

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Credit Card Fraud Detection using Logistic Regression and Bayesian Network

Reetu Bala¹, Deepika Garg²

P.G. Student, Department of CSE, AITM at Palwal, Haryana, India¹

HOD, Department of CSE, AITM at Palwal, Haryana, India²

ABSTRACT: The primary target of this plan is to perform prescient investigation on credit cards fraud detection dataset utilizing AI procedures and distinguish the deceitful exchanges from the given dataset. The center is to distinguish if an exchange goes under ordinary class or fake class utilizing prescient models. Diverse examining methods will be executed to handle the class irregularity/deceitful exchange issue and arrangement of AI calculations like Logistic Regression and Bayesian Network will be actualized on the dataset, and the outcomes will be accounted for with estimating the information credited utilizing proposed calculations with more exactness and adequacy.

KEYWORDS: Artificial Intelligence, Machine Learning, Logistic Regression, Bayesian Network, Credit Card Fraud Detection.

I. INTRODUCTION

The credit card fraud misrepresentation/recognizable proof strategies are gathered into two general classes: extortion examinations (misuse acknowledgment) and client conduct examination (irregularity or oddity discovery). The essential social affair of frameworks oversees managed arrangement task in exchange level. In these methodologies, trades are set apart as phony or conventional reliant on past bona fide data. This dataset is then used to make game plan models which can foresee the state (regular or deception) of new records. There are different model creation methodologies for an ordinary two class gathering undertakings, for example, rule enlistment, choice trees, and neural frameworks. This methodology is exhibited to constantly recognize most deception or charge card misrepresentation traps which have been seen previously, it generally called extortion examination. The subsequent strategy oversees unsupervised ways of thinking which rely upon record lead. In this system, a trade is perceived beguiling in case it is intriguing with the user's customary direct. This is in light of the fact that we don't expect fraudsters act proportionate to the record owner or think about the led model of the owner. To this point, we need to remove the genuine customer social model (e.. customer profile) for each record and thereafter recognize false activities according to it. Contrasting new practices and this model, adequately particular activities are perceived as cheats. The profiles may contain the activity information of the record, for instance, seller types, total, region and time of trades. This system is generally called abnormality identification. It is basic to include the key complexities between customer direct examination and blackmail examination approaches. The extortion identification strategy can perceive acknowledged deception traps, with a low false positive rate. These systems remove the signature and model of blackmail traps showed in prophet dataset and can then adequately choose unequivocally which fakes, the structure is at present experiencing. If the test data does not contain any extortion denotes, no alert is raised. Along these lines, the counterfeit positive rate can be diminished incredibly. Be that as it may, since learning of a deception examination structure (for instance classifier) relies upon confined and unequivocal distortion records, It can't distinguish novel fakes. As needs be, the bogus negative rate might be high depending upon how keen are the fraudsters. Client conduct examination, on the other hand, gigantically addresses the issue of perceiving novel fakes. These strategies don't check for unequivocal blackmail plans however rather, take a gander at moving toward exercises with the assembled model of real customer direct. Any development that is adequately interesting in connection to the model will be considered as a possible distortion. Nonetheless, customer direct examination methodologies are earth shattering in identifying creative fakes, they truly

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

experience the ill effects of high rates of a bogus alarm. Furthermore, if a distortion occurs in the midst of the readiness arrange, this phony direct will be entered in standard mode and is believed to be regular in the further investigation. In this section we will quickly present some present coercion ID techniques which are associated with charge card blackmail area errands, in like manner, essential good position will be evaluated using Logistic Regression and Bayesian Network.

II. RELATED WORK

Aleskerov, Freisleben, and Rao [1] developed a fraud detection system called Cardwatch that is built upon the neural network learning algorithm. This system is aimed towards commercial implementation and therefore can handle large datasets, and parameters of an analysis can be easily adjusted within a graphical user interface. Cardwatch uses three main neural network learning techniques: conjugate gradient, backpropagation, and batch backpropagation. This system is a useful product for large financial institutions due to its ease of implementation with commercial databases. Unfortunately, the disadvantage of this system is the need to build a separate neural network for each customer. This results in a very large overall network that requires relatively higher amounts of resources to maintain.

