

Fraudulent transactions identification using a machine learning approach

Silvia Vázquez-Noguera¹, Miguel García-Torres², Sebastián Grillo³, Francisco Gómez-Vela², Katherin Arrua¹, Ricardo R. Palma⁴, and Lorena Andrea Bearzotti⁵

¹ Universidad Americana, Asunción, Paraguay

² Data Science and Big Data Lab, Pablo de Olavide University,
ES-41013, Seville, Spain (mgarcia@upo.es)

³ Universidad Autónoma de Asunción, Asunción, Paraguay

⁴ Universidad Nacional de Cuyo, Argentina

⁵ Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Abstract. The rise of e-commerce, driven by digital business models, has made online shopping easier and more convenient. While this benefits both businesses and shoppers, the increase in online transactions also attracts fraudsters. In this work, we address the identification of fraudulent transactions using a machine learning approach. For this purpose we analyze the DataCo Supply Chain Dataset and apply a two-step pre-processing workflow to identify non-informative features. The predictive power of the resulting dataset is tested and compared with classification models induced after applying a feature selection approach. Results suggest that a small number of factors are helpful to build highly predictive models capable of identifying fraudulent transactions.

Keywords: Supply chain, classification, feature selection

1 Introduction

In e-commerce, the analysis of customer behavior like browsing time, purchase history, and credit card usage can help identify fraudulent transactions. However, it's a complex task because fraudsters adapt their tactics. They take a series of steps leading up to a purchase, and constantly change their patterns to avoid detection. While specialists use data analysis to find patterns in fraudulent behavior, the challenge lies in striking a balance as being too cautious might disrupt legitimate purchases [14].

Concerning the techniques used to detect fraud in transactions, we can make the following classification. Firstly, we can differentiate between approaches that use hypothesis-driven mathematical models and data-driven models. The former approach uses mathematical models for assessing the risk of fraudulent transactions [9]. The latter approach deals with the difficulty of unbalance data due to the lack of fraudulent cases.

Within data-driven models we can find semi-supervised [22] and supervised [18] approaches. The second type of techniques have become very popular and we can find the application of classical predictive models such as Logistic Regression, Naive Bayes, Decision trees, Random Forest or Support Vector Machine to detect fraud in the payments [4, 16, 10]. In recent years, proposals based on Artificial Neural Networks or particularly in Deep Learning have also emerged in fraud detection [17, 5]. Some particular architectures in this approach are Competitive Graph Neural Networks, which have been applied to detect suspicious behavior in e-commerce platforms [23] and Convolutional neural networks, which have been applied for the detection of financial fraud on supply networks, under large volumes of data that require a distributed approach [24]. Furthermore, new machine learning approaches have also been applied to this problem as it is the case of Deep Forest [20] and explainable artificial intelligence [6].

In another regard, fraud detection in transactions may involve the evaluation of a large amount of information or attributes per case. In this sense, to improve the computational complexity, it is necessary to perform a feature selection process in order to remove those features that not contribute to the fraud detection. Feature selection is the process of choosing a relevant subset of features from a larger set for machine learning tasks. It helps to improve the performance of learning algorithms by reducing irrelevant and redundant information in high-dimensional data. There are two main categories of feature selection algorithms: filter models and wrapper models. Filter models are computationally cheap but may not be optimal for a specific learning algorithm, while wrapper models consider the chosen learning algorithm but are more expensive to compute [11].

In this work we address the identification of fraudulent transaction using a machine learning approach. The DataCo Supply Chain Dataset [4] was considered to perform this study. We first preprocess the data using a two step approach. First, some irrelevant features were identified and removed manually. Then, all non-informative features, according to information theory, were removed. The predictive power of the dataset was test using Naive Bayes, C4.5, JRIP [2] and oneR [1]. Finally, we applied feature selection to identify a small subset of highly predictive features. For this purpose, Scatter Search algorithm [12] was applied. We can summarize the contributions of this paper as follows:

- We propose a a two-step preprocessing workflow to remove non-informative features,
- Then, we study the predictive power of several classification algorithms to identify fraudulent transactions, and
- Finally, we find out a small subset of highly predictive and non-redundant features.

