A Result Paper on Anomaly Detection in Credit Card Transactions Using Machine Learning

Kavita Bisht[1], Mr. Ashok Kaushik[2], Mr. Shashi Ranjan Parmar[3]

[1]Research Associate Department of Computer Science, Bhiwani Institute of Technology & Science, Bhiwani, Haryana
[2]HOD, Department of Computer Science, Bhiwani Institute of Technology & Science, Bhiwani, Haryana
[3]Chairperson Department of Computer Science, Bhiwani Institute of Technology & Science, Bhiwani, Haryana

ABSTRACT

In the last years, credit and debit cards usage has significantly increased. However a non negligible part of the credit card transactions are fraudulent and billions of Euros are stolen every year throughout the world. Credit card fraud detection presents several characteristics that make it a challenging task. First, the feature set describing a credit card transaction usually ignores detailed sequential information which was proven to be very relevant for the detection of credit card fraudulent transactions. Second, purchase behaviors and fraudster strategies may change over time, making a learnt fraud detection decision function irrelevant if not updated. This phenomenon named dataset shift (change in the distribution p(x, y)) may hinder fraud detection systems to obtain good performances. We conducted an exploratory analysis in order to quantify the day by day dataset shift and identified calendar related time periods that show different properties. Third, credit card transactions data suffer from a strong imbalance regarding the class labels which needs to be considered either from the classifier perspective or from the data perspective (less than 1% of the transactions are fraudulent transactions). To conclude, this work leads to a better understanding of what can be considered contextual knowledge for a credit card fraud detection task and how to include it in the classification task in order to get an increase in fraud detection. The method proposed can be extended to any supervised task with sequential datasets.

Keywords: Credit Card, Anomaly Detection, Machine Learning, Fraud Detection, Big Data.

1 INTRODUCTION

In this era of internet the trading all around the world is developed, the main motivation behind the quick development is exchanging. Exchanging through the web is supposed to be a web-based business, where merchandise gets traded through online administrations. Fundamentally, online business is a stage where individuals get electronic URLs as their shop for buying merchandise and things to purchase or pay through the web. For purchasing they need to pay cash with that they pay through various administrations accessible most likely through Visas or charge cards with these two cards we can pay cash to the broker yet when cards come into the play it carries another term with it that is network safety.

It is common fraud, widespread around the world. A substantial amount may be withdrawn in a short span of time without the owner's acknowledgement that’s why it simple fraud for fraudsters without causing any risks. Fraudsters put lots of efforts to do fraudulent transactions look legitimate, making fraud detection difficult and time-consuming.

i. What are Anomalies?

Inconsistencies are designs in the information that don't adjust to a well-defined thought of ordinary conduct. Figure 1 shows oddities in a straightforward 2-dimensional informational index. The information has two ordinary areas, X and Y since most perceptions lie in these

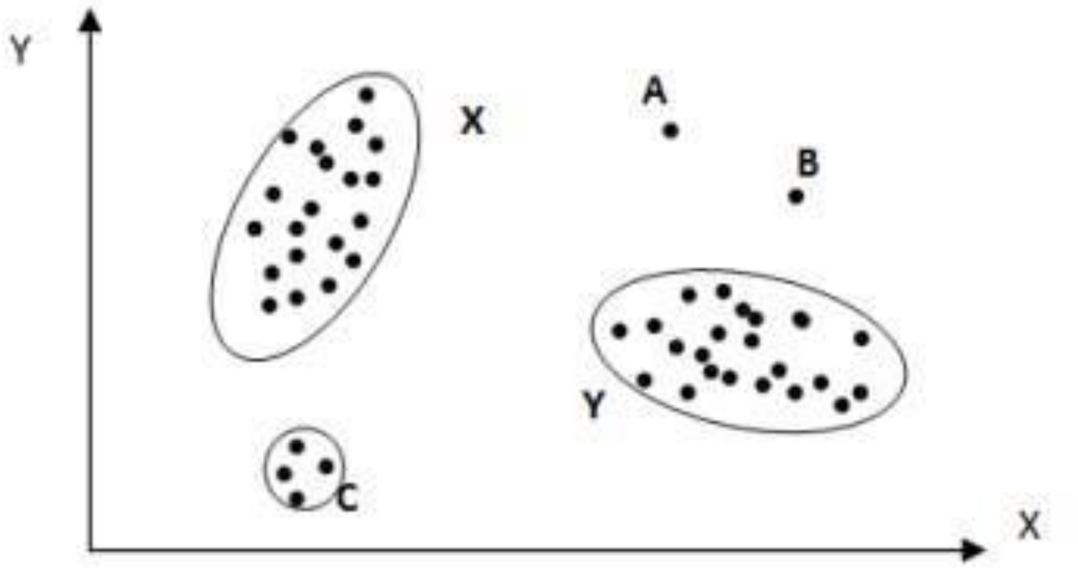


Fig 1. demonstrates an example of Point Anomalies

For a variety of causes, such as vengeful action, e.g., Visa extortion, digital disruption, fear mongering action, or the collapse of a system, peculiarities may be introduced in the data; however, all of the explanations have one thing in common: they all pique the examiner's interest. A major highlight is the "intriguing character" or true significance of inconsistencies. The "intriguing quality" or genuine significance of inconsistencies is a key highlight of peculiarity identification.

ii. Types of Anomalies

As per Chandola u. a. (2009), irregularities can be partitioned into three classes:

• Point is the most straightforward sort of abnormality. Given that perception is portrayed by a few highlights (credits, attributes), the point abnormality just influences one of them, in this manner treating any element autonomously of the others. For example, a credit choice is made dependent on the candidate's compensation as it were. A compensation is, accordingly, a component to consider. A pay that is very high regarding the remainder of compensations is in this way the point inconsistency.

• Context is an expansion of the point abnormality when a few highlights are considered without a moment's delay, in this way expecting certain reliance among highlights. By expanding the model above, pay as well as a nation of home is considered. In this case, compensation is connected to the nation where an advance candidate resides: a compensation of 10,000 rupees can be viewed as too high in one nation while in another country a similar sum compares to a common (normal) pay.