Dorronsoro, and others [2] developed a neural network based fraud detection system called Minerva. This system's main focus is to imbed itself deep in credit card transaction servers to detect fraud in real-time. It uses a novel nonlinear discriminant analysis technique that combines the multi-layer perceptron architecture of a neural network with Fisher's discriminant analysis method. Minerva does not require a large set of historical data because it acts solely on immediate previous history, and is able to classify a transaction in 60ms. The disadvantage of this system is the difficulty in determining a meaningful set of detection variables and the difficulty in obtaining effective datasets to train with.

Kokkinaki [3] suggested to create a user profile for each credit card account and to test incoming transactions against the corresponding user's profile. The attributes that were used to construct these profiles are: credit card numbers, transaction dates, type of business, place, amount spent, credit limit and expiration time. Kokkinaki proposed a Similarity Tree algorithm, a variation of Decision Trees, to capture a user's habits. The analyses found that the method has a very small probability for false negative errors. However, in this approach the user profiles are not dynamically adaptive and therefore continual updates are needed when user habits and fraud patterns change.

Chan and Stolfo [4] studied the class distribution of a training set and its effects on the performance of multi-classifiers on the credit card fraud domain. It was found that increasing the number of minority instances in the training process results in fewer losses due to fraudulent transactions. Furthermore, the fraud distribution for training was varied from 10% to 90% and it was found that maximum savings were achieved when the fraud percentage used in training was 50%.

Brause and others [5] looked specifically at credit card payment fraud and identified fraud cases by combining a rule-based classification approach with a neural network algorithm. In this approach the rule-base classifier first checked to see if a transaction was fraudulent, and then the transaction classification was verified by a neural network. This technique increases the probability for the diagnosis of "fraud" to be correct and therefore it is able to decrease the number of false alarms while increasing the confidence level.

Wheeler and Aitken [6] developed a case-based reasoning system that consists of two parts, a retrieval component and a decision component, to reduce the number of fraud investigations in the credit approval process. The retrieval component uses a weighting matrix and nearest neighbor strategy to identify and extract appropriate cases to be used in the final diagnosis for fraud, while the decision component utilizes a multi-algorithm strategy to analyze the retrieved cases and attempts to reach a final diagnosis. The nearest-neighbour and Bayesian algorithms were used in the multi-algorithm strategy. Initial results of 80% non-fraud and 52% fraud recognition from Wheeler and Aitken suggest that their multi-algorithmic case-based reasoning system is capable of high accuracy rates.

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

III. PROPOSED WORK

The fraud detection flow starts with data collection and migrating the data to data repository first, thereafter the noise will be removed using data pre-processing techniques, hereinafter the feature selection module will define the attributes and classes responsible to define data sets into training sets and segregate the complexities. Subsequently, the rule engine is responsible to define the behaviour pattern to be evaluated and define goals to achieve whereas the Logistic Regression will evaluate the risk model based on rules defined using MLE (Maximum Likelihood) and pass the relevant information to Bayesian network for classification and to detect the fraud level and performance evaluation to the below mention workflow depicts the procedural dimensions to ensure that credit card fraud should be trapped ineffective and inappropriate time below figure depicts the proposed workflow:-

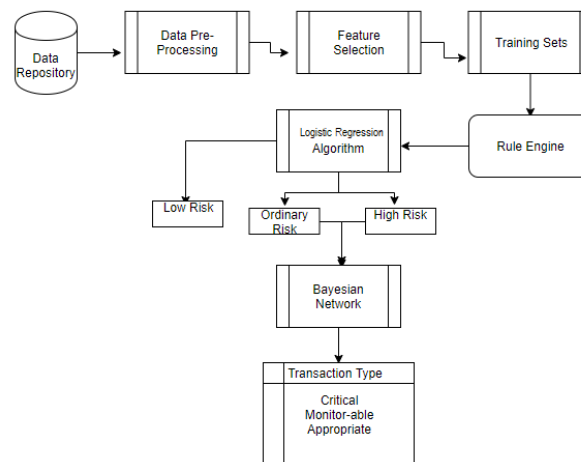


Figure 1: Work Flow model of Credit Card Fraud Detection System

IV. PROPOSED ALGORITHM

Logistic Regression : Because logistic regression predicts probabilities, rather than just classes, we can fit it using likelihood. For each training data-point, we have a vector of features, x_i , and an observed class, y_i . The probability of that class was either p , if $y = 1$, or $1 - p$, if $y = 0$. The likelihood is then :- $\beta = (X^T * X)^{-1} * X^T * Y$ we will evaluate

