to detect anomaly in credit card transaction records based on the assumption that the pattern of a fraud would depend on the past transaction. However, unsupervised learning does not ignore the fact that the fraudsters could change their approaches based on customers' behaviors and patterns. In this study, three unsupervised methods were presented including autoencoder, one-class support vector machine, and robust Mahalanobis outlier detection. The dataset used in this study is based on real-life data of credit card transaction. Due to the availability of the response, fraud labels, after training the models the performance of each model was evaluated. The performance of these three methods is discussed extensively in the manuscript. For one-class SVM and auto encoder, the normal transaction labels were used for training. However, the advantages of robust Mahalanobis method over these methods is that it does not need any label for its training.**

Keywords*—*Credit card fraud; anomaly detection; SVM; Mahalanobis distance; autoencoder; unsupervised techniques

I. INTRODUCTION

Fraud detection in large scale is one of the biggest challenges in fraud investigation. Annually credit fraud resulted in a loss of billions of dollars [1]. Fraud could be defined as any wrongful or criminal activity that could result in financial loss to a card holder, and personal gain of the fraudsters [2]. The two main approaches of avoiding fraud is through fraud prevention and fraud detection [3]. An application of fraud detection becomes practical when the fraudsters exceed the fraud prevention systems and start a fraudulent transaction activity. Thus, a responsibility of a fraud detection would be defined as checking any transaction with the objective of preventing a fraudster from a fraudulent activity. Credit card fraudulent activity is a most notorious activity in a financial system.

Fraudulent transaction or Outliers could be divided into two main groups: global and local outliers. A global outlier is a measured observation which has a very high or very low value relative to other observation in the dataset. On the other hand, a local outlier is a sample point with a value within a normal range of the whole dataset but compared with surrounding points, it is usually high or low. An efficient fraud detection system should be able to detect the frauds accurately and also to adjust its performance based on the changes in the behaviors of fraudsters.

Machine learning techniques are primarily methods in identifications of frauds. These techniques could be divided into two groups: supervised and unsupervised methods. In supervised machine learning techniques, a model would be trained on a past sample of fraudulent and legitimate transactions in order to classify new transactions as fraudulent or legitimate. In other words, the supervised learning uses the whole labeled dataset for training. The labels are known since card holders did identify the mismatch of a transaction, or an unusual transaction being identified by credit card agency and confirmed by a credit card holder. The supervised methods have this disadvantage that if fraudsters change their patterns, these models might not be able to detect them based on the old observations.

On the other hand, unsupervised techniques acquire information from new transactions and the anomalies would be based on updated transactions. In unsupervised fraud detection anomalies or unusual transactions would be identified as possible cases of fraudulent transactions. The advantage of unsupervised learning is that a machine does not need the knowledge of the fraud labels to train itself on, and a decision on identifying a transaction as an outlier would be made based on the distribution of the transaction. However, normal transaction labels are needed for most of the unsupervised methods so machine learning techniques would be trained on normal transaction so it can differentiate between normal and fraud for the upcoming transactions.

This study is based on real-life data of transaction from an international credit card corporation. Frauds happen barely compared with the total number of transactions so due to having an imbalanced dataset, under-sampling method was conducted to have balanced categories for comparison. Also, although unsupervised machine learning techniques do not need labeled data for the whole dataset, the data labels were used in this study for performance evaluation of different models.

II. LITERATURE REVIEW

Machine learning techniques for fraud/outlier detection could be divided into two main approaches: supervised and unsupervised approaches. The supervised method needs the whole data to be labeled for fraud identification meaning that it should be clarified in the dataset whether a transaction is fraud or legitimate. In many cases it is not clear whether a transaction is fraudulent or not as that transaction was not completed by the system. For these cases the analysis would favor unsupervised method. This study will go over few studies conducted on fraud detection using various supervised and unsupervised learning and, then, it moves to the application of the unsupervised learning techniques implemented in this study.

A study conducted to compare a performance of different supervised and unsupervised method for studying credit card fraud detection [4]. Four unsupervised anomaly detection methods including one-class support vector machines (SVM), restricted Boltzmann machine, and generalized adversarial network were used as unsupervised methods. The performance across different models was compared using area under the curve (AUC).

Hidden markov model (HMM) was used to detect credit card frauds [5]. The model was initially trained with the normal behavior of a cardholder, and then evaluated on incoming credit card transactions. If an upcoming credit card transaction is not accepted by the trained HMM with high probability, it would be considered as fraud. Supervised and unsupervised machine learning techniques were combined to detect credit frauds [6]. The results showed that the hybrid technique is efficient and could improve the accuracy of detection.