• Collective irregularities can be recognized by noticing an assortment of perceptions. The setting irregularity actually sees every perception to some degree in detachment from different perceptions. The aggregate abnormality shows itself just when a couple of perceptions are together broke down. Every perception without anyone else doesn't look odd, yet a gathering of such perceptions that happened together comprise the peculiarity. Usually such a gathering is framed while considering the time measurement. For instance, someone didn't yet reimburse one advance however as of now got the following one, despite the fact that the standards unequivocally preclude this. In disconnection, getting credit is an ordinary occasion, however getting two advances when the first advance isn't yet completely reimbursed is probably going to be strange. The aggregate oddity is out of the extension in this work as it requires approaches radically not quite the same as those used to address the initial two sorts of peculiarities.

iii. Key Components of Anomalies.

Analysts have embraced ideas from various teaches like measurements, AI, information mining, data hypothesis, unearthly hypothesis, and have applied them to specific issue plans. Figure 1.3 shows the previously mentioned key parts related to any abnormality identification method.

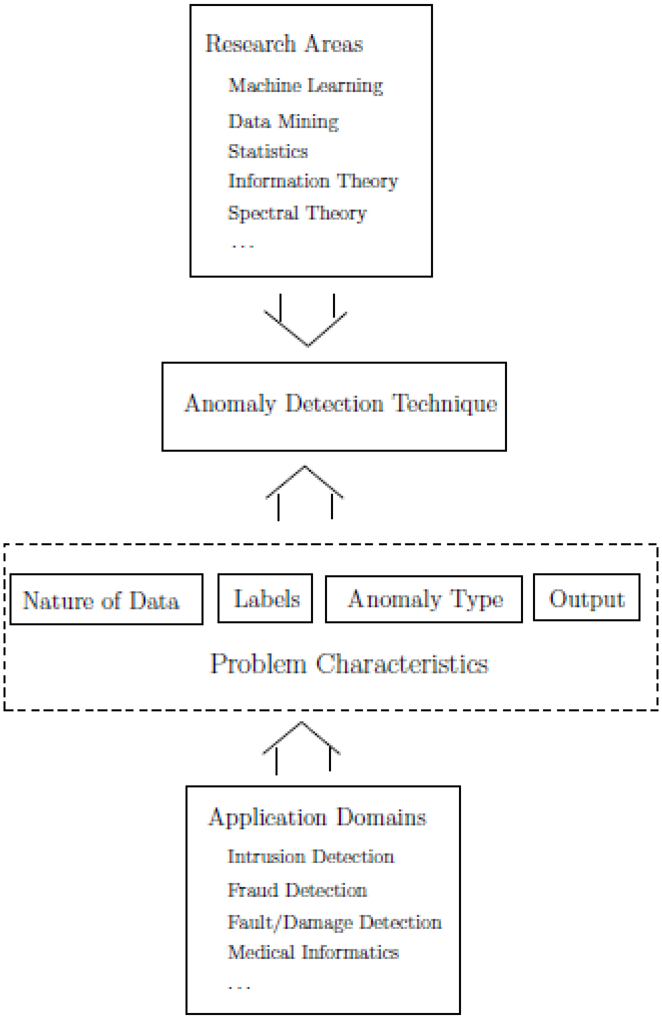


Fig 2. key Components in detecting anomalies

iv. Comparison of Classification and prediction Methods:

1. **Accuracy:** The accuracy or correctness of a classifier indicates that how well the given classifier could classify the new test tuples those are new to the classifier itself. In case of prediction, the accuracy works on the norms of how well the classifier could predicts the value of a new test case that too also unknown to the classifier. More the size of the test and training data more will be the accuracy of the classifier.
2. **Speed:** This usually indicated the computational cost joints in using the provided prediction or classifier.
3. **Robustness:** this term indicates to provide the correct prediction while given a noisy or missing data.
4. **Scalability:** The buildup module would be applicable and capable enough to suit on the huge amount of data also.
5. **Interpretability:** this is usually the understanding and insight level that is given by the predictor or classifier.

2 METHODS AND MATERIALS

The initial step of this examination was to utilize an accessible news keywords dataset and contrast distinctive AI calculations with comprehend their presentation dependent on various execution measurements.

In the subsequent advance, scope for the examination and its motivation was characterized. The extent of the examination was characterized according to consider reason, assets and calendar. The examinations design was to comprehend the AI calculations conduct specifically informational indexes and to attempt to induce the outcomes. In the third step, the essential information to begin with the examination was characterized .For that few online assets were utilized and basics for the investigation were comprehended.