$\beta = (X^T * W * X)^{-1} * X^T * W * U$ where $U_i = X_i^T * \beta + \frac{y_i - \mu_i}{W_{ii}}$ and μ_i is our estimate for p , which we previously saw

$$W_{ii} = \mu_i * (1 - \mu_i)$$

could be written as $\mu_i = \frac{1}{1 + e^{-X_i^T * \beta}}$. The weights W_{ii} are nothing but the standard deviation of our own prediction. In

general, if $X \sim \text{Bin}(p, n)$ then $\text{Var}(X) = n * p * (1 - p)$ and since we have a Bernoulli trial we have $n = 1$ so the variance becomes $W_{ii} = \mu_i * (1 - \mu_i)$.

Bayesian Pseudo Code : Note that what we have just examined is very limited since we have only considered when each piece of evidence affects only one hypothesis.

- This must be generalized to deal with “m” hypotheses H_1, H_2, \dots, H_m and “n” pieces of evidence E_1, \dots, E_n , the situation normally encountered in real-world problems. When these factors are included, Equation (2) becomes

$$(1) P(H_i | E_{j1} \cap E_{j2} \cap \dots \cap E_{jk}) = \frac{P(E_{j1} \cap E_{j2} \cap \dots \cap E_{jk} | H_i) P(H_i)}{P(E_{j1} \cap E_{j2} \cap \dots \cap E_{jk})}$$

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

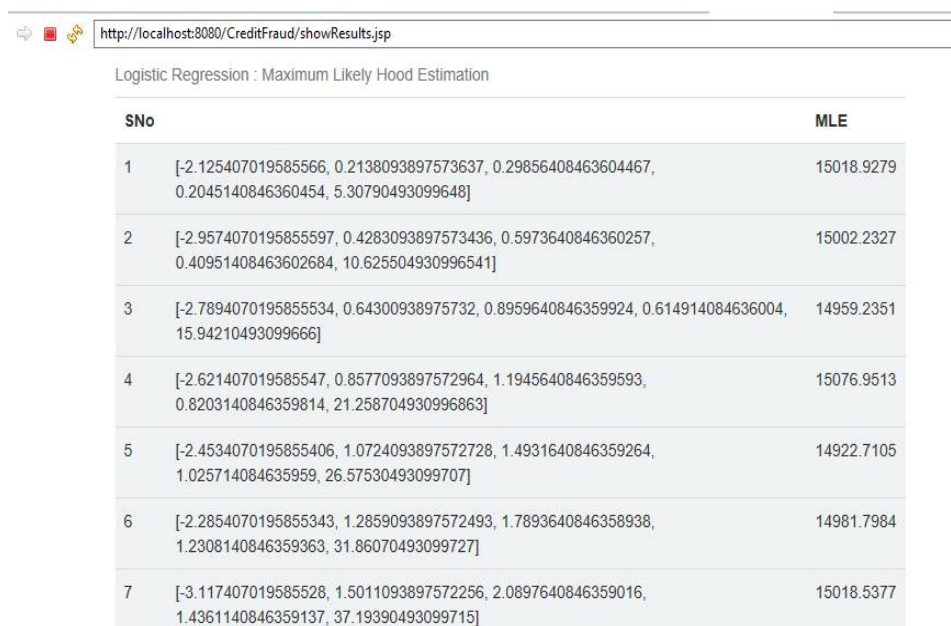
$$(2) P(H_i | E_{j_1} \cap E_{j_2} \cap \dots \cap E_{j_k}) = \frac{P(E_{j_1} \cap E_{j_2} \cap \dots \cap E_{j_k} | H_i) P(H_i)}{\sum P(E_{j_1} | H_1) P(E_{j_2} | H_1) \dots P(E_{j_k} | H_1) P(H_1)}$$

The hypotheses H_1, \dots, H_m , $m \geq 1$, are mutually exclusive. Furthermore, the hypotheses H_1, \dots, H_m are collectively exhaustive.

