مجلة بنها للعلوم الإنسانية
العدد (1) الجزء (2) السنة (2022) , (503-497)

الترقيم الدولي الموحد للطباعة:(2537-0170)
الترقيم الدولي الموحد الإلكتروني : (2537-0189)

https://bjhs.journals.ekb.eg

طرق الكشف عن الاحتيال في بطاقة الائتمان - التنقيب عن البيانات

ايه سامى شريف- رحاب شحته محمود- محمد جودة هنداوى
قسم الاحصاء- كلية التجارة- جامعة بنها

الملخص:

أدى وباء كورونا (Covid-19) إلى تقييد حركة الناس إلى مستوى ما، مما جعل من المستحيل شراء المنتجات والخدمات في وضع عدم الاتصال، مما أدى إلى ثقافة الاعتماد المتزايد على خدمات الإنترنت. ويعد الاحتيال أحد أهم الاهتمامات المتعلقة باستخدام بطاقات الائتمان، وهو أمر صعب بشكل خاص في مجال الشراء عبر الإنترنت. نتيجة لذلك، هناك حاجة ماسة لاكتشاف أفضل استراتيجية لاستخدام خوارزميات استخراج البيانات لمنع جميع معاملات بطاقات الائتمان الاحتيالية تقريبًا. لذلك أدى نمو تكنولوجيا المعلومات إلى وجود عدد كبير من قواعد البيانات والمعلومات في مختلف المجالات. يتم إجراء العديد من الدراسات من أجل تغيير هذه البيانات المهمة لاستخدامها في المستقبل. تم استخدام تقنية SMOTE للإفراط في أخذ العينات نظرًا لأن مجموعة البيانات كانت غير متوازنة بشدة. علاوة على ذلك، تم اختيار الميزة، وتم تقسيم مجموعة البيانات إلى جزأين: بيانات التدريب وبيانات الاختبار. الخوارزميات المستخدمة في التجربة هي (Ada Boost (ADB تظهر النتيجة أنه يمكن استخدام كل خوارزمية للكشف عن الاحتيال في بطاقة الائتمان بدقة عالية .كما يمكن استخدام النموذج المقترح للكشف عن المخالفات الأخرى.

الكلمات المفتاحية:
بيانات التعدين- احتيال

I. Introduction:

With the rapid advancement of technology, the world is turning to credit cards rather than cash in their daily lives, which opens the door to numerous new possibilities for dishonest people to use these cards in an unethical manner. Global card losses are likely to hit $35 billion by 2020, according to Nilson research. To safeguard the protection of these credit card customers, the credit card issuer should provide a service that protects consumers from any danger they may encounter (Dal Pozzolo et al., 2017).

The dataset for this study was gathered through research cooperation between Worldline and the Université Libre de Bruxelles's Machine Learning Group on the issue of big data mining and fraud detection. It is made up of numerous transactions made by European cardholders in September of 2013. After PCA transformation, the information is presented as numerical variables to ensure user confidentiality and identification. It is made up of the time between transactions and the quantities of money involved in the transactions (Anis, M., & Ali, M., 2017).

The credit card dataset is substantially unbalanced since it contains more legal transactions than fraudulent ones. That is, without identifying a fraudulent transaction, the prediction will have a very high accuracy score. Class distribution, i.e., sampling minority classes, is a preferable technique to deal with this type of situation. In minority sampling, class training examples can be increased in proportion to the majority class to boost the algorithm's chances of the right prediction (Brownlee, J., 2020).

Several studies are being conducted to identify fraudulent transactions using deep neural networks. These models, on the other hand, are computationally costly and perform better on bigger datasets. This strategy may provide excellent results, as seen by certain studies, but we can obtain the same, or even better, results with fewer resources. So, our major objective is to demonstrate that with proper preprocessing, several machine learning algorithms may provide satisfactory results(Kazemi, Z., & Zarrabi, H.,2017).

As a result, the (Ada Boost) algorithm, according to our findings, brings the highest results, i.e., better determines whether transactions are fraudulent or not. This was assessed using a variety of criteria, including recall, accuracy, and precision. For this type of circumstance, having a high recall value is crucial. The significance of feature selection and dataset balance in producing significant results has been proven.

II. Related Work:

The purpose of data analytics is to uncover hidden patterns and utilize them to make better judgments in a variety of situations. With the growth of updated technology, credit card fraud has increased substantially, making it an easy target for fraudsters. The publicly accessible datasets on credit card fraud are heavily skewed. In the last section, we discussed strategies for identifying credit card fraud. It also goes into its introduction and operation.

Maniraj et al. (2019) focused on data set analysis and preprocessing, as well as the application of multiple anomaly detection techniques to PCA-transformed credit card transaction data, such as the Local Outlier Factor and Isolation Forest Algorithm.