A study was concerned with behavioral fraud through the analysis of longitudinal data [7]. This study implemented an unsupervised method which used changes in behavior or unusual transaction. Another study conducted with the help of unsupervised method of improved nearest neighbor method to detect intrusion [8]. The Minkowski's distance was modified and used as a means for intrusion detection.

A discussion was made about the shortcoming of one-class SVM due to its high false positive rate [9]. Thus, a new approach named enhanced SVM was proposed, which combine traditional SVM with one-class SVM to create an unsupervised machine learning. Genetic Algorithm (GA) was used as a feature selection method for extracting optimized information from raw dataset.

The following paragraphs will highlight the studies that focused on one-class SVM, autoencoder and Robust Mahalanobis distance methods respectively.

One-class SVM was used to detect anomaly problems [10]. As this method is sensitive to outliers, ramp loss function was introduced to this paper to address this issue. The objective of this function was to make sparse semi-supervised algorithm. The obtained results showed an improvement in outlier detection. In another study, one-class SVM was used for detecting anomalous windows registry accesses using registry anomaly detection (RAD). The system was compared with probabilistic anomaly detection (PAD), and the results showed that PAD outperformed the SVM model possibly due to hieratical prior incorporated on the PAD algorithm. Different machine learning techniques such as SVM, random forest, and the logistic regression model were compared for detection of credit card fraud [11]. The models performances were compared based on different metrics such as precision, sensitivity, and specificity. Two-class and one-class SVM were used and compared for detection of fraudulent credit card transactions [12]. These models were considered and evaluated using different Kernels. The results showed the superiority of one-class SVM for the anomaly detection problem over two class SVM.

Turning to studies used Autoencoder for anomaly detection, Autoencoder based on ensemble model was used as an anomaly detection method in building energy data [13]. A comparison was made across ensembles of different auto encoder models. The threshold for normal versus anomalous observations was based on the assumption that 5% of the data are anomaly candidates. Credit card fraud detection was proposed using regular autoencoder and variational autoencoder (VAE), defined as a variant of autoencoder that uses a probabilistic graph as a basic for anomaly detection [14]. Reconstruction error was used as an anomaly score for the autoencoder and a reconstruction probability. It was found that a simple regular autoencoder outperform the VAE for detecting credit card fraud. Another study used anomaly detection to identify anomaly related to deviation of practical building operation due to existence of operating faults and improper control strategies [13]. An autoencoder-based ensemble method was developed for anomaly detection in this study, A number of autoencoder, and autoencoder-based ensemble, were stacked with different architecture. Root mean square was used as a metric for the model evaluation.

One of the ways to implement multivariate outlier detection (MVO) is through Mahalanobis, or Cook's distance. Robust Mahalanobis distance has been used extensively for anomaly detection in the literature review. When Mahalanobis is used for MVO, a large (squared) Mahalanobis distance would be considered as Multivariate outliers [15]. However, Mahalanobis distance is sensitive to the presence of outliers [16] due to sensitivity of arithmetic mean and sample covariance matrix to outliers [17]. The solution of this problem could be achieved by estimating the mean and covariance matrix in a robust manner, resistance against the impact of outlying observations [15]. The minimum covariance determinant (MCD) is most commonly used estimator of multivariate location and scatter due to having a computationally fast algorithm [18]. This matrix is calculated by the subset of observation of size h, which minimize the determinant of the covariance matrix.

Multivariate outlier detection was used in exploration of geochemistry [15]. The method was able to distinguish between extreme values of a normal distribution and values obtained from different distribution. In this study Mahalanobis was used by robust estimates, which downgrade the impact of the extreme values in the solution. In order to simplify the visualization of the outliers spatially on a map, a multivariate outlier plot was introduced, which uses different symbols to illustrate the distance measures from the center of the distribution. Different colors were also used to highlight the magnitude of the distance from the center. The same methodology was implemented for interpretation of multivariate outlier for compositional dataset [19]. The isometric logratio(ilr) transformation was implemented on the data before conducting any analysis.

In this study, a robust geographically weighted method was used for multivariate spatial outlier detection [20]. Also a large Mahalanobis distance was used as a means for anomaly detection, and in order to have a robust estimate of distance, MCD was used.

