Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 8, Issue. 3, March 2019, pg.257 – 260

Credit Card Fraud Detection using Machine Learning Methodology

Hamzah Ali Shukur
Department Of Information Technologies
Faculty of Engineering and Natural Sciences, Altınbaş University-Turkey
Hamzamarsoomi@gmail.com

Sefer Kurnaz
Department of Computer Engineering, Altinbas University
Faculty of Engineering and Natural Sciences, Altınbaş University-Turkey
sefer.kurnaz@altinbas.edu.tr

Abstract— The speedy participation in online primarily based transactional activities raises the fallacious cases everywhere and causes tremendous losses to the personal and financial business. [1] Although several criminal activities are occurring in commercial business, fraudulent e-card activities are among the foremost prevailing and disturbed regarding by online customers. Data processing techniques were used to check the patterns and characteristics of suspicious and non-suspicious transactions supported normalized and anomalies knowledge. On the opposite hand, machine learning (ML) techniques were used to predict the suspicious and non-suspicious transactions mechanically by victimization classifiers [2][5]. This paper discusses the supervised based mostly classification. When preprocessing the dataset using normalization and Principal element Analysis, all the classifiers achieved over 95.0% accuracy compared to results reached before preprocessing the dataset.

Keywords— ML, Classification, Data processing, supervised, learning

I. INTRODUCTION

As businesses still move into the online community which currency is transacted dynamically in cash-less banking finance, adequate anomalies detection stay an important factor for bank systems. Not for the reason to stop the explicit cost obtained with counterfeit activities although verify that automated and manual reviews don't adversely wedge legitimate customers [3]. In deposit or withdraw trade, illegal transactions on card happens once someone abducts information from the card to undertake to purchases while no permission given from the holder and conjointly the detection of these dishonourable transactions has become a significant activity for payment processors.

A typical fraud detection systems encompass associate academic degree automatic tool and a manual technique. The automatic tool depends on fraud detection rules. It analyses all the new incoming transactions and assigns a fallacious score. Fraud investigators produce the manual technique [6]. They concentrate on transactions with a high fallacious score and turn back binary feedback (fraud or legal) on all analysed activity. The fraud detection systems are supported professionally driven rules, knowledge-driven rules or a combination of every style of rules [4].

The created rules try to verify specific things of fraud discovered by the fraud investigators. A state of affairs of fraud is "a cardholder can avoid dealing throughout a given country and, among the 2 next weeks, he can another dealing for a given amount in another given country. If this example is detected among transactions, then the anomaly detection system will manufacture an academic degree alert. Machine learning algorithms rules. They learn the fallacious patterns and check out to find them during a data-stream of new incoming transactions. The usually used machine learning algorithms embody supply regression, SVM Fraud detection may be a problematic machine learning for many reasons.

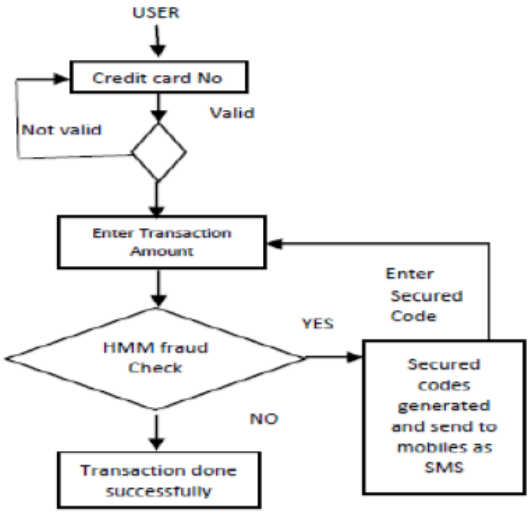
II. CONTRIBUTION

The task in this project is to classify the fraud activity and the normal activity as good as possible. The main challenge is the dataset is highly imbalanced. There are 492 frauds out of 284,807 transactions. It is difficult to lower the false negative rate (misclassify fraud activity as normal) while the false positive rate (misclassify normal activity as fraud) is still kept reasonably low. In the analysis of prediction on highly imbalanced dataset, recall score and auc are some good indicators. The indicator recall score is defined as true positive/(true positive + false negative). The indicator auc under ROC curve is the area under the ROC curve [Some people use area of another curve as auc]. Usually, high recall score and high auc under ROC can conclude as an accurate classification. My goal in this project is to build up a model with high recall score and acc.