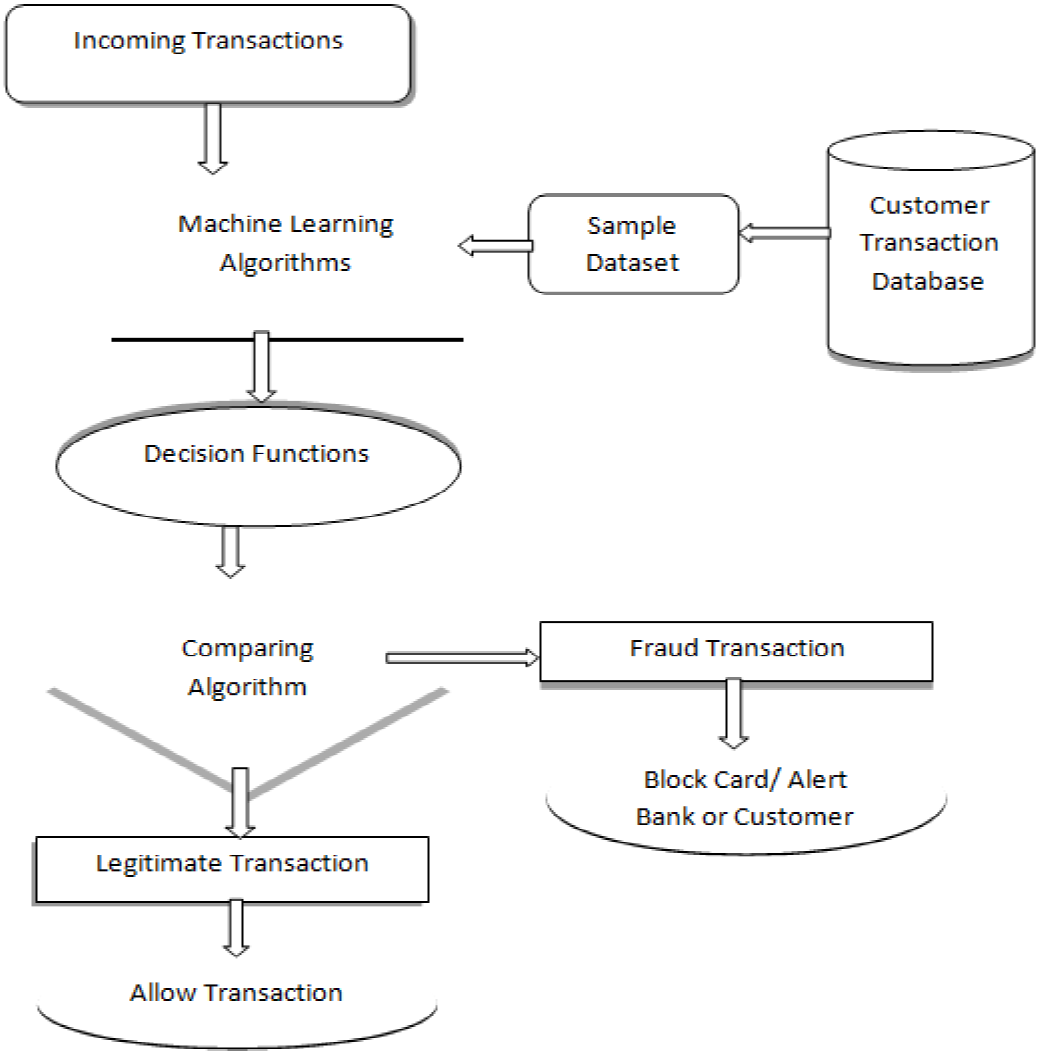
In fourth step, information was gathered online from AI vault.

In the fifth step, elucidating and exploratory information investigation just as information cleaning was done according to the information picked up from step3.It assisted with understanding the information all the more, for example, highlights notwithstanding ,connection between's highlights, missing qualities and anomalies .

In the 6th stage, distinctive writing was contemplated identified with this informational collection .The writing was accessible on the web and the discoveries of the written works were assessed.

In the seventh stage, diverse AI calculations were prepared and the outcomes were acquired according to various execution measurements.

In the eighth stage, results were deciphered and contrasted and other existing writing regarding the matter .In ninth and last stage, ends were determined dependent on the outcomes got



i. Taxonomies of Supervised and Unsupervised ML Algorithms

Most of useful ML utilizes administered learning. Regulated learning is a procedure of taking in a calculation from the preparation dataset where the information factors and yield factors are accessible. The summed up scientific classification of administered and unaided AI calculations is given in Figure. Likewise, Table 1 gives a summed up examination of the premise and prominent properties, just as focal points and restrictions for every calculation of the sub-spaces of a regulated and solo ML. In the accompanying subsections, a nitty-gritty conversation is introduced.

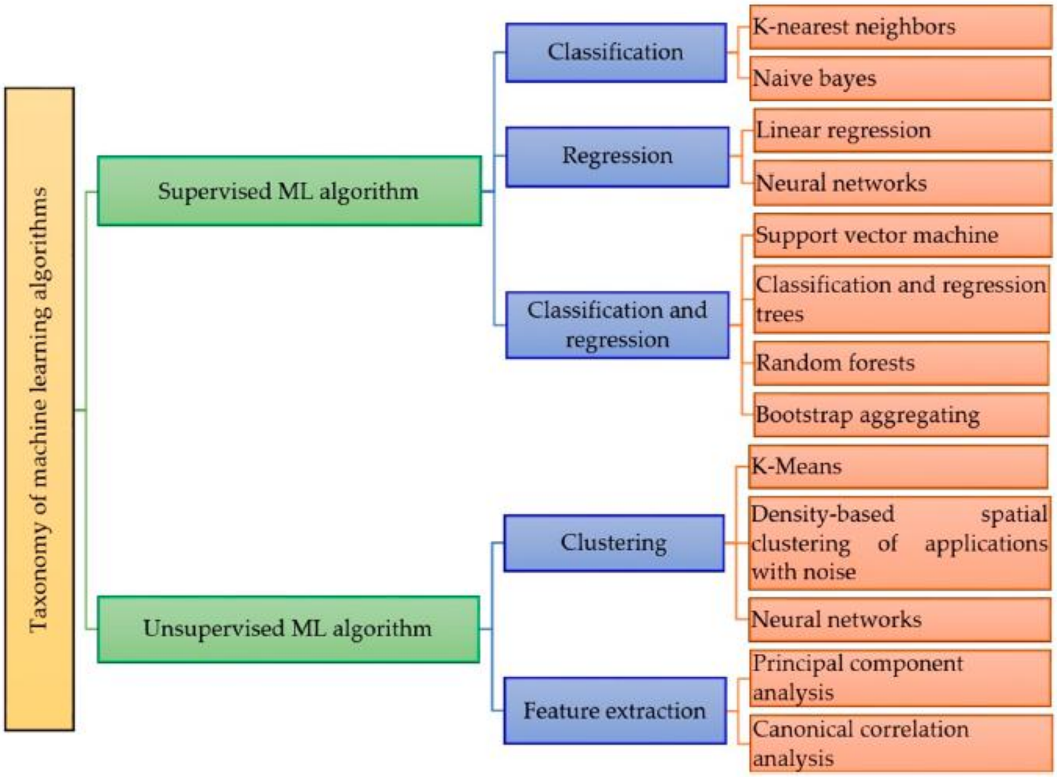
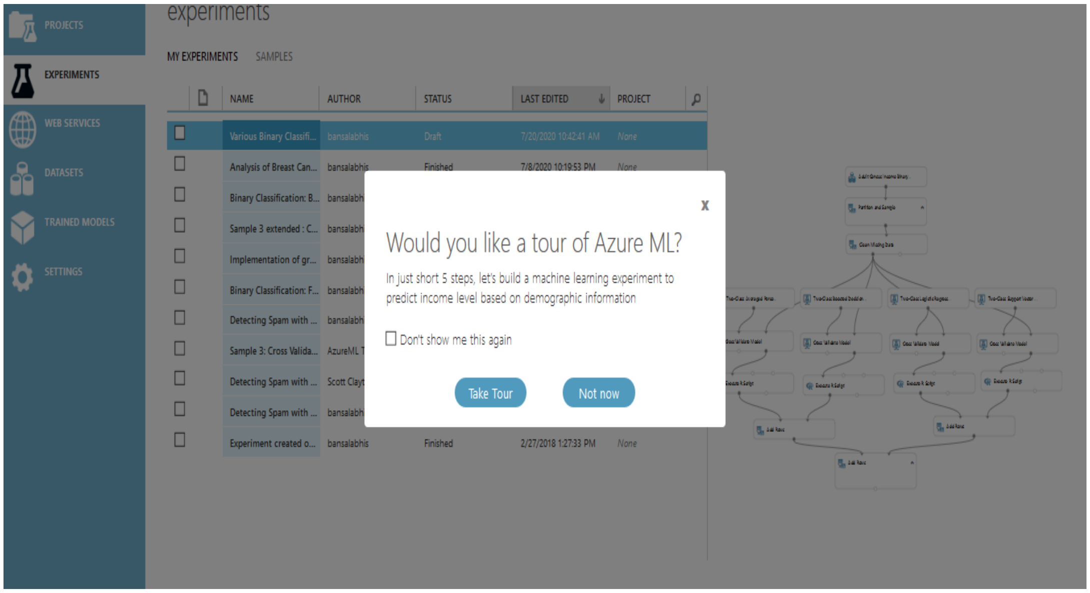