The rest of this paper is organized as follows. The data description will talk about the data used in this study. The method will go over the three unsupervised methods being implemented for detection of credit card fraud. The remaining sections will talk about results and discussion.

III. Data Description

The input data consists of numerical values resulted from principal component transformation to preserve confidentially. The response is binary classes with 1 in case of fraud and 0 otherwise. Time was in seconds indicating amounts passed between each transaction, and the first transaction in the dataset. Amount refers to transaction amount. There were 29 predictors in the dataset, including the response variables. Before including the predictors in the model, all the distribution for fraudulent and normal transaction were plotted and compared to see if it is necessary to include all the predictors in the analyses. For instance, the distribution of two predictors are depicted in Fig. 1, time on the left and amount spent on the right. The first one, left, is the time for both true or normal transaction and fraudulent or false for the response (class) whether a transaction was fraud or normal.

As can be seen from Fig. 1 on the left, the two transactions have almost similar distribution indicating that the time variable would not be much of a help for prediction of transaction types. On the other hand, for the "amount" predictor as expected the two distributions are different across different transactions as a response.

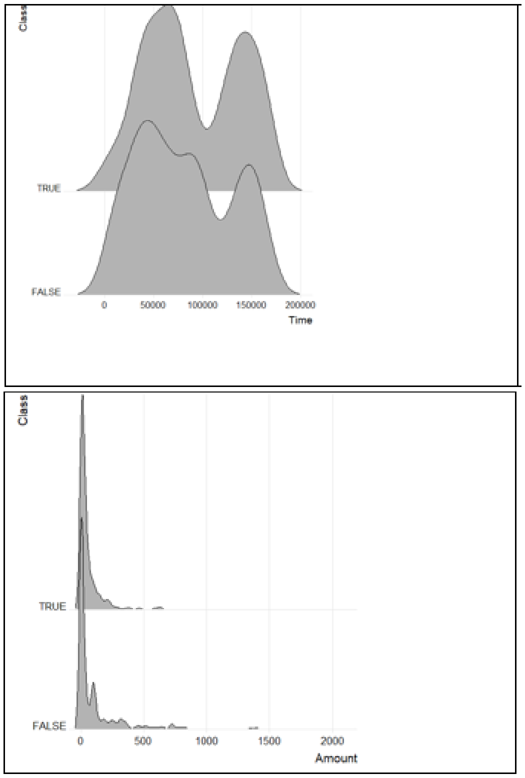

Fig. 1. Distributions of Time and Amount Spent as a Transaction Versus different Class Categories as Response.

Machine learning techniques tend to produce unsatisfactory predictions for the minority class when the data is imbalanced. Generally, when the minority class events accounts for less than 5% of the all response category in the dataset, that dataset is called imbalanced. For this dataset, fraudulent observation accounts for 0.17% of all observations. For imbalanced dataset even best of algorithms are incapable of detection of fraud from legitimate transactions, and they would face the problem of identifying too many false positive, legitimate transaction, as fraudulent ones [21]. In order to address the problem of imbalanced dataset, under sampling technique was used to convert the dataset into a balanced dataset. It has been found that random under sampling of the majority class to be generally batter than other sampling methods [22]. This method has been used in credit card fraud detection often [11]. Thus his in this study, this method was taken advantage for balancing the dataset.

IV. Method

The following section would describe different methods applied in this study to identify credit cards frauds.

A. One-Class SVM

One-class SVM is particularly useful for imbalanced dataset where there are many cases of normal data and not many cases of outliers (anomalies). The objective if this method is to see if test data is a member of a class of the training dataset or not.

The method could be viewed as a quadratic optimization problem, minimizing an objective functionω, with an objective of identifying a best algorithm to maximize the accuracy of the training dataset. To enhance an application of the model trained on test dataset, the distance between the margins, support vectors, needs to be maximized.

In SVM, the anomalies in the positive data which was used in negative samples would be identified. The one-class problem could be written as follows [23]:

$$\mathrm{f(x)}\begin{cases}+1, if\ x \in S\\ -1, if\ x \in S\end{cases} \qquad (1)$$

For this method, the algorithm maps the data into a feature space H using kernel function, and then a hyperplane would try to separate the mapped vector with maximum margin. The hyperplane could be written as:

$$\omega^T x + b = 0 \qquad (2)$$

Where ω is the normal vector to the hyperplane, x is a feature, and b is an intercept.

For this model all the data points for class –1 are on one side and all the data points for class 1 are on the other side. The hyperplane searches for the maximal margin between the classes.