This document is organized as follows. In the Section 2 we present the techniques used to obtain the results. In the Section 3 we detail the application and results obtained. Finally, in the Section 4 we present our conclusions.

2 Methodology

This section presents the techniques and the dataset used in this work. We first describe the problem of classification followed by the feature selection problem and, finally, we describe the dataset.

2.1 Classification

In classification task, the aim is to build a function to relate input and output factors. Formally, given a set $\mathcal{T}$ of n instances described by $(\mathbf{x}_i, y_i)$, with $\mathbf{x}_i$ a vector described by d features, and y_i its corresponding class label. Then, the classification problem can be defined as the map $f : \mathbf{X} \rightarrow \mathcal{Y}$. The different classification algorithms are described below:

Naive Bayes (NB). This algorithm, unlike more complex models, assumes that features are independent. Despite this seemingly unrealistic assumption, Naive Bayes performs well in practice, even against more sophisticated techniques. The key to this success lies in the fact that optimal classification does not require a perfect match to the underlying probability distribution. As long as both the real and estimated models agree on the most likely class for a given example, Naive Bayes achieves good results. This is supported by theoretical evidence that it is an optimal solution for specific types of problems with large feature dependencies [13].

C4.5 . It is an algorithm that builds decision trees for classification tasks. It uses a training dataset where each sample has features (attribute values) and a known class label. The algorithm works by recursively splitting the data at each node of the tree based on the feature that best separates the samples according to their class. This separation is measured using normalized information gain, which essentially tells us how much learning about a feature reduces uncertainty about the class. The process continues until all samples at a node belong to the same class, no feature provides further separation or a new class is encountered (handled by assigning the most likely class). This approach allows C4.5 to learn complex decision rules from data [15].

JRIP. This classification algorithm is a rule-learning algorithm well-suited for large datasets with imbalanced classes and noisy data. It builds a set of rules by iteratively separating the classes based on frequency. For each class, it uses a sequential covering approach to find rules that distinguish it from all others. The algorithm employs a general-to-specific strategy to build rules, stopping when negative examples are included. It then prunes the rule based on validation set performance and removes covered instances from the training data. Finally, JRIP stops adding rules when a minimum description length threshold is met or the validation set error exceeds 50%. Additionally, it can optimize the rule set by potentially replacing existing rules with better alternatives [2].

oneR. This is a simple supervised learning algorithm which is better understood as "One Feature" rather than "One Rule." It focuses on finding a

single feature that best separates data based on class labels. Despite its limitations in prediction accuracy, OneR is valuable for establishing baseline performance [1].

2.2 Factor identification

To identify the subset of factors relevant to the classification problem we apply a two step approach. In the first step we identify, manually, those factors that are irrelevant while in the second one, we apply feature selection.

In feature selection, the aim is to select a small subset of highly predictive features. Therefore, it seeks to reduce the number of attributes in the dataset, while keeping (or improving) the predictive power of the classification model induced. To assess a subset of features we consider an objective function, $J(S)$, with S the feature subset of size $|S|$.

Information theory-based metrics are gaining attention due to the ability of such measures to quantify the uncertainty. In this approach entropy measures the average amount of information contained of a features. Given two features X and Y , the Information Gain (IG) is another key metric defined as:

$$I(X|Y) = H(X) - H(X|Y), \quad (1)$$

with $H(X|Y)$ the conditional entropy, which measures the uncertainty about the value of X given the value of Y . IG computes the increase in information contained in a variable Y when the value of another feature X is known. Despite the popularity of this metric, it suffers from bias towards features with more values. To address this, symmetrical uncertainty is presented in the literature as a normalized alternative, ranging from 0 (indicating independence between feature and class) to 1 (where knowing one perfectly predicts the other). The Symmetrical Uncertainty for X and Y is defined as

$$SU(X, Y) = 2 \left[\frac{H(X) - H(X|Y)}{H(X) + H(Y)} \right]. \quad (2)$$

This concept allows to objectively assess feature relevance in the context of machine learning tasks, ultimately selecting features that are most informative for predicting the desired outcome [7].