Taxonomy of machine learning algorithms
Supervised ML algorithm
Unsupervised ML algorithm
Classification
Regression
Classification and regression
Clustering
Feature extraction
K-nearest neighbors
Naive bayes
Linear regression
Neural networks
Support vector machine
Classification and regression trees
Random forests
Bootstrap aggregating
K-Means
Density-based spatial clustering of applications with noise
Neural networks
Principal component analysis
Canonical correlation analysis

Fig 3. taxonomy of machine learning algorithms

ii. **Microsoft Azure**

The Microsoft Azure stage was utilized to lead all tests on classifiers clarified in this paper. Sky blue Machine Learning is a cloud-based condition that you can use to prepare, execute, computerize, oversee, and track ML models. Microsoft Azure Machine Learning can be utilized for any sort of AI, from old style ml to profound learning, unaided, and solo learning. Regardless of whether you like to compose zero-code/low-code choices like Python or R code or Designer, you can assemble, train and track the most exact AI and profound learning models in the Azure Machine Learning Workspace



	NAME	AUTHOR	STATUS	LAST EDITED	PROJECT
□	Various Binary Classifi...	bansalabhis	Draft	7/20/2020 10:42:41 AM	None
□	Analysis of Breast Can...	bansalabhis	Finished	7/8/2020 10:19:53 PM	None
□	Binary Classification: B...	bansalabhi			
□	Sample 3 extended : C...	bansalabhi			
□	Implementation of gr...	bansalabhi			
□	Binary Classification: F...	bansalabhi			
□	Detecting Spam with ...	bansalabhi			
□	Sample 3: Cross Valida...	AzureML T			
□	Detecting Spam with ...	Scott Clayt			
□	Detecting Spam with ...	bansalabhi			
□	Experiment created o...	bansalabhis	Finished	2/27/2018 1:27:33 PM	None

Would you like a tour of Azure ML?

In just short 5 steps, let's build a machine learning experiment to predict income level based on demographic information

□ Don't show me this again

Fig 4. Interface of Microsoft Azure

iii. **Process Cycle:**

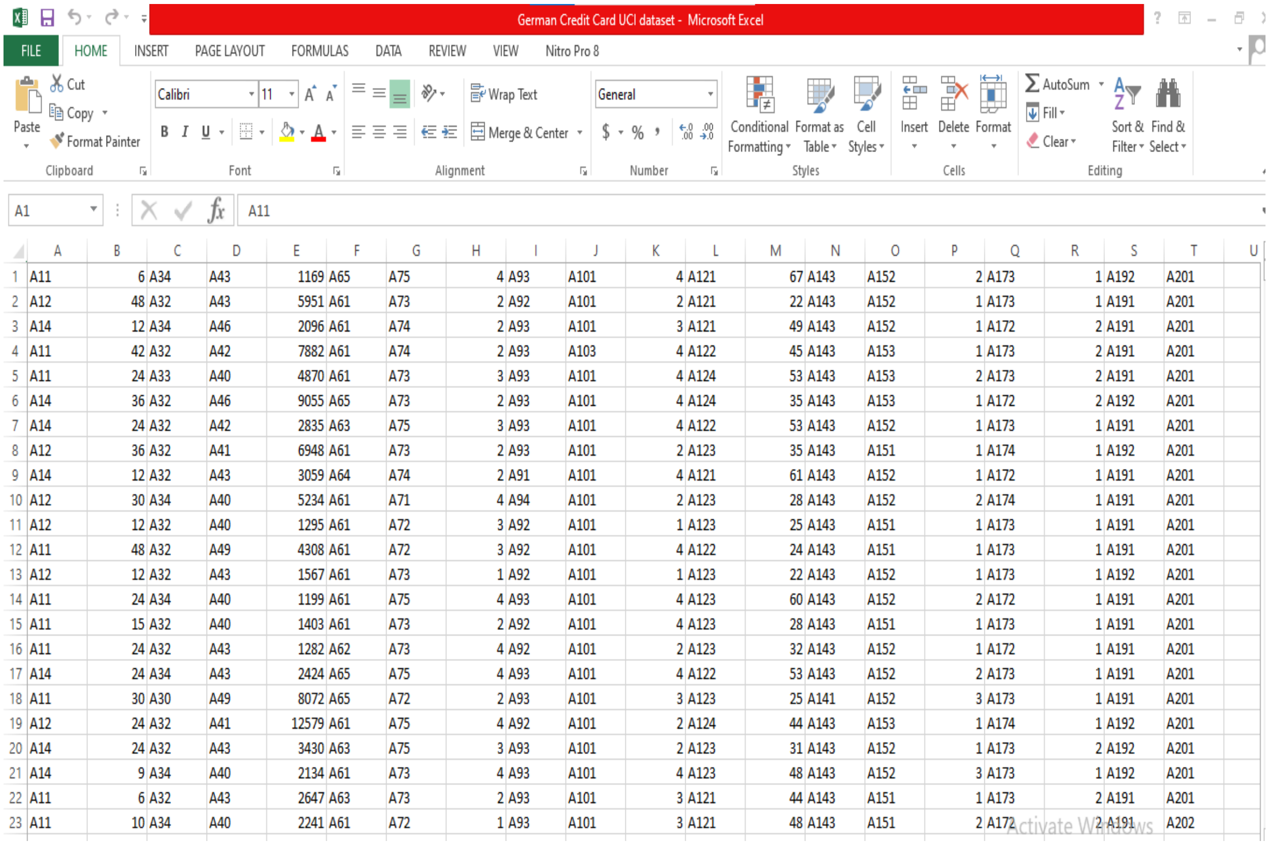
1. Alter Metadata: Assign col21 as Label

