International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Multi-Level Credit Card Fraud Detection

V. Sobanadevi, G. Ravi

Abstract: ***Fraud detection in credit card transactions is one of the major requirements of the current business scenario due to the huge amount of losses associated with the domain. This work presents a multi-level model that can provide highly effective fraud detection in credit card transactions. The model is based on the amount for which the transaction is committed. The proposed MLFD model identifies the nature of the transaction and depending on the significance level of the transaction, the appropriate learning model is selected. Experiments were performed with the standard benchmark data and comparisons were performed with existing model in literature. Results shows that the proposed model exhibits high performance compared to the existing model.***

Keywords: Credit card fraud detection, Decision Tree, Multi-Level Modelling , Naïve Bayes, Random Forest.

I. INTRODUCTION

The process of fraud detection in credit card transactions is used to determine if a transaction is fraudulent or legitimate. The detection process is governed by the historical transactions. The past labelled transaction data is used to train the model for predicting the current transactions. This is a challenging task mainly due to the change in the spending behaviors of the users. Spending behaviors of users often tend to vary with time, which is considered as seasonal changes and also gradually varies over long time periods, which can be treated as trends in the behaviors. This results in the historical models becoming obsolete due to the absence of recent data. Further, fraudsters are also creating evolving models to compete with the recent advancements in the detection strategies.

Recent research shows that credit card fraud losses have resulted in significant losses for organizations and also for banks. It was observed that the losses due to credit card increased by 300% in five years (2010 to 2015) [1]. By the year 2020, the global loss due to fraud is expected to touch $31.67 billion. The two major sources of card frauds were identified as counterfeit card based frauds and frauds at point-of-sale regions. Both these span globally and are not confined to bank premises. Hence they inadvertently require effective prediction mechanisms to stop the fraud from happening at those locations itself. This calls for a model that operates in real-time and also accurately. However, when it was viewed from a business perspective, not all predictions are to be considered significant. Every prediction involves a certain cost. Identifying a transaction as fraudulent requires business models to raise fraud alarms to customers, while falsely flagging a fraudulent transaction as legitimate incurs losses. It would be more appropriate to concentrate on higher Valued transactions, which can help balance the cost factors. This work presents a multi-level fraud detection model that aims to consider the amount involved in the transaction to determine the type of prediction model to be used for prediction. The amount directly corresponds to significance of the transaction. Transactions with low significance are predicted using fast predictors, while transactions with higher significance are predicted with better prediction models. This enables the models to provide cost sensitive and faster prediction. This remainder of this work is presented as follows: Section II presents a review of literature, Section III presents a detailed description of the working of the proposed model, Section IV presents the results and discussion and Section V concludes the work.

II. LITERATURE REVIEW

Credit card fraud detection is considered to be one of the major domains of analysis in the research world. This is due to the huge amount of losses associated with the domain. This section presents some of the significant works in the fraud detection domain. Risk based modelling involving cost has become a major factor in the current fraud detection models. A risk based model was proposed by Bahnsen et al. [2]. This method is based on Naïve Bayes to identify the risk levels to be used for the final prediction. The work by Pozollo et al. [3] aims to provide testing mechanisms for models handling concept drift. A cost based model was proposed by Mahmoudi et al. [4]. This technique considers cost as a major factor for the prediction process involving credit card transactions. The method is based on Fisher Discriminant Analysis. A Game Theory based analysis for credit card fraud detection was proposed by Gianini et al. [5]. A Big Data based credit card fraud detection model that utilizes Big Data based techniques was proposed by Vaughan [6]. A customer incentive based model was proposed by Wang et al. [7]. Feature engineering is one of the mostly incorporated methods into the credit card fraud detection strategy. The behavioral change in users has made fraud detection in credit card transactions a complex process. In order to leverage the best rules, it becomes mandatory to mine as much information as possible from the transaction data. Feature engineering strategies perform the exam same process.

Manuscript published on November 30, 2019.
* Correspondence Author
V. Sobanadevi*, Research Scholar, Department of Computer Science, Jamal Mohamed College, Trichy, Tamilnadu, India.
G. Ravi, Associate Professor and Head, Department of Computer Science, Jamal Mohamed College, Trichy, Tamilnadu, India.