The training dataset is $(x_1, x_2, ..., x_n) \in R^n \times (\pm 1)$, where a Kernel map could be written as $\Phi: R^n \rightarrow H$, which transforms the data into feature space H. The model, then, minimize the objective function as follows:

$$\min_{\omega,b,\varepsilon} \frac{||\omega||^2}{2} + C \sum_{i=1}^{n} \varepsilon_i \qquad (3)$$

Slack of ε_i would be introduced to the model to prevent the SVM from overfitting with noisy data, on the other hand, the constant of $C > 0$ governs the trade-off across maximizing the margin and the number of training data points within the margin, and $\frac{2}{||\omega||}$ is a distance between the two support vector, which is subject to:

$$y_i(\omega^T(\varphi(x_i) + b) \geq 1 - \varepsilon_i \; for\, all\; i = 1, \ldots., n \tag{4}$$

$$\varepsilon_i \geq 0 \; for\, all\; i = 1, \ldots., n \tag{5}$$

The equation for on-class SVM is a bit different. This model separates all the data points from the origin with the objective of maximizing the distance from the hyperplane to the origin. The results of the above function would be positive for a small region and negative elsewhere. The quadratic minimization function could be written as follows:

$$\underset{\omega,b,\varepsilon}{min} \frac{||w||^2}{2} + \frac{1}{\nu n}\sum_{i=1}^{n} \varepsilon_i - \rho \tag{6}$$

$\nu \in (0,1)$ is an important Parameter that characterize the solution for the machine by controlling the trade-off between maximizing the distance from the region and containing most of the data in the region created by hyper plane.

B. Deep Autoencoder Network

Autoencoder is a branch of neural network that could be used to learn data in an unsupervised manner. The goal of this model is to learn a representation (encoding) of a dataset. This method can be applied on various objectives such as dimensionality reduction or anomaly detection. Typically, this method is trained over number of iterations with an optimizer and an objective of minimizing the cost function such as mean square error (MSE) or reconstruction error. This model performance in fraud identification is based on the assumption that the distribution of normal transaction is different that the distribution for fraudulent ones.

This model has two parts: the encoder f (mapping $X\ to\ F$), and the decoder g (mapping $F\ to\ \acute{X}$) [24]. Generally, autoencoders are symmetric, with the first half of the autoencoder is considered as encoder and the other half is considered as decoder.

It can be said that autoencoder follows the same principal as feed forward neural network and the same technique could be applied to this model. The goal of this model is to construct the inputs by minimizing the difference between the input and output, compared with feed forward neural network having an objective of predicting output as Y given input x. It should be noted for this model, the number of inputs is equal to the number of outputs. The model can be depicted in a simple way consisting of encoder and decoder functions:

$$\theta: X \rightarrow F \tag{7}$$

$$\varphi: F \rightarrow \acute{X} \tag{8}$$

$$\theta, \varphi = \underset{\theta,\varphi}{argmin} ||X - (\varphi.\theta)X||^2 \tag{9}$$

Where θ and φ are transactions where θ map the input X to F (encoder), and transaction φ which map F back to the input (decoder), and the objective is to find transactions that would minimize the objective function, which is MSE.

F, which is a Map of X, is referred as code and can be written as follows:

$$F = \sigma(Wx + b) \tag{10}$$

where $\sigma, W, and\ b$ are activation function, weight and bias, respectively. These values are initialized randomly and then updated during back propagation. After this stage the decoder, maps back F to the reconstruction $\acute{X}$

$$\acute{X} = \acute{\sigma}(\acute{W}F + \acute{b}) \tag{11}$$

As mentioned earlier the objective function or OF is to minimize the cost function such as MSE”

$$OF = minimize||X - \acute{X}||^2 = ||X - \acute{\sigma}(\acute{W}(\sigma(Wx + b)) + \acute{b}||^2 \tag{12}$$

Where $\acute{\sigma}$ and σ are activation functions and the other parameters were defined earlier.

Before conducting a main analysis, different hyper parameters were tuned to get a model with a better performance or lower error rate. In this model, fraud observation from normal observation was distinguished by mean square error (MSE), with the assumption that fraudulent transaction would have a higher MSE’s. However, for the sake of a comparison, there should be a unique threshold k, MSE>K, so an observation higher than this threshold would be considered as a fraud transaction.

It is expected that frauds transaction to have a different distribution than normal transaction. In other words, autoencoder will have higher reconstruction errors on frauds than on normal transactions. Thus, the reconstruction error can be used to distinguish fraudulent transactions from the normal ones.