A Bhanusri. et al. (2020) had used Support Vector Machine (SVM), Artificial Neural Networks (ANN), Bayesian Network, K-Nearest Neighbor (KNN), Hidden Markov Model, Fuzzy Logic Based System, and Decision Trees are some of the approaches available for a fraud detection system. A detailed evaluation of current and proposed models for credit card fraud detection has been conducted, as well as a comparison study of different strategies using quantitative metrics such as accuracy, detection rate, and false alarm rate.

Siddhant Bagga et al. (2019) examined the performance of logistic regression, K-nearest neighbors, random forest, naïve Bayes, multilayer perceptron, ada boost, quadrant discriminative analysis, pipelining, and ensemble learning.

Vaishnavi Nath Dornadula et al. (2019) created and designed an unique fraud detection approach for Streaming Transaction Data, with the goal of analyzing customers' prior transaction information and extracting behavioral patterns. Then, using a sliding window method, aggregate the transactions done by cards from different groups in order to derive the behavioral patterns of the groupings. Following that, distinct classifiers are trained on each group independently. The classifier with the highest rating score can then be selected as one of the best approaches for predicting fraud.

Abdulsattar et al. (2020) examined the binary classification problem in situations where the transaction might be either fraudulent or genuine. The objective is to categorize transactions using five different machine learning algorithms: SGD, DT, RF, and J48. Following the application of classifiers, the results are compared to determine which methods perform the best.

J. O. Awoyemi et al. (2017) had used highly skewed credit card fraud data, analyses the performance of naive Bayes, k-nearest neighbor, and logistic regression. The credit card transaction dataset is obtained from European cardholders and contains 284,807 transactions. On the skewed data, a hybrid strategy of under- and over-sampling is used.

R. Sailusha et al. (2020) aimed to concentrate mostly on machine learning methods the random forest algorithm and the Ada boost method were utilized. The results are evaluated using the accuracy; precision, recall, and F1 score of the two approaches. The confusion matrix is used to plot the ROC curve. The Random Forest and Ada boost methods are compared, and the approach with the highest accuracy, precision, recall, and F1-score is regarded the best one for detecting fraud.

Selvani Deepthi Kavila et al. (2018) evaluated and compared machine learning techniques used to identify fraud in credit card systems such as logistic regression, decision trees, and random forests. The suggested system's performance is evaluated using sensitivity, specificity, accuracy, and error rate. The logistic regression, decision tree, and random forest classifiers have accuracy values of 90.0, 94.3, and 95.5 respectively.

III. Materials and Techniques:

A. Dataset:

In this research we used Credit Card Fraud Detection dataset. These datasets contain purchases performed by European cardholders in two days in September 2013. There are 31 numerical features in the dataset. Because some of the input variables contained financial information, the PCA transformation of these input variables was conducted to ensure that the data remained anonymous (v1…v28). Three of the specified characteristics were not converted. The "Time" feature displays the time between the first transaction and each successive transaction in the dataset. The "Amount" feature displays the total amount of credit card transactions. The label is represented by the feature "Class" which has only two values: 1 in the case of a fraudulent transaction and 0 otherwise. The experiment system environment is (Windows 10) operating system and the software operating environment is Google Collab, a scientific python development environment, which is part of the Anaconda platform. Used libraries include NumPy, pandas, matplotlib, sklearn and imblearn, Tensorflow.

B. Our work and results

The proposed technique (Ada Boost Data Mining technique) presented in this thesis could give a good insight into the detection of credit card fraud. We can conclude the following advantages of the proposed technique (Ada Boost Classifier):

1) It is difficult to learn from an unbalanced dataset and the sampling procedure used to balance it. We used 70% of the data is used for training and 30% used for the testing set.
2) There were a few Nan values where the classifier couldn't detect even a single true positive or true negative value. Contributions to future development should be made. Studying resampling approaches that will assist us in minimizing the datasets imbalance ratio and, moreover, using Nan values, remove and improve classifier skewers. Using skewed datasets for improved classification results
3) Anomaly Detection is to eliminate "extreme outliers" from features having a high correlation with our classes (v5, v6, v7).
4) An under-sampling approach was used to balance the data. To compare the models, we employed Accuracy, F1-Score, Recall, Precision, FPR, TRP, and Specificity.
5) The supervised method assists in identifying the label on past transactions; however, it does not detect prior fraud patterns, whereas the unsupervised method aids in detecting the kind of transaction.
6) We will utilize the ANOVA test to select features from a given dataset. The ANOVA test, also known as the Analysis of Variance test, is a statistical tool for comparing the means of two groups of data sets and determining how much they differ. The "Linear Model" is the underlying concept behind the Analysis of Variance, as seen in the following figure(1) proving that ANOVA selected the best 20 datasets.