The pieces of evidence E_1, \dots, E_n , $n \geq 1$, are conditionally independent given any hypothesis H_i , $1 \leq i \leq m$. *Conditional independent*: The events E_1, E_2, \dots, E_n , are *conditionally independent* given an event H if $P(E_{j_1} \cap \dots \cap E_{j_k} | H) = P(E_{j_1} | H) \dots P(E_{j_k} | H)$ for each subset $\{j_1, \dots, j_k\} \subseteq \{1, \dots, n\}$.

V. SIMULATION RESULT

Logistic regression is one of the most popular machine learning algorithms that is used for classification. Although the term 'regression' appears in the name, it is not a regression algorithm. Logistic regression has its name as it was built on another very popular machine learning algorithm, which is used for regression problem. In logistic regression, the prediction is expressed in terms of probability of outcome belonging to each class. In a linear regression model, the real-valued outputs are predicted by combining the input variables (x) with the weights. To be more clear, consider there is just one input or independent variable ' x ' and a dependent variable ' y '. Then the hypothesis of linear regression can be expressed as $y = a_0 + a_1 * x$ where a_0 is a bias term, and a_1 is the weight for single input variable x . These weights are learned during the training. In this case, the value of the hypothesis can be less than 0 or greater than 1. Logistic regression also uses such a linear equation. However, since it should predict the probability of the outcome of belonging to each class, it uses a sigmoid function or a logistic function which is shown in equation 2.2, to squash the predicted real values between the range of 0 and 1 $\text{sigma}(z) = 1 / 1 + e^{-z}$. Figure below shows how the sigmoid function provides the result as algorithm proposed:-



SNo	MLE
1	[-2.125407019585566, 0.2138093897573637, 0.29856408463604467, 0.2045140846360454, 5.30790493099648]
2	[-2.9574070195855597, 0.4283093897573436, 0.5973640846360257, 0.40951408463602684, 10.625504930996541]
3	[-2.7894070195855534, 0.64300938975732, 0.8959640846359924, 0.614914084636004, 15.94210493099666]
4	[-2.6214070195855547, 0.8577093897572964, 1.1945640846359593, 0.8203140846359814, 21.258704930996863]
5	[-2.4534070195855406, 1.0724093897572728, 1.4931640846359264, 1.025714084635959, 26.57530493099707]
6	[-2.2854070195855343, 1.2859093897572493, 1.7893640846358938, 1.2308140846359363, 31.86070493099727]
7	[-3.117407019585528, 1.5011093897572256, 2.0897640846359016, 1.4361140846359137, 37.19390493099715]

Figure 2: Confusion Matrix using Logistic Regression with MLE

The parameters of the network are the local probability distributions attached to each variable. The structure and parameters taken together encode the joint probability of the variables. Let $U = \{X_1, \dots, X_n\}$ represent a finite set of discrete random variables. The set of parents of X_i are given by $pa(X_i)$. The joint distribution represented by a

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Bayesian network over the set of variables U is $p(U) = \prod_{i=1}^n p(X_i | pa(X_i))$ where n is the quantity of factors in U and $p(X_i | pa(X_i)) = p(X_i)$ when X_i has no guardians. The joint circulation for the arrangement of factors $U = \{C, S1, S2, S3\}$ from Figure 0 is determined as $p(U) = p(C)p(S1|C)p(S2|C)p(S3|C)$ the results are as follows:-