A feature engineering based model for detecting frauds in credit card transactions was proposed by Lucas et al. [8]. This work aims at using HMM for the prediction process. Multiple perspectives are obtained by identifying several features that constitute behavioral nature of the customer. The method proposed by Minegishi et al. [9] selects and handles only the raw features of the data for the fraud detection process. The necessity of using feature engineering methodologies and additional enhanced attributes was presented by Bolton and Hand in their work [10]. Other significant contributions include models by Zhang et al. [11].

Ensemble based methods are the type of models that use multiple models for analysis. The increase in complexity of base data has resulted in the requirement of complex models. This is satisfied by ensemble modelling techniques. A framework of ensemble based modelling technique was proposed by Kim et al. [12]. The major advantage of this model is its usage of realistic metrics and the employment of real-world constraints. Other similar ensemble based modelling techniques include models by Zareapoor et al. [13] and Vlasselaer et al. [14]. A hybrid model that uses a combination of supervised and unsupervised models was proposed by Carcillo et al. [15]. Other similar hybrid techniques include models by Yamanishi et al. [16] and Veeramachaneni et al. [17].

III. MULTI LEVEL FRAUD DETECTION (MLFD)

Fraud detection in credit card transactions is a complex task that requires balancing the tradeoff between false alerts and missed frauds. This work presents a Multi-Level Fraud Detection (MLFD) model that aims to provide an effective balance and also enhanced predictions. The MLFD architecture is shown in figure.

A. Data Preprocessing

Data preprocessing is the initial phase of the proposed MLFD model. Data obtained from credit card transactions usually constitute several features that help determine the nature of the transaction. The data should hence be preprocessed to eliminate the labelling contents. Further, several categorical values are also contained in the dataset. Such data should be converted to numerical formats prior to machine learning. Hence one-hot encoding is applied to the data. This results in a huge expansion in the dataset. The end of preprocessing phase prepares the data for te machine learning phase.

B. Data Segregation and Model Training

Data segregation is the next phase in the architecture. The input data is split into training and test data. The split is performed in 7:3 ratio basis. 70% of the input data is used for training and 30% of the data is used for testing purposes.

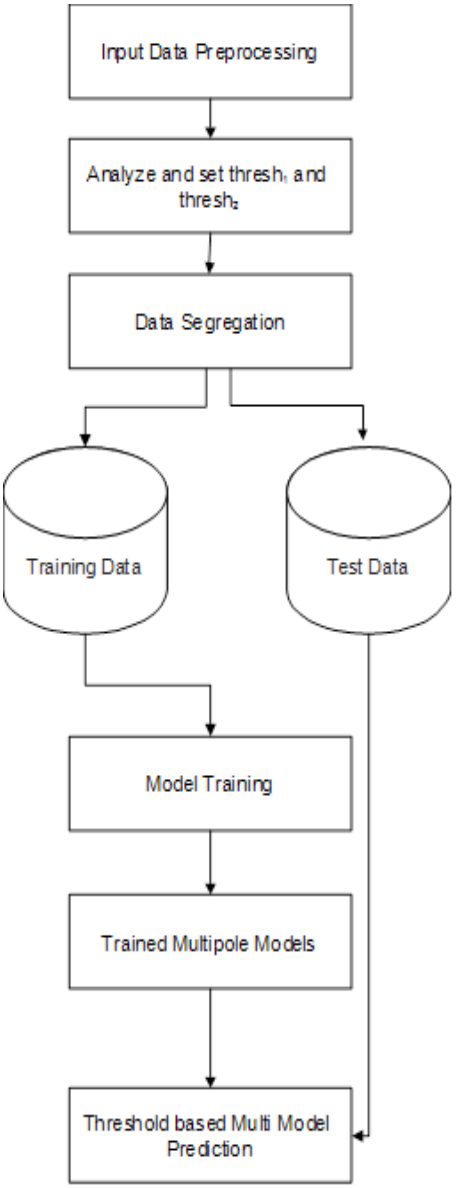

Fig. 1. MLFD Architecture

The training data is passed to the machine learning models for training. This work uses three base models for the prediction process. A fast predictor to suit the real-time requirements, a predictor model that is highly effective in predicting fraudulent transactions (best fraud predictor) and a best overall predictor. The classifier models used in this work includes Naïve Bayes (fast predictor) , Decision Tree (best fraud predictor) and Random Forest (best overall predictor). The models used in this work have been identified by analysis on the training data. Each model is used in a different prediction scenario and for a different purpose. Their usage is dependent on the amount involved in the transaction. Naïve Bayes is a probability-based model. The major advantage of this model is that it considers the property of attribute independence. This property regulates that every attribute or feature contained in the data is only related to the class attribute and no other.