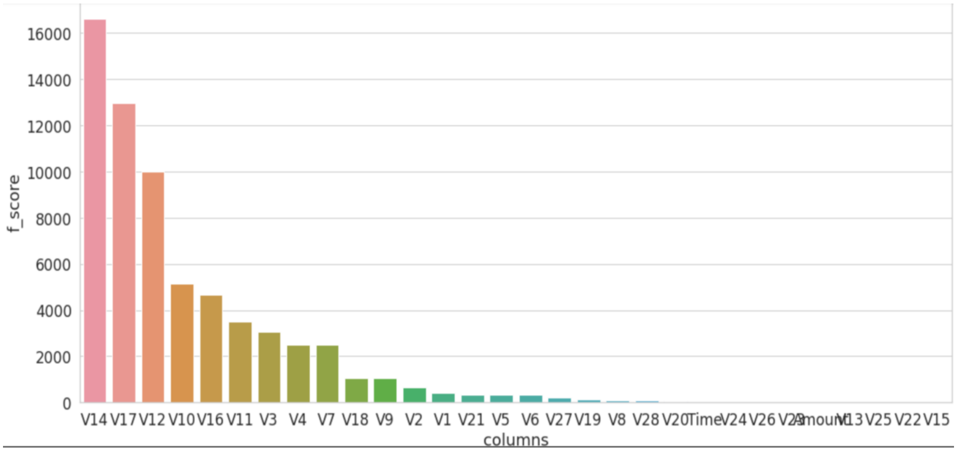

Figure 1 . ANOVA Test

7) Ensemble learning (also known as meta-classifier) improves the predicted outcomes by merging numerous machine learning classifiers. To assess the performance of classification models, we look at various metrics such as F1-Score, Precision, TPR,

FPR, Recall, and Specificity. All these assessment metrics properly reflect the study's validity.

8) Ada Boost (Adaptive Boosting) classifier combines weak classifiers to create a strong classifier. If a poor classifier has high accuracy, it is given greater weight. It is a way of ensemble learning. To enhance accuracy, random forests and XG Boost estimators can be utilized. To reduce the training error, the weak classifier is given a coefficient. This type of boosting is used in conjunction with other algorithms to increase their performance. It uses the three Matthews Correlation Coefficients (MCC) to assess the problem's quality. A score of +1 indicates that the prediction is exact, whereas a score of 1 indicates absolute disagreement.
9) Overall, the stacking classifier, which uses Ada Boost as a meta classifier, appears to be the most promising for identifying fraud transactions in the dataset, followed by XGB, and LR classifiers.

Evaluation metric for final ML Model /n

	Metrics	Results
0	Accuracy	0.968365
1	Precision	0.045664
2	Recall	0.948529
3	F1_score	0.087133
4	AUC	0.958463

Table 1. Performance evaluation of Ada Boost Model

Further Suggestions:

This study may be expanded in several ways to include future research points. The following suggestions are examples of future research:

(1) To get better outcomes, future research should concentrate on other machine learning techniques, such as genetic algorithms and different types of stacked classifiers, as well as broad feature selection.

(2) The vote classifier will be used in future studies, and its performance will be compared against other ML learning approaches, the combined size of the training and testing datasets.

(3) We may work on the top 10 characteristics to determine the accuracy, recall, precision, and confusion matrix and compare it to our previous results.

(4) Based on existing data mining and machine learning approaches, we will create efficient CC fraud detection solutions.

References

[1] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). “Credit card fraud detection: a realistic modeling and a novel learning strategy”. IEEE transactions on neural networks and learning systems, 29(8), 3784-3797.

[2] Anis, M., & Ali, M. (2017). “Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets”. European Scientific Journal, 13(33), 341-353.

[3] Brownlee, J. (2020). “Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning”, Machine Learning Mastery.

[4] Kazemi, Z., & Zarrabi, H. (2017, December). “Using deep networks for fraud detection in the credit card transactions”. In 2017 IEEE 4th International conference on knowledge-based engineering and innovation (KBEI) (pp. 0630-0633). IEEE.

[5] **Maniraj, S. P., Saini, A., Ahmed, S., & Sarkar, S. (2019).** Credit card fraud detection using machine learning and data science. International Journal of Engineering Research, 8(9), 110-115.

[6] Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., & Rohith, R. (2020). “Credit card fraud detection using Machine learning algorithms”. Journal of Research in Humanities and Social Science, 8(2), 04-11.

[7**] Bagga, S., Goyal, A., Gupta, N., & Goyal, A. (2020).** Credit card fraud detection using pipeling and ensemble learning. *Procedia Computer Science*, *173*, 104-112.

[8] **Dornadula, V. N., & Geetha, S. (2019).** Credit card fraud detection using machine learning algorithms. Procedia computer science, 165, 631-641.

[9] Abdulsattar, K., & Hammad, M. (2020, December). “Fraudulent Transaction Detection in FinTech using Machine Learning Algorithms”. In 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT) (pp. 1-6). IEEE.

[10] **Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October).** Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 international conference on computing networking and informatics (ICCNI) (pp. 1-9). IEEE.

[11] **Sailusha, R., Gnaneswar, V., Ramesh, R., & Rao, G. R. (2020, May).** Credit card fraud detection using machine learning. In 2020 4th international conference on intelligent computing and control systems (ICICCS) (pp. 1264-1270). IEEE.