Again, as this method uses the same concept as neural network, and neural network models assign a higher importance to variable with higher values, the data needs to be reprocessed before conducting any statistical modeling. This scaling was conducted by dividing each cell by the difference of maximum and minimum of that column.

Two autoencoder layouts are commonly have been employed in the literature review: a bottleneck or under complete, where the number of nodes in the hidden layer is less than the number of nodes in the input layer. The second option is called over-complete layer where the number of nodes in the hidden layer is higher than the number of nodes in the input layer. Beside these two methods, denoising autoencoder (DAE) could be considered. This method objective is to achieve a good representation through changing the reconstruction criteria [25]. After creating partially corrupted input, the model would try to recover the original uncorrupted input. The amount of corrupted input that would be added to the model would be typically 30%. There are three common methods used for denoising. In the first approach, isotropic Gaussian noises would be added to the input layer based on Gaussian model. The second approach randomly selects a fraction of input variables and set them as either zero or one. For masking noise, the third approach, a fraction of input values would be randomly selected, and their values would be masked at zeros.

Although a basic version of the autoencoder consists of three fully connected layers including one input, one output, and one hidden layer in between, the performance of auto encoder might be improved by adding new hidden layers or increasing the number of nodes. The autoencoder deployed in this study consists of four fully connected layers including one input layer, two hidden layer and one output layer. Generally, the performance of this model due to being unsupervised is evaluated by the normality of reconstruction residuals and the quality of the features extracted by autoencoder. For instance, shapiro-wilk test could be applied on reconstructed residuals to see if the distribution is normal or not.

In this study, hyperbolic tangent function (tanh) function was used for encoding and decoding the input to the output. Then, the backpropagation was used to reconstruct the error. The possible activation functions could be Relu, tanh, or sigmoid. As data contain negative value, in case of using Relu activation function, negative values would be converted to zero which, in return, would block gradient information from learning. On the other hand, since there were negative input data, the sigmoid activation function is not appropriate as it is likely to face a vanishing gradient due to its formula $(\frac{1}{1+\exp(-x)})$.

For this model stacked autoencoder was used where the number of nodes per layer decrease from a layer to a layer and would increase back in the decoder. Number of hidden layers and number of nodes are dependent on the structure of the data, number of features and number of observations. As undersampling technique was used, resulting in a reduction of the observation, only one hidden layer was selected. On the other hand, number of nodes for input layer often follows the below equation:

$$NN=2*N+1 \tag{13}$$

Where NN is a number of nodes, and N is number of features. As there were 28 features in the data set, a value of 15 was chosen for the number of input node and output layers. The number of nodes in hidden layer was selected as slightly smaller than the number of input nodes as 10.

As the dataset become balanced, in order to identify the threshold between normal and fraudulent transaction, the quality of the prediction could be evaluated by recall or precision. In the analysis not much different was identified across these two measures so the results for precision would be presented in Fig. 2. The equation of precision can be written as follows:

$$Precision = \frac{Real\ fraud\ category \cap Predicted\ as\ fraud}{Fraud\ transaction + True\ transaction} \tag{14}$$

The threshold which would result in highest precision would be selected as a boundary between fraud and normal transaction. However, after conducting few trial and errors, a value slightly lower than an identified value in Fig. 2 was identified for the threshold. For this study a value of 0.42 was chosen, Thus MSE greater than this value would be considered as abnormal (fraud) transaction and a value less than this cutting point would be considered as normal.

C. Multivariate Outliers' Detection

Usually a traditional outlier detection refers to measuring the distance between a point and distribution to which that point belongs. A classic Mahalanobis distance is a common method of measuring this distance. This method measures how many standard deviations away a point is from the mean of a distribution. The Mahalanobis distance of an observation x=(x1, x1, x1,…, x1) from a set of observation with mean of $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_n)$ would be written as:

$$MD = \sqrt{(x-\bar{x})^T C^{-1} (x-\bar{x})} \tag{15}$$

where, C is the estimated covariance matrix. $\bar{x}$ Is the estimated multivariate location, or the multivariate arithmetic mean or the centroid.