This ensures that the influence of an attribute over other attributes is very minimal, hence the prediction process based on an attribute is determined only by it and no other. Credit card transaction data, being very sensitive in nature, is greatly benefitted by this approach. Further, the prediction time is also very low resulting in the model operating up to the real-time expectations.

Decision Tree is a tree based model that operates by building decision rules. The rules are identified by the entropy levels exhibited by each attribute. At every level, attributes with highest entropy are used to create the divisions. Each division in the tree creates a different decision flow. Every intermediate node, starting from the root node contains conditions. The final leaf nodes correspond to predictions. The prediction process follows these rules starting from the root node and finally results in the predictions. The Decision Tree follows a slightly complicated process compared to Naïve Bayes. They are effective in classifying data with dependent attributes. The prediction times required by decision trees are also low. Hence they also provide real-time predictions.

Random Forest is an ensemble based modelling technique. Random forest creates multiple training data bags and multiple instances of the prediction model to create a complex prediction model. This method enables better predictions, however, the model is computationally complex. The predictions are also time consuming. However, when used appropriately, the model provides a huge boost in the predictions.

The training data is passed to each of these prediction models and the models are trained. The prediction efficiency of each of the models in terms of predicting legitimate transactions and in terms of fraudulent transactions. The best predictors of each of these categories is used for the prediction process appropriately.

C. Multi-Level Prediction

The prediction process is multi-level and prediction is performed based on the amount involved in the transaction. Transaction amount plays a vital role in determining the significance of a transaction. The existing models considers transactions as single and equally significant entities. This however is not the case in organizational environments. Every transaction is provided a different importance depending on the customer and the amount for which the transaction is performed. The general classification architecture should be modified to incorporate this ideology into the prediction process. The prediction process in this work hence follows a workflow that determines the significance of transactions and then determines the type of prediction model required. Algorithm for the multi level prediction process is provided below.

Algorithm

1. Set threshold1 and threshold2 based on the training data
2. For each transaction t in the test data
a. If amount involved > thresh1
i. Perform prediction using the best overall prediction model
ii. If the transaction is predicted as fraudulent, raise the alarm for secondary verification
iii. If the transaction is predicted as legitimate
1. If the amount involved > thresh2
a. Predict using the best fraud predictor
b. Raise fraud alarm if prediction is fraudulent else consider the transaction to be legitimate
2. If amount involved < thresh2 consider the transaction to be legitimate
b. If amount involved < thresh1 predict using the fast predictor
c. Raise fraud alarm if prediction is fraudulent else consider the transaction to be legitimate

The amount threshold values thresh1 and thresh2 are set initially. These are determined by analyzing the training data and is decided by the domain expert. For each test data provided for prediction, the amount involved in the transaction is identified. If the amount is less than thresh1 the fast predictor is used for prediction. If the amount involved is greater than thresh1, the best overall predictor is used for prediction. If the prediction labels the transaction as legitimate and if the amount involved is greater than thresh2, the transaction is again predicted using the best fraud predictor. If the transaction amount involved is less than thresh1, the transaction is predicted using the fast predictor. The obtained prediction is verified and if it is fraudulent, a fraud alarm is raised, and secondary verifications are initiated. If the transaction is predicted as legitimate, the transaction is allowed to happen.

The additional rules and the model selection process enables the model to provide faster predictions when appropriate and also secure transactions for the customers.

IV. RESULTS AND DISCUSSION

The proposed MLFD model has been implemented using Python. The experiments were performed using BankSim dataset, which is a simulated and benchmark dataset for fraud detection models. The results obtained are compared with work by Vaughan et al. [6]. The results are examined in terms of their true prediction levels, false prediction levels and ROC plots.

The true prediction levels of the proposed model is shown in figure 2. True prediction levels correspond to True Positive Rate (TPR) and True Negative Rate (TNR). A good prediction model requires both these metrics to exhibit the highest possible values. It could be observed that the proposed MLFD model exhibits good TPR and TNR levels. Further, comparison also indicates that the proposed model exhibits better prediction levels compared to the model by Vaughan et al.