2. Split Information: Data is divided multiple times. (a) The first occasion when it is STRATIFIED 75/25: Train/Test split. (b) There is two Second Split, the principal split is again 75/25: Train/Test proportion. The subsequent split uses Regular Expression Split, \"Label" ^1
3. This one likewise follows the Regular Expression Split, \"Label" ^1.
4. Convey One-Class Support Vector Machine (SVM) and Principal Component Analysis (PCA) Based Anomaly Detection modules to make two distinct abnormality recognition models.
5. In the two cases apply Tune Model Hyper parameters and select "Whole Grid" and F-Score and Mean Square Error as measurements.
6. Convey Train Anomaly Detection.
7. Convey Score Models and pass pertinent information.
8. Standardize Data: Transforms information through standardization to a typical scale, without misshaping contrasts in the scopes of qualities or losing data. Here the Logistic change is applied tp Score Probabilities.
9. Assessment Model: Finally the models are assessed as far as ROC/AUC and disarray measurements.

3. PROPOSED METHODOLOGY:

i. Select Dataset

This dataset contains various details regarding credit card transactions of various companies. In this research work dataset of German Credit card UCI dataset is used. The dataset contains various attributes representing with numeral data for example A11 to A14 is a qualitative attribute in which status of existing checking accounts represented with various integral numbers (i.e. A14 No checking Account), Attribute 2 representing Numerical attribute i.e. Duration in month in that manner this database have lots of attributes with some detailed information.



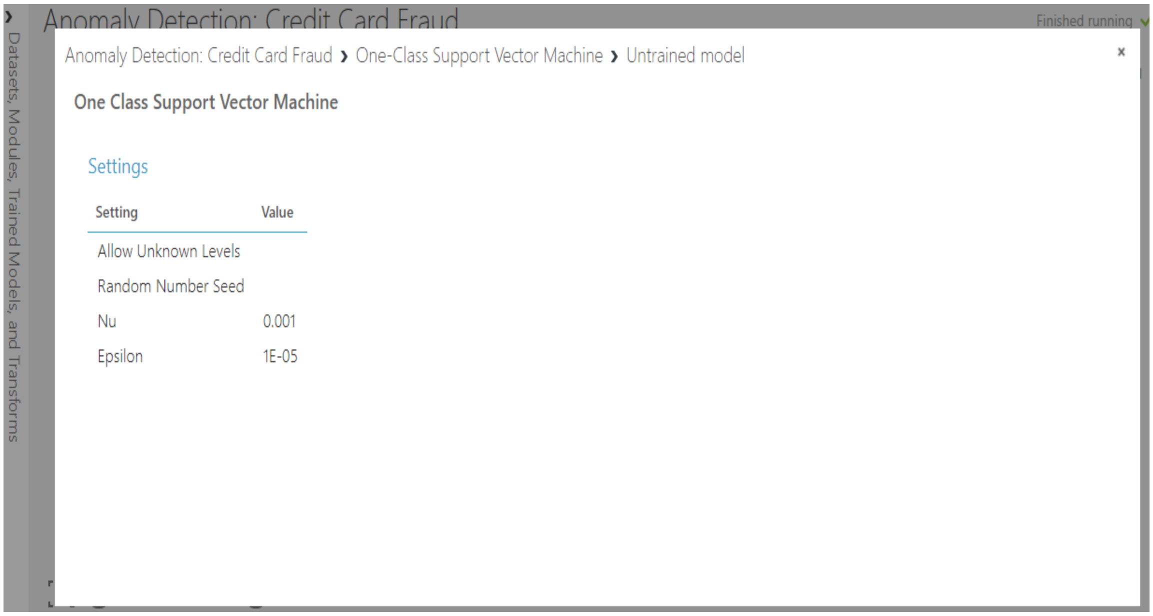
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201	
2	A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201	
3	A14	12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201	
4	A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	
5	A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201	
6	A14	36	A32	A46	9055	A65	A73	2	A93	A101	4	A124	35	A143	A153	1	A172	2	A192	A201	
7	A14	24	A32	A42	2835	A63	A75	3	A93	A101	4	A122	53	A143	A152	1	A173	1	A191	A201	
8	A12	36	A32	A41	6948	A61	A73	2	A93	A101	2	A123	35	A143	A151	1	A174	1	A192	A201	
9	A14	12	A32	A43	3059	A64	A74	2	A91	A101	4	A121	61	A143	A152	1	A172	1	A191	A201	
10	A12	30	A34	A40	5234	A61	A71	4	A94	A101	2	A123	28	A143	A152	2	A174	1	A191	A201	
11	A12	12	A32	A40	1295	A61	A72	3	A92	A101	1	A123	25	A143	A151	1	A173	1	A191	A201	
12	A11	48	A32	A49	4308	A61	A72	3	A92	A101	4	A122	24	A143	A151	1	A173	1	A191	A201	
13	A12	12	A32	A43	1567	A61	A73	1	A92	A101	1	A123	22	A143	A152	1	A173	1	A192	A201	
14	A11	24	A34	A40	1199	A61	A75	4	A93	A101	4	A123	60	A143	A152	2	A172	1	A191	A201	
15	A11	15	A32	A40	1403	A61	A73	2	A92	A101	4	A123	28	A143	A151	1	A173	1	A191	A201	
16	A11	24	A32	A43	1282	A62	A73	4	A92	A101	2	A123	32	A143	A152	1	A172	1	A191	A201	
17	A14	24	A34	A43	2424	A65	A75	4	A93	A101	4	A122	53	A143	A152	2	A173	1	A191	A201	
18	A11	30	A30	A49	8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A201	
19	A12	24	A32	A41	12579	A61	A75	4	A92	A101	2	A124	44	A143	A153	1	A174	1	A192	A201	
20	A14	24	A32	A43	3430	A63	A75	3	A93	A101	2	A123	31	A143	A152	1	A173	2	A192	A201	
21	A14	9	A34	A40	2134	A61	A73	4	A93	A101	4	A123	48	A143	A152	3	A173	1	A192	A201	
22	A11	6	A32	A43	2647	A63	A73	2	A93	A101	3	A121	44	A143	A151	1	A173	2	A191	A201	
23	A11	10	A34	A40	2241	A61	A72	1	A93	A101	3	A121	48	A143	A151	2	A172	2	A191	A202	