One of the methods that could be used to identify outliers based on Mahalanobis is distance-distance plot, which plots the classical Mahalanobis distance of the data against robust Mahalanobis distance based on the minimum covariance determination (MCD). The MCD is a highly robust estimators of multivariate location and scatter [18], [26]. It should be noted one of the main advantages of MCD is its resistance to outliers which also makes its use practical for different multivariate techniques such as principal component and factor analyses. It should be noted that Mahanabolis method is applicable when all the variables are continuous. However, when comparison is conducted on categorical predictors another proposed method in the literature could be applied [27]. This method is based on simple matching coefficient (SMC) which could be used for comparing similarity and diversity of sample sets for a binary predictor as follows:

$$SMC = \frac{Number\ of\ matching\ attribuates}{Number\ of\ attribuates} = \frac{M_{00}+M_{11}}{M_{00}+M_{01}+M_{10}+M_{11}} \tag{16}$$

M's measures similarity between different binary predictors, for instance, M_{00} is the total number of attributes where both binary attributes A and B have a same value of zero.

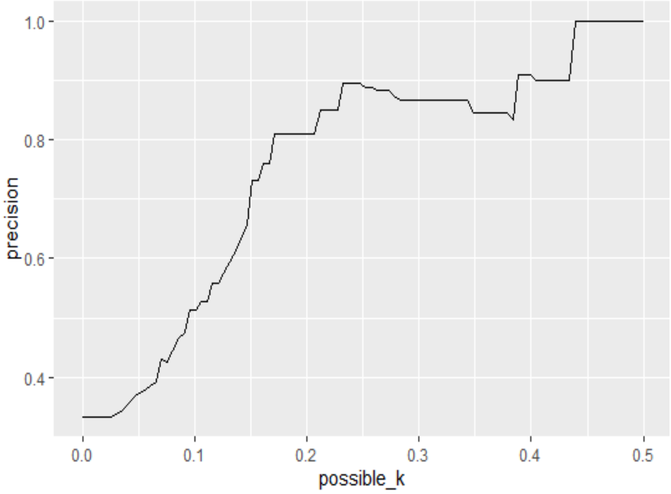

Fig. 2. Identification of Threshold for Autoencoder.

On the other hand, a Gower distance should be used for distance between two entity whose attributes have a mixed of categorical and numerical [28]. This model uses Manhattan distance for calculating distance between continuous datapoints, and Sørensen–Dice coefficient (DSC) for calculating distance between categorical datapoints. The DSC could be calculated as follows:

$$DSC = \frac{2|X\cap Y|}{|X|+|Y|} \qquad (17)$$

where, |X| and |Y| are the cardinalities of the two sets.

For autoencoder dataset was normalized with minmax normalization. For one-class SVM, normalizing the data was conducted with scale function in the syntax of the model.

V. Performance Measure

Different metrics could be implemented to evaluate the performance of different models. The performance measures were used not to compare the performance of models to come up with a better model, since these models are performing differently, but to have some ideas how each model is performing in identification of frauds observation. It should be noted that although SVM and auto-encoder needs true positive labeled data for training, Mahanabolis method does not need labeled data and this model would be conducted on the whole dataset. Therefore, it would not be a fair comparison to compare these three models.

These metrics are "true positive" (TP), "false positive" (FP), True negative (TN), and "false negative" (FN). For instance, TF represents the number of normal transaction that are classified/predicted as normal. On the other hand, false positive, are the crashes that are classified as fraud while they were normal transaction.

VI. Results

A. One-Class SVM

For this algorithm, the features were scaled and centered with a logical vector of "scale" in R. Different kernel types were available for this model including linear, and nonlinear such as polynomial and radial basis. The results of using different kernel function indicated that linear basis function kernel would result in an optimal result. This model is based on squared Euclidean distance between two features vectors. As this model needs to be trained on the normal transaction, the data labels were used and divided into two sets of data, normal versus fraudulent transactions. The model was trained on the normal dataset. It, then was examined across the whole dataset including normal and fraudulent observations.

B. Autoencoder

Keras package was used for construction of autoencoder. The construction of this model in Keras is similar to multilayer perceptron models. Similar to SVM, the model was trained over normal dataset and then tested over the whole dataset. The inputs were transformed into 15 nodes and they were transferred into two hidden layer with 10 and 15 nodes. The model, then, recreated the input from the transferred output. In the compilation section of the model, the mean square error (MSE) was set as a loss function to identify the outlier based on higher values of MSE, while adam was set as an optimizer. Due to highlighted reasons in a previous section, tanh activation functions were set for the three layers, input, hidden and output layer. For the model fit section of Keras package, the input and output were set as the normal transactions. After identification of the threshold of normal and possible fraudulent transactions, each transformed input was defined based on its value versus the threshold. The MSE of each observation (transaction) was calculated from the below equation:

$$\text{MSE}_{observation} = \sum_{i=1}^{966}\sum_{j=1}^{29}(Real\ observation - predicted\ observation)^2 \qquad (18)$$

where, i is a number of observations, which was 966 transactions, and j is a number of columns or features.