In this work we consider as $J(\cdot)$ the Correlation based Feature Selection (CFS) [8], which is defined as follows:

$$J(S) = \frac{|S| \cdot \langle SU_{cf} \rangle}{\sqrt{|S| + |S| \cdot |S| - 1} \cdot \langle SU_{ff} \rangle} \quad (3)$$

with $\langle SU_{cf} \rangle$ the mean feature-class correlation ($f \in S$), and $\langle su_{ff} \rangle$ is the average feature-feature inter-correlation.

In this work we apply the Scatter Search (SS) [12]. Scatter Search is an evolutionary population-based metaheuristic that tackles complex optimization problems by combining the strengths of existing solutions. It leverages two key

ideas: 1) extracting new information by combining existing solutions, and 2) employing additional problem-specific methods to evaluate and refine these new possibilities.

2.3 Dataset

In this work we use the DataCo Supply Chain Dataset [3], which has been provided by the company DataCo Global for its analysis. It consist of around 180k samples about transaction from supply chains for 3 years. The two aims of this dataset are to detect fraud transactions and late delivery of orders. In this work we focus on the first objective. The dimensionality of the dataset is 32 (after removing the identifiers) plus the class label.

3 Computational results

This section presents the experiments conducted. The aims of the experiments are:

- Identify the irrelevant factors.
- Analyze the predictive power of the classification models.
- Identify a subset of highly predictive factors.

To compare the different predictive models we use the k -fold cross-validation with $k = 5$.

3.1 Data preprocessing

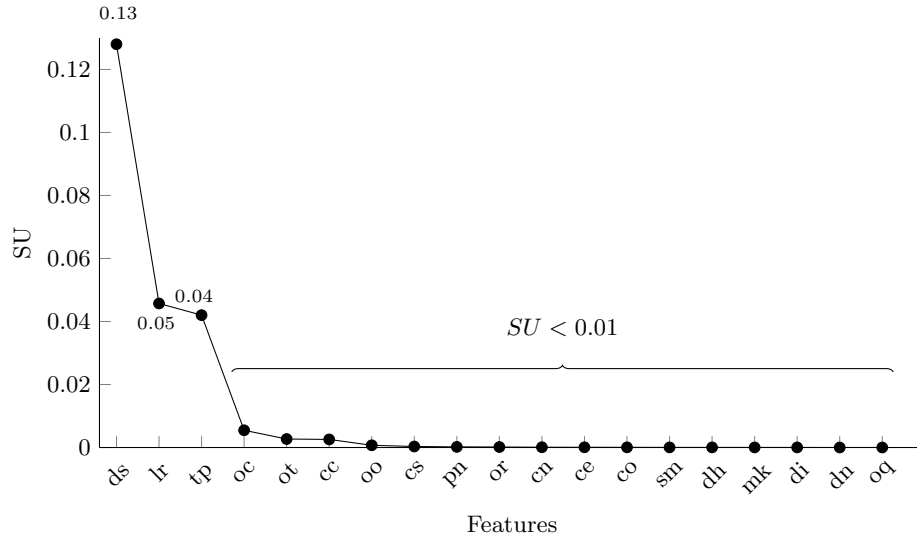
This section is focused on selecting a subset of irrelevant features manually. To remove them, it is necessary to check that they do not contain information about the class. This is done by calculating SU between each feature and the label.

All the identifiers were removed. The list of features identified as irrelevant is shown in Table 1. The different columns refer to the reason for which it was removed. Column Unique refers to the case in which the feature has only a unique value. Then, Distinct column is associated to column in which the number of categorical values is too large