III. PROBLEM DESCRIPTION

Our goal is to implement 3 different machine learning models in order to classify, to the highest possible degree of accuracy, credit card fraud from a dataset gathered in Europe in 2 days in September 2013 [7]. After initial data exploration, we knew we would implement a logistic regression model, a k-means clustering model, and a neural network. Some challenges we observed from the start were the huge imbalance in the dataset: frauds only account for 0.172% of fraud transactions. In this case, it is much worse to have false negatives than false positives in our predictions because false negatives mean that someone gets away with credit card fraud. [9]False positives, on the other hand, merely cause a complication and possible hassle when a cardholder must verify that they did, in fact, complete said transaction (and not a thief).

Figure 9: Proposed model of credit card fraud detection after training during detection.



Equations

There is a sever algorithm that we will use in our paper which we need to explain how exactly works the first one that we will start with:

- **Isolation Forest**

Outlier detection formula of an anomaly score is required for decision prediction. For Isolation Forest it is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where h(x) is the path length of observation x, c(n) is the Avg path length of failed search in a BST (Binary Search Tree) and (n) is the number of other nodes.

Each n observation is given an anomaly score and therefore the following call are often created on its basis:

- Score near to 1 precise the outlier
- Score less than 0.5 show legal transactions
- In condition of scores which they are near to 0.5 than the rest of sample does not seem clearly detect anomalies

- **The local outlier factor**

The LOF depends on a thought of limited density, which section is given by KNN, where the distance is used to predict the solidity or denseness [12]. By inspecting the real density of correlated object to the local densities of its neighbours, one can give areas of the approximately same density and points that have a significantly lower density than other neighbours. These are thought of to be exception [8] [6].

The standard distance calculates the local density at that defined point is often "reached" from its neighbours. The definition of "reachability distance" used in LOF is a further measure to provide additional stable results in clusters.

IV. DATA PROCESSING

One challenging aspect of the dataset was that there were 30 features, but in order to protect confidentiality, all but 28 of them had been PCA transformed and had unknown labels. The known, non-transformed features were 'Time', which measures the seconds between the transaction and the first transaction in the 2-day time period, and 'Amount', which is the cost of the transaction, presumably in Euros. In order to offset the imbalance in the dataset, we oversampled the fraud (class = 1) portion of the data, adding Gaussian noise to each row.

A. Modeling

The 3 models we used were a fully connected LOF, Isolation Forest, and logistic regression. In terms of the neural network, performing principal component analysis on the oversampled data before splitting it into training and test sets resulted in a jump from 50% accuracy to 94.56% accuracy For LOF the reached accuracy after PCA was 0.996% and for Isolation Forest 0.997%. Before PCA, nothing we tried was able to push the accuracy past 50%. After this step, however, adjustments to the number of layers, activation functions, and neurons in each layer did not do much to change the accuracy, which hovered at just below 95%. Furthermore, choosing only 2 neurons for the first dense layer (called a "bottleneck effect") forced the model to really reduce to only the most necessary features and decrease the likelihood of over fitting. For the K-means model, principal component analysis was used to reduce the dimensionality of the data from 31 to 2. Only the two features with the most variance were used to train the model. The model was set to have 2 clusters, 0 being non-fraud and 1 being fraud. We also experimented with different values for the hyperparameters, but they all produced similar results. Changing the 2 dimensionality of the data (reducing it to more dimensions than 2) also made little difference on the final values [10]. The last model we used was logistic regression, as it was a good candidate for binary classification. On each logistic regression model we trained, we made the constant C to be $1 * 10{-}5$. We trained three different configurations: a vanilla logistic regression with no preprocessing whatsoever, a logistic regression with oversampling on the scarce fraudulent data points and data scaling, and a logistic regression with balanced weights for each class (which greatly helped the data unbalance).

B. Comparison of Models

Logistic regression outperformed both the K-means and neural network. We believe that it is because of how the decision boundary changed with the class weights features. The neural network was next, and K-means performed the poorest. We believe this is due to the fact that clustering relies entirely on the similarities and differences of features of the dataset. Since fraud transactions can look very similar to regular transactions, it is difficult to put them into a separate group based on features alone.