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_A
1	20000.0000	2	2	1	24	2	2	-1	-1	-2	-2	3913.0000	3102.0000	689.0000	
2	120000.0000	2	2	2	26	-1	2	0	0	0	2	2682.0000	1725.0000	2682.00	
3	90000.0000	2	2	2	34	0	0	0	0	0	0	29239.0000	14027.0000	13559.0	
4	50000.0000	2	2	1	37	0	0	0	0	0	0	46990.0000	48233.0000	49291.0	
5	50000.0000	1	2	1	57	-1	0	-1	0	0	0	8617.0000	5670.0000	35835.0	
6	50000.0000	1	1	2	37	0	0	0	0	0	0	64400.0000	57069.0000	57608.0	
7	500000.0000	1	1	2	29	0	0	0	0	0	0	367965.0000	412023.0000	445007.0	
8	100000.0000	2	2	2	23	0	-1	-1	0	0	-1	11876.0000	380.0000	601.0000	
9	140000.0000	2	3	1	28	0	0	2	0	0	0	11285.0000	14096.0000	12108.0	
10	20000.0000	1	3	2	35	-2	-2	-2	-2	-1	-1	0.0000	0.0000	0.0000	
11	200000.0000	2	3	2	34	0	0	2	0	0	-1	11073.0000	9787.0000	5535.00	
12	260000.0000	2	1	2	51	-1	-1	-1	-1	-1	2	12261.0000	21670.0000	9966.00	
13	630000.0000	2	2	2	41	-1	0	-1	-1	-1	-1	12137.0000	6500.0000	6500.00	
14	70000.0000	1	2	2	30	1	2	2	0	0	2	65802.0000	67369.0000	65201.0	
15	250000.0000	1	1	2	29	0	0	0	0	0	0	70887.0000	67060.0000	63561.0	
16	50000.0000	2	3	3	23	1	2	0	0	0	0	50614.0000	29173.0000	28116.0	
17	20000.0000	1	1	2	24	0	0	2	2	2	2	15378.0000	18010.0000	17428.0	
18	320000.0000	1	1	1	49	0	0	0	-1	-1	-1	253286.0000	246536.0000	194663.0	
19	360000.0000	2	1	1	49	1	-2	-2	-2	-2	-2	0.0000	0.0000	0.0000	
20	180000.0000	2	1	2	29	1	-2	-2	-2	-2	-2	0.0000	0.0000	0.0000	
21	130000.0000	2	3	2	39	0	0	0	0	0	-1	38358.0000	27688.0000	24489.0	
22	120000.0000	2	2	1	39	-1	-1	-1	-1	-1	-1	316.0000	316.0000	316.0000	
23	70000.0000	2	2	2	26	2	0	0	2	2	2	41087.0000	42445.0000	45020.0	
24	450000.0000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512.0000	19420.0000	1473.00	
25	90000.0000	1	1	2	23	0	0	0	-1	0	0	4744.0000	7070.0000	0.0000	

Figure 3 Results using Bayesian Network and Logistic Regression : Red Line Depicts the Fraud Transactions Occurred in E-Commerce Business using Credit Cards

VI. CONCLUSION

In this proposed scheme, we applied machine learning techniques to predict whether a credit card transaction is fraudulent or not. For this, we collected a publicly available dataset provided by the machine learning group of Machine Learning Repository from UCI, which contains the record of credit card transactions made by Taiwanese cardholders and occurred in 6 Months in September 2005 to backward month wise respectively. It contains 30000 transactions out of which only 3182 approximately are fraudulent. The dataset is highly unbalanced as the positive class accounts for only 10.67% of the total transactions. When providing input data of a highly unbalanced class distribution to the predictive model, the model tends to be biased towards the majority samples. As a result, it tends to misrepresent a fraudulent transaction as a genuine transaction. To tackle this problem, we implemented a data level approach which includes regression technique namely, Logistic Regression which originally formed the confusion matrix and graded or illustrated the MLE (Maximum Likelihood Estimation). In addition, we implemented the algorithmic model based on Bayesian Network for classification by which we can find out which transaction is fraudulent with evaluating the positive/negative binary weights and relation formed therein based on attributed and their respective features. For this, For future work, a cost-sensitive learning approach can be implemented by considering the misclassification costs. The cost for misclassifying a fraudulent class as a legitimate class (False Negative), which corresponds to the fraud amount (can be from few to thousands of dollars) is much higher than the cost for misclassifying a legitimate class as a fraudulent class (False Positive), which corresponds to the cost related to analysing the transaction and contacting the cardholder. So, this type of learning deals with classifying an example into a class that has the minimum expected cost. Credit card fraud is related to the non-stationary nature of transaction distributions in which the fraudsters usually always comes with a new way to attempt the fraudulent activities. Therefore, it becomes essential to consider these changing behavior as well while developing a predictive model. Hence, a detailed study on dealing with non-stationary

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

nature in credit card fraud detection can be performed. However, this study requires a huge amount of data which can be processed using streaming services like Apache Spark in future scenarios.