Figure 5. Dataset in .csv format for implementation on Azure

ii. Implemented Algorithms

a. One Class Support Vector Machine:

Support vector machines (SVMs) are administered learning models that examine the information and perceive designs, and that can be utilized for both grouping and relapse errands. Commonly, the SVM calculation is given a bunch of preparing models marked as having a place with one of two classes. A SVM model depends on partitioning the preparation test focuses into independent classifications by as wide a hole as could really be expected while punishing preparing tests that fall on some unacceptable side of the hole. The SVM model then, at that point makes expectations by doling out focuses aside of the hole or the other.


Anomaly Detection: Credit Card Fraud
Finished running ✓
Datasets, Modules, Trained Models, and Transforms
×

Anomaly Detection: Credit Card Fraud › One-Class Support Vector Machine › Untrained model

One Class Support Vector Machine

Settings

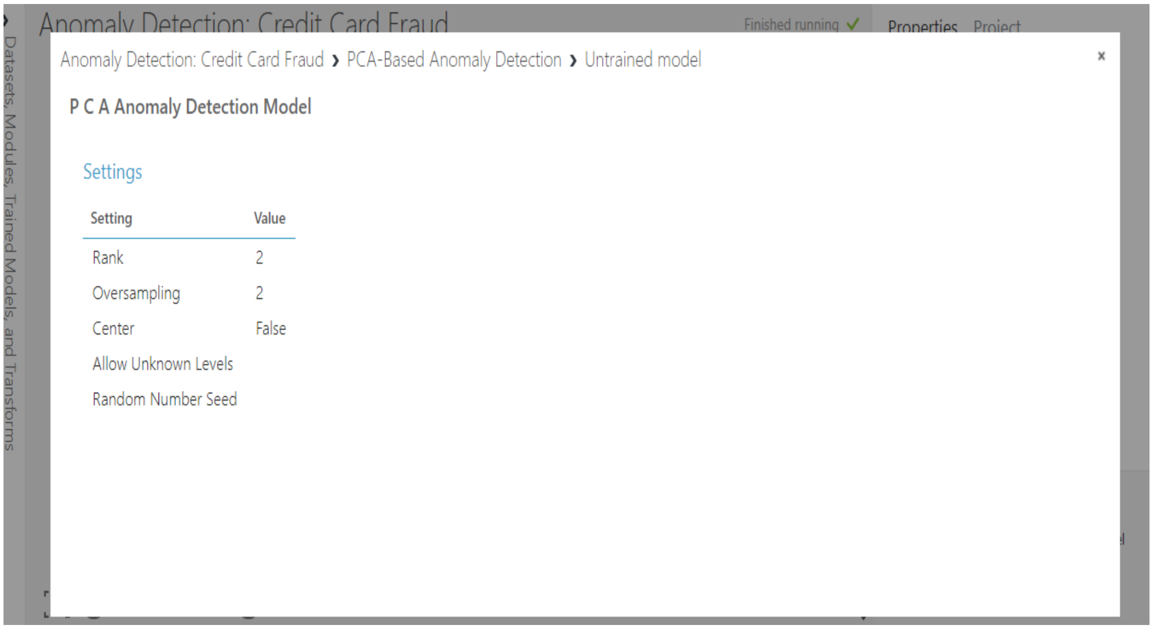
Setting	Value
Allow Unknown Levels	
Random Number Seed	
Nu	0.001
Epsilon	1E-05

Fig. 6. One Class Support Vector Machine in Anomaly Detection

b. PCA Based Anomaly Detection Algorithm

Principal Component Analysis, which is regularly condensed to PCA, is a setup method in AI. PCA is as often as possible utilized in exploratory information analysis since it uncovers the internal design of the information and clarifies the difference in the information.

For oddity recognition, each new information is investigated, and the irregularity location calculation processes its projection on the eigenvectors, along with a standardized remaking mistake. The standardized blunder is utilized as the peculiarity score. The higher the blunder, the more peculiar the occurrence is.


Anomaly Detection: Credit Card Fraud
Finished running ✓
Properties
Project
Datasets, Modules, Trained Models, and Transforms
×

Anomaly Detection: Credit Card Fraud › PCA-Based Anomaly Detection › Untrained model

P C A Anomaly Detection Model

Settings

Setting	Value
Rank	2
Oversampling	2
Center	False
Allow Unknown Levels	
Random Number Seed	

Fig. 7. PCA Based Anomaly Detection in Credit Card Fraud

4. EVALUATE MODEL

We utilize assess model to quantify the precision of a prepared model. This model will process a set of standard assessment measurements for this we give an informational collection containing scores created from a model. Assess demonstrate restore the frameworks that are relied upon sort of model that we use for assessment these different kinds of models are:

- Classification display
- Clustering model
- Regression display

Utilize preparing information

We interface a dataset that contains a set of contributions to assess a model and if no information is accessible we can utilize unique information.