For Local outlier factor as the result showed that the accuracy was much higher than the previous model but it's still a bit smaller than isolation forest this one showed a powerful result with 0.997% of accuracy. As result we can say that Isolation forest is the best model for this study case

ACKNOWLEDGEMENT

While the isolation forest had a high accuracy, its biggest pitfall was that within its 0.03% inaccuracy rate. The fact that this neural network missed 4.60% of frauds is enough to make this model infeasible compared to the other methods, where the false negative rate was much lower. Interestingly, a switch from a sigmoid to tanh activation function reduced the false negative rate by about 1%. The K-means clustering model produced a low accuracy of 54.27%. Of the wrongly predicted transactions, 99.75% were false positives, giving only 0.24% false negatives, or 0.11% of the validation set. However, the false negative rate was only so low due to the extremely low proportion of frauds in the dataset. In reality, 112 of the 176 frauds were misclassified as non-frauds, giving this a true accuracy rate of 36.36%. Therefore, K-means would not be the preferred model for this dataset, as it did not correctly predict frauds and it also produced a lot of false positives. The logistic regression gave us the best results. The vanilla logistic regression gave us a great accuracy rate of 99.88%, with 0.079% of the validation set being false negatives (or 0.49% of the number of misclassifications). The logistic regression with oversampling gave us an interesting result, as they performed worse than the vanilla logistic regression. The accuracy was 98.01%, with

1.56% of the validation set being false negatives (or 3.12% of the misclassifications). Lastly, the logistic regression with balanced weights achieved the best results: although the accuracy was 97.5%, just 0.011% of the validation set resulted in 3 false negatives (or 0.44% of the misclassifications).

CONCLUSION

Given more time, we would have liked to research and experiment more with adjustments to the layers of the neural network. When looking at other Kaggle submissions, it was clear that some people used very sophisticated techniques to try to optimize these parameters, while our method was mostly guess and check. Many Kagglers found that a random forest model was often the best classifier, so implementing that would be another next step. Additionally, we would like to try to implement an autoencoder or try our hand at an SVM to see how that performed.

REFERENCES

[1] J. Hand, G. Blunt, M.G. Kelly, and N.M. Adams, "Data Mining for Fun and Profit," Statistical Science, vol. 15, no. 2, pp. 111-131, 2000.

[2] "Statistics for General and On-Line Card Fraud," http://www.epaynews.com/statistics/fraud.html, Mar. 2007.

[3] S. Ghosh and D.L. Reilly, "Credit Card Fraud Detection with a Neural-Network," Proc. 27th Hawaii Int?l Conf. System Sciences:Information Systems: Decision Support and Knowledge-Based Systems, vol. 3, pp. 621-630, 1994.

[4] M. Syeda, Y.Q. Zhang, and Y. Pan, "Parallel Granular Networks for Fast Credit Card Fraud Detection," Proc. IEEE Int?l Conf. FuzzySystems, pp. 572-577, 2002.

[5] S.J. Stolfo, D.W. Fan, W. Lee, A.L. Prodromidis, and P.K. Chan, "Credit Card Fraud Detection Using Meta-Learning: Issues and InitialResults," Proc. AAAI Workshop AI Methods in Fraud and Risk Management, pp. 83-90, 1997.

[6] S.J. Stolfo, D.W. Fan, W. Lee, A. Prodromidis, and P.K. Chan, "Cost-Based Modeling for Fraud and Intrusion Detection: Results fromthe JAM Project," Proc. DARPA Information Survivability Conf. and Exposition, vol. 2, pp. 130-144, 2000.

[7] Abdelhalim, A, and I Traore. "Identity Application Fraud Detection using Web". International Journal of Computer and Network Security1, no. 1 (October 2009): 31-44.

[8] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms."Machine Learning, 1991: 37-66.

[9] Aleskerov, Emin, Bernd Freisleben, and Bharat Rao. "Card watch: A neural network based database mining system for credit card frauddetection." Computational Intelligence for Financial Engineering. Piscataway, NJ: IEEE, 1997. 220-226.

[10] Ali, K., and M. Pazzani. "Error reduction through learning multiple descriptions." Machine Learning 24, no. 3 (1996): 173-202.

[11] Basel Committee on Banking Supervision. "Basel Accords II." Basel, Switzerland: Bank for International Settlements Press &Communications, June 2006.

[12] Bolton, R, and D Hand. "Unsupervised Profiling Methods for Fraud Detection." Credit Scoring and Credit Control VII, 2001.