After construction of predicted input, a decision of normal versus fraudulent transactions would be made based on following equations:

Normal transaction: $\text{MSE}_{observation}$<K (19)

Fraudulent transaction: $\text{MSE}_{observation}$>K (20)

Where k is a threshold calculated from Fig. 2.

C. Multivariate Outlier Detection

This section would highlight the application of Multivariate Outlier Detection for credit card fraud detection. The ordered squared robust Mahalanobis distances of the observations against the empirical distribution function is presented in Fig. 3. The Mahalanobis distance of the data point against the robust Mahalanobis distance was plotted based on MCD estimators [29].

Alpha, amount of observations used for calculating the adjusted quantile, was set as .80. Value of 0.8 was chosen as a lower value was resulting in computationally singular output. This resulted from invertible design matrix which in return is due to multicollinearity across predictors being used for MCD estimations.

On the other hand, amount of observation used for calculating the adjusted quantile was also set as 0.8. This value, Adjusted quantile, is a new threshold that separate outliers from non-outliers [30]. An approximation of 97.5 percentile would be obtained by estimation of mean and standard deviation of each variable and computing the values of mean ±standard deviation [30]. An observation would be considered as an outlier if that observation falls in this extreme 2%. In Fig. 3, the horizontal and vertical lines are plotted at values equal to the cutoff, where the default is square root of the 97.5 distribution.

Table I presented different error rates for different categories across the three implemented machine learning techniques. This study is not discussing the input of Table I in detail as these tree models are performing differently. It should be noted although robust Mahanabolis resulted in a worst performance contrary to the other two methods this method did not used any label for training.

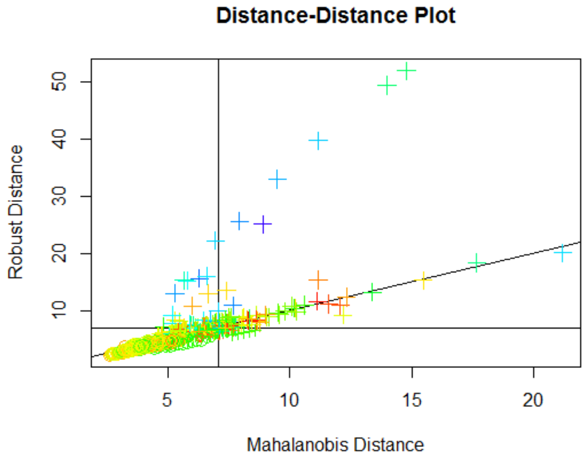

Fig. 3. Distance-Distance Plot: the Robust Distance Versus the Classical Mahalanobis Distance.

TABLE. I. ERROR RATE OF DIFFERENT INCLUDED MODELS

One-class SVM			Autoencoder			Robust Mahanabolis		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
	0	1		0	1		0	1
0	279	195	0	430	44	0	360	114
1	67	425	1	63	429	1	245	247

VII. CONCLUSION

Due to huge loss to banks, individuals and insurance companies, credit card fraud detection is considered as one of most explored domains of fraud detection. In data evaluation, anomaly is referred to any observation that does not conform to the expected distribution/pattern of the other items. At the age of computer, this could refer to an adverse event such as network intrusion, bank or credit frauds. The problem with supervised leaning techniques is that they need labels for all the observations to predict the future transactions. This would create a problem when fraud transactions need to be detected and no label is available for these observations. Moreover, fraudsters change their habit constantly which make it difficult for supervised techniques to be prepared for those transactions. However, unsupervised techniques need only labels for one-class, usually normal class, and it could predict the future observations based on distance from normal observations.

On the other hand, some supervised learning techniques even do not need the label for one-class, and they can identify the outliers on the whole dataset with no labels. In this study we took advantages of the aforementioned techniques and we used the available data labels to check the model performance.

This paper presents application of three unsupervised methods in detection of credit card frauds. For unsupervised methods SVM and autoencoder the training would be achieved from past normal transactions to predict future transactions, normal versus fraud. However, the advantage of mahalanobis method over the other two method is that this method does not need to be trained on labeled data and it can identify the anomalies just based on the minimum covariance determinant matrix. As the performance and training the three models are different, no comparison was made across these three models. However, to have a vision about the performance of these models the available labels were used for models' performance evaluations.