Utilize the testing information:

We utilize a split module to isolate the informational collection into preparing and testing informational collection.

Think about the score of different models:

Assess show is valuable to think about the consequence of different models on similar information, and the score may be a shared assessment set or set of result from various machine learning method on the similar informational index.

Measurements for arrangement models

While assessing arrangement display the accompanying measurements are accounted for. By utilizing these metrics we look at different models and discover which show accomplish the best outcome for an order of spam or ham.

- **Accuracy**: this will gauge the level of the right consequence of an order to demonstrate.
- **Precision:** this is a level of genuine forecast that is right.
- **Recall:** this s a small amount of positive occurrence that was anticipated as positive and give all the right outcome returned by demonstrating.
- **F-Score:** it is figured as the heaviness of accuracy and reviews normally.

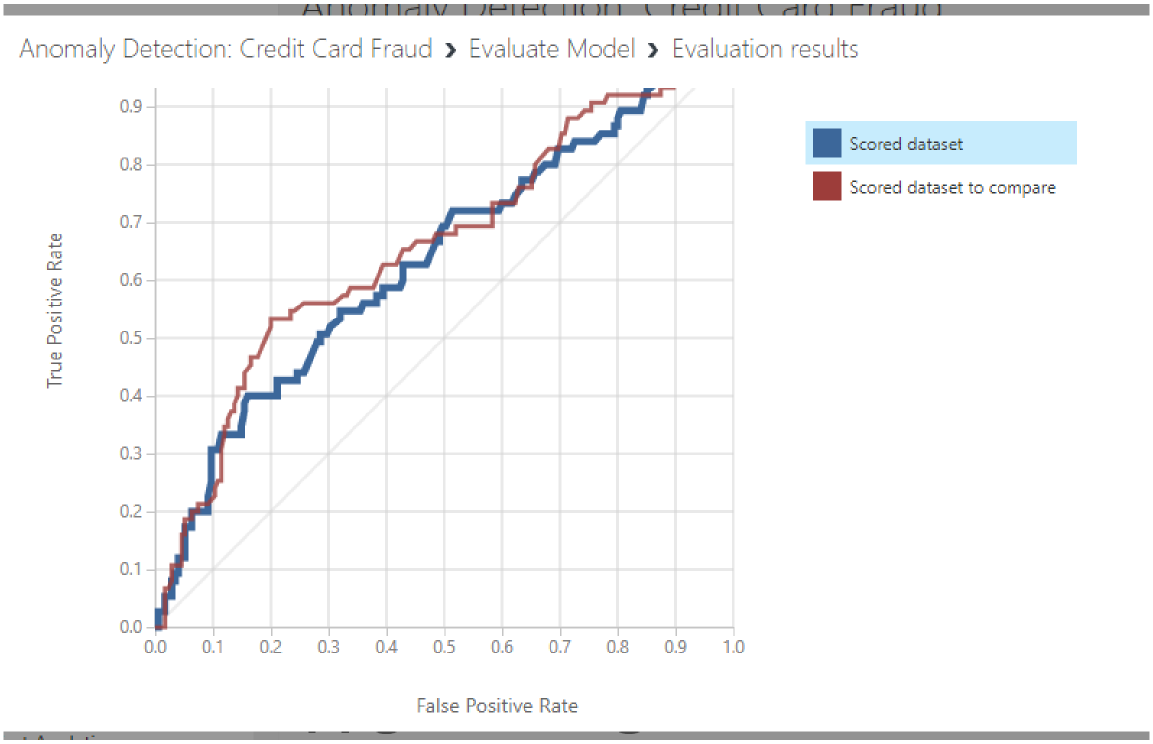
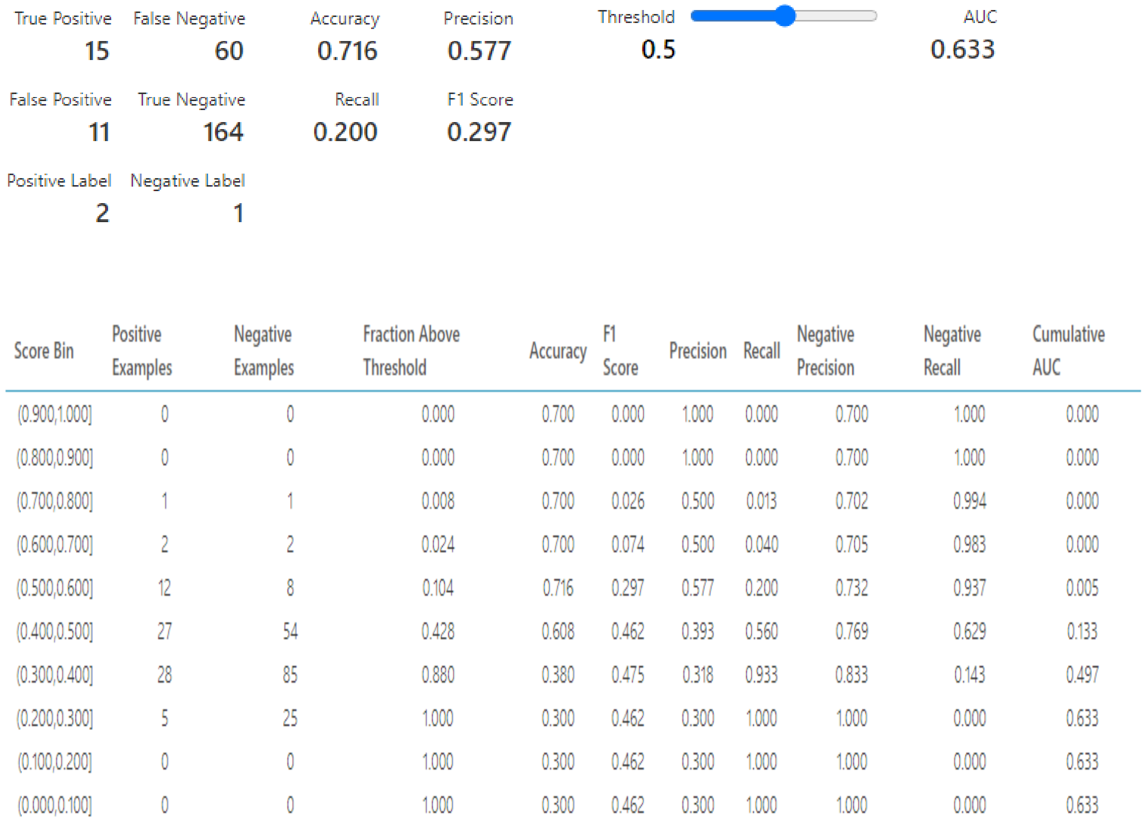