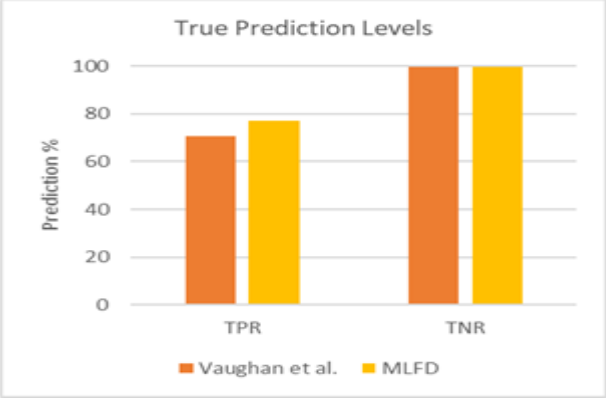

Fig. 2. Comparison of True Prediction Levels

A comparison of the false prediction levels is shown in figure 3. False prediction levels correspond to False Positive Rate (FPR) and False Negative Rate (FNR). FaPR correspond to false alarms and FNR correspond to a missed fraud. A good prediction model should exhibit low FPR and FNR levels. These metrics also correpond to the direct losses incurred due to fraud. It could be observed that the proposed MLFD model exhibits less than 1% false alarms and <30% of missed frauds. Comparisons also shows that the proposed model performs better than the model by Vaughan et al.

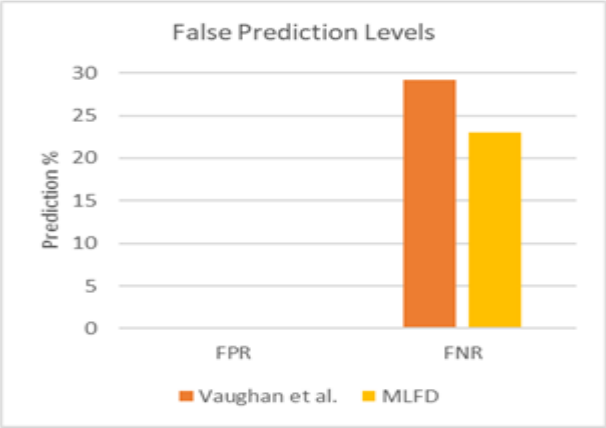

Fig. 3. Comparison of False Prediction Levels

ROC curve for the proposed model is shown in figure. ROC curve constitutes FPR levels in the x axis and TPR levels in the y axis. The model exhibiting the highest TPR levels and the lowest FPR levels is considered to be the best predictor. It could be observed that the proposed MLFD model exhibits higher TPR levels and lower FPR levels compared with the model by Vaughan et al.

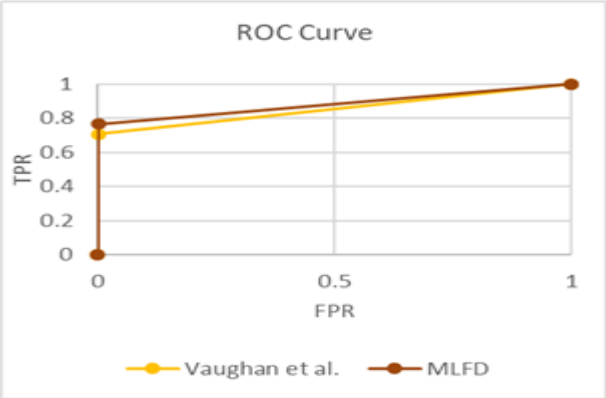

Fig. 4. Comparison of ROC curve

A tabulated view of the results is shown in table 1. A comparison of TPR, FPR, TNR and FNR are shown in the table. The best predictions are marked in bold. Both MLFD and Vaughan et al. models exhibits very low FPR and very high TNR. However, when it comes to TPR levels, which corresponds to the fraud predictions, the proposed model exhibits 7% higher prediction level compared to Vaughan et al. Similarly, on FNR levels, which corresponds to the highest loss, the proposed model exhibits better predictions. This exhibits the high efficiency of the proposed MLFD model.

Table- I: Performance Comparison with Vaughan et al.