REFERENCES

- [1] Aleskerov, Emin & Freisleben, Bernd & Rao, Bharat. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFER). 220 - 226. 10.1109/CIFER.1997.618940.
- [2] J. R. Dorronsoro, F. Ginel, C. Sgnchez and C. S. Cruz, "Neural fraud detection in credit card operations," in IEEE Transactions on Neural Networks, vol. 8, no. 4, pp. 827-834, July 1997.
- [3] Kokkinaki, A. 1997. On Atypical Database Transactions: Identification of Probable Frauds using Machine Learning for User Profiling, Proc. of IEEE Knowledge and Data Engineering Exchange Workshop; 107-113.
- [4] P. K. Chan, W. Fan, A. L. Prodromidis and S. J. Stolfo, "Distributed data mining in credit card fraud detection," in IEEE Intelligent Systems and their Applications, vol. 14, no. 6, pp. 67-74, Nov.-Dec. 1999.
- [5] R. Brause, T. Langsdorf, M. Hepp, Neural Data Mining for Credit Card Fraud Detection, November 1999 ICTAI '99: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence.
- [6] Wheeler, Richard & Aitken, Stuart. (2000). Multiple Algorithms for Fraud Detection. Knowledge-Based Systems. 13. 93-99. 10.1016/S0950-7051(00)00050-2.
- [7] Abdelhalim, Amany & Traore, Issa. (2009). Converting Declarative Rules into Decision Trees. Lecture Notes in Engineering and Computer Science. 2178.
- [8] Ngai, E.W.T. & Hu, Yong & Wong, Y.H. & Chen, Yijun & Sun, Xin. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems. 50. 559-569. 10.1016/j.dss.2010.08.006.
- [9] F. Ogwueleka, "Data mining application in credit card fraud detection system," Journal of Engineering Science and Technology, Vol. 6, No. 3, 2011.
- [10] R. Patidar and L. Sharma, "Credit Card Fraud Detection Using Neural Network," International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-1, Issue NCAI2011, June 2011.
- [11] J. Dara and L. Gundemoni, "Credit Card Security and E-Payment," 2006 [15] P. Chan, W. Fan, Prodromidis and Salvatore, "Distributed Data Mining in Credit Card Fraud Detection," IEEE December 1999.
- [12] John O. Awoyemi, Adebayo Olusola Adetunmbi, and Samuel Adebayo Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNi), pages 1–9, 2017.
- [13] Emin Aleskerov, Bernd Freisleben, and R. Bharat Rao. Cardwatch: a neural network based database mining system for credit card fraud detection. In CIFER, 1997. Kevin W. Bowyer, 19. Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. J. Artif. Intell. Res., 16:321–357, 2002
- [14] Galina Baader and Helmut Krcmar. Reducing false positives in fraud detection: Combining the red flag approach with process mining. International Journal of Accounting Information Systems, 2018.
- [15] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations, 6:20–29, 2004.
- [16] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition, 30:1145–1159, 1997. [CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In KDD, 2016
- [17] Fabrizio Carcillo, Andrea Dal Pozzolo, Yann-A'el Le Borgne, Olivier Caelen, Yannis Mazzer, and Gianluca Bontempi. Scarff: a scalable framework for streaming credit card fraud detection with spark. Information Fusion, 41:182–194, 2018.
- [18] Thomas G. Dietterich. Multiple classifier systems. In Lecture Notes in Computer Science, 2000.
- [19] Pedro M. Domingos. Metacost: A general method for making classifiers cost-sensitive. In KDD, 1999.