Fig.8. ROC PRECISION/RECALL LIFT

.



True Positive 15	False Negative 60	Accuracy 0.716	Precision 0.577	Threshold 0.5	AUC 0.633
False Positive 11	True Negative 164	Recall 0.200	F1 Score 0.297		
Positive Label 2	Negative Label 1				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.700	0.000	1.000	0.000	0.700	1.000	0.000
(0.800,0.900]	0	0	0.000	0.700	0.000	1.000	0.000	0.700	1.000	0.000
(0.700,0.800]	1	1	0.008	0.700	0.026	0.500	0.013	0.702	0.994	0.000
(0.600,0.700]	2	2	0.024	0.700	0.074	0.500	0.040	0.705	0.983	0.000
(0.500,0.600]	12	8	0.104	0.716	0.297	0.577	0.200	0.732	0.937	0.005
(0.400,0.500]	27	54	0.428	0.608	0.462	0.393	0.560	0.769	0.629	0.133
(0.300,0.400]	28	85	0.880	0.380	0.475	0.318	0.933	0.833	0.143	0.497
(0.200,0.300]	5	25	1.000	0.300	0.462	0.300	1.000	1.000	0.000	0.633
(0.100,0.200]	0	0	1.000	0.300	0.462	0.300	1.000	1.000	0.000	0.633
(0.000,0.100]	0	0	1.000	0.300	0.462	0.300	1.000	1.000	0.000	0.633

Fig 9. Result Table

CONCLUSION

We present a procedure in this undertaking that displays an stunning capacity to recognize oddities and simple inliers by making a few choice trees for each datum point. For better assessment of our procedure, we use Area under Accuracy Recall bend (AUC), which shows better outcomes than the Area under ROC bend. In conclusion, we show the effectiveness of our methodology in an extortion identification model seen to be 71.6 %, which demonstrates an altogether preferable methodology over other extortion location procedures. The solitary limit to the misrepresentation location framework is the inaccessibility of the reasonable dataset for preparing purposes furthermore, the lack of the dataset. On the off chance that the monetary establishments make accessible the basic informational collection of different false exercises, the examination result will be more effective and subjective.

REFERENCES

.

[1]. R. Kumari, Sheetanshu, M. K. Singh, R. Jha and N. K. Singh, “Anomaly detection in network traffic using K-mean clustering,” 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, 2016, pp. 387-393. doi: 10.1109/RAIT.2016.7507933.

[2]. S. Papadopoulos, A. Drosou and D. Tzovaras, “A Novel Graph-Based Descriptor for the Detection of Billing-Related Anomalies in Cellular Mobile Networks,” in IEEE Transactions on Mobile Computing, vol. 15, no. 11, pp. 2655-2668, Nov. 1 2016. doi: 10.1109/TMC.2016.2518668

[3]. S. Zhang, B. Li, J. Li, M. Zhang and Y. Chen, “A Novel Anomaly Detection Approach for Mitigating Web-Based Attacks Against Clouds,” 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing, New York, NY, 2015, pp. 289-294. doi: 10.1109/CSCloud.2015.46

[4]. K. I. Hassan and A. Abraham, “Modeling consumer loan default prediction using neural netware,” 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE), Khartoum, 2013, pp. 239-243. doi: 10.1109/ICCEEE.2013.6633940

[5]. Y. Jin and Y. Zhu, "A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, 2015, pp. 609-613.doi: 10.1109/CSNT.2015.25

[6]. J. Xu, D. Chen and M. Chau, "Identifying features for detecting fraudulent loan requests on P2P platforms," 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, 2016, pp. 79-84. doi: 10.1109/ISI.2016.7745447

[7]. Birla, K. Kohli and A. Dutta, "Machine Learning on imbalanced data in Credit Risk," 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2016, pp. 1-6. doi: 10.1109/IEMCON.2016.7746326

[8]. G. Sudhamathy and C. J. Venkateswaran, "Analytics using R for predicting credit defaulters," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016, pp. 66-71. doi: 10.1109/ICACA.2016.7887925.

[9]. Vinod Kumar L, Natarajan S, Keerthana S, Chinmayi K M and Lakshmi N, "Credit Risk Analysis in Peer-to-Peer Lending System," 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA), Singapore, 2016, pp. 193-196. doi: 10.1109/ICKEA.2016.7803017

[10]. Lei Xia and Jun-feng Li, "Analysis on Credit Risk Assessment of P2P" *2016 E. Qi et al. (eds.), Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management 2015*, DOI 10.2991/978-94-6239-180-2_86

[11]. Fahmida E. Moula, Chi Guotai, Mohammad Zoynul Abedin, "Credit default prediction modeling: an application of support vector machine", DOI 10.1057/s41283-017-0016-x

[12]. Chandola, V., Banerjee, A., and Kumar, V. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI = 10.1145/1541880.1541882

[13]. http://www.investopedia.com/terms/d/defaultrisk.asp .

[14]. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concept and Techniques".

[15]. Sam Maes, Karl Tulys, Bram Vanschoenwinkel, Bernard Manderick, "credit card Fraud Detection. Applying Bayesian and Neural Network", 2016.

[16]. Ravinder Reddy, B.kavya, Y Ramadevi (Ph.D.), "A Survey on SVM Classifier for Intrusion Detection", 2014.