	Vaughan et al.	MLFD
TPR	70.8	**77**
FPR	**0.2**	**0.2**
TNR	**99.8**	**99.8**
FNR	29.2	**23**

V.CONCLUSION

This work presents an effective model to handle frauds in credit card transactions. The major advantage of this model is the rule based prediction process. Although multiple models are included in the architecture, prediction is performed by the best model identified by the rules. Hence the architecture is found to be robust and efficient. The downside of the model is that it exhibits moderate fraud detection rates. Although these rates are better than existing models, there exists scope for improvement in this section. Further enhancements of the model are the possibilities to add additional rules that can identify fraud repetitions. Further, feature engineering methodologies can be included in the model to provide enhancements.

REFERENCES

1. Robertson D. The Nilson report. 2016. https://www.nilsonreport.com/upload/ content _ promo/The _ Nilson _ Report _ 10- 17- 2016.pdf .
2. Bahnsen, Alejandro Correa, Aleksandar Stojanovic, Djamila Aouada, and Björn Ottersten. "Cost sensitive credit card fraud detection using Bayes minimum risk." In 2013 12th international conference on machine learning and applications, vol. 1, pp. 333-338. IEEE, 2013.
3. Dal Pozzolo, Andrea, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. "Learned lessons in credit card fraud detection from a practitioner perspective." Expert systems with applications 41, no. 10 (2014): 4915-4928.
4. Mahmoudi, Nader, and Ekrem Duman. "Detecting credit card fraud by modified Fisher discriminant analysis." Expert Systems with Applications 42, no. 5 (2015): 2510-2516.
5. Gianini, Gabriele, Leopold Ghemmogne Fossi, Corrado Mio, Olivier Caelen, Lionel Brunie, and Ernesto Damiani. "Managing a pool of rules for credit card fraud detection by a Game Theory based approach." Future Generation Computer Systems 102 (2020): 549-561.
6. Vaughan, Gregory. "Efficient big data model selection with applications to fraud detection." International Journal of Forecasting (2018).
7. Wang, Deshen, Bintong Chen, and Jing Chen. "Credit card fraud detection strategies with consumer incentives." Omega 88 (2019): 179-195.

8. Lucas, Yvan, Pierre-Edouard Portier, Léa Laporte, Liyun He-Guelton, Olivier Caelen, Michael Granitzer, and Sylvie Calabretto. "Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs." Future Generation Computer Systems 102 (2020): 393-402.
9. T. Minegishi, A. Niimi, Proposal of credit card fraudulent use detection by online-type decision tree construction and verification of generality, Int. J.Inf. Secur. Res. 1 (4) (2011).
10. R. Bolton, D.J. Hand, Unsupervised profiling methods for fraud detection, in: Credit Scoring and Credit Control VII, 2001.
11. Zhang, Xinwei, Yaoci Han, Wei Xu, and Qili Wang. "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture." Information Sciences (2019).
12. Kim, Eunji, Jehyuk Lee, Hunsik Shin, Hoseong Yang, Sungzoon Cho, Seung-kwan Nam, Youngmi Song, Jeong-A. Yoon, and Jong-il Kim. "Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning." Expert Systems with Applications 128 (2019): 214-224.
13. Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia Computer Science, 48 , 679-685.
14. Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, 665 B. (2015). Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75 , 38-48.
15. Carcillo, Fabrizio, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. "Combining unsupervised and supervised learning in credit card fraud detection." Information Sciences (2019).
16. K. Yamanishi , J.I. Takeuchi , Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 389-394 .
17. K. Veeramachaneni , I. Arnaldo , V. Korrapati , C. Bassias , K. Li , AI 2 : training a big data machine to defend, in: Big Data Security on Cloud (BigDataSe- curity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), IEEE, 2016, pp. 49-54.

AUTHORS PROFILE

V. Sobanadevi is a research scholar working in the field of big data analytics. She has done her M.Phil at Jamal Mohamed college in the area for network security and has done her M.sc (IT) in SRM college Chennai. Her research area is based on Machine Intelligence based Zero Event Anomaly Detection in Big Data using Human Guided Interactive Visualization.

Dr. G. Ravi is working as associate professor and Head in Department of computer science, Jamal Mohamed College (Autonomous), Tiruchirapalli, Tamil Nadu, India. He has more than 31 years of experience in teaching field. His areas of interest include Artificial Intelligence, Network Management, Big Data Analytics and Cloud computing. His current area of research is wireless sensor Networks.