For the future studies the information related to cardholder behaviors and their historical transaction history need to be taken into consideration for achieving a higher accuracy. In other words, both global and local outliers need to be considered for those studies.

REFERENCES

[1] P. K. Chan et al, "Distributed data mining in credit card fraud detection," IEEE Intelligent Systems, (6), pp. 67-74, 1999.

[2] A. Hobson, The Oxford Dictionary of Difficult Words. 2004.

[3] Y. G. Şahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines," 2011.

[4] X. Niu, L. Wang and X. Yang, "A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised," arXiv Preprint arXiv:1904.10604, 2019.

[5] A. Srivastava et al, "Credit card fraud detection using hidden Markov model," IEEE Transactions on Dependable and Secure Computing, vol. 5, (1), pp. 37-48, 2008.

[6] F. Carcillo et al, "Combining unsupervised and supervised learning in credit card fraud detection," Inf. Sci., 2019.

[7] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," Credit Scoring and Credit Control VII, pp. 235-255, 2001.

[8] S. Jiang et al, "A clustering-based method for unsupervised intrusion detections," Pattern Recog. Lett., vol. 27, (7), pp. 802-810, 2006.

[9] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," Inf. Sci., vol. 177, (18), pp. 3799-3821, 2007.

[10] Y. Tian et al, "Ramp loss one-class support vector machine; A robust and effective approach to anomaly detection problems," Neurocomputing, vol. 310, pp. 223-235, 2018.

[11] S. Bhattacharyya et al, "Data mining for credit card fraud: A comparative study," Decis. Support Syst., vol. 50, (3), pp. 602-613, 2011.

[12] M. Hejazi and Y. P. Singh, "One-class support vector machines approach to anomaly detection," Appl. Artif. Intell., vol. 27, (5), pp. 351-366, 2013.

[13] C. Fan et al, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data," Appl. Energy, vol. 211, pp. 1123-1135, 2018.

[14] T. Sweers, T. Heskes and J. Krijthe, "Autoencoding Credit Card Fraud," 2018.

[15] P. Filzmoser, R. G. Garrett and C. Reimann, "Multivariate outlier detection in exploration geochemistry," Comput. Geosci., vol. 31, (5), pp. 579-587, 2005.

[16] P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage points," Journal of the American Statistical Association, vol. 85, (411), pp. 633-639, 1990.

[17] F. R. Hampel et al, Robust Statistics: The Approach Based on Influence Functions. 2011196.

[18] P. J. Rousseeuw, "Multivariate estimation with high breakdown point," Mathematical Statistics and Applications, vol. 8, (283-297), pp. 37, 1985.

[19] P. Filzmoser, K. Hron and C. Reimann, "Interpretation of multivariate outliers for compositional data," Comput. Geosci., vol. 39, pp. 77-85, 2012.

[20] P. Harris et al, "Multivariate spatial outlier detection using robust geographically weighted methods," Mathematical Geosciences, vol. 46, (1), pp. 1-31, 2014.

[21] M. Krivko, "A hybrid model for plastic card fraud detection systems," Expert Syst. Appl., vol. 37, (8), pp. 6070-6076, 2010.

[22] J. Van Hulse, T. M. Khoshgoftaar and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in Proceedings of the 24th International Conference on Machine Learning, 2007,.

[23] B. Schölkopf et al, "Estimating the support of a high-dimensional distribution," Neural Comput., vol. 13, (7), pp. 1443-1471, 2001.

[24] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning. 2016.

[25] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning. 2016.

[26] P. J. Rousseeuw, "Least median of squares regression," Journal of the American Statistical Association, vol. 79, (388), pp. 871-880, 1984.

[27] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining." DMKD, vol. 3, (8), pp. 34-39, 1997.

[28] J. C. Gower, "A general coefficient of similarity and some of its properties," Biometrics, pp. 857-871, 1971.

[29] P. Filzmoser, R. Maronna and M. Werner, "Outlier identification in high dimensions," Comput. Stat. Data Anal., vol. 52, (3), pp. 1694-1711, 2008.

[30] P. Filzmoser, R. G. Garrett and C. Reimann, "Multivariate outlier detection in exploration geochemistry," Comput. Geosci., vol. 31, (5), pp. 579-587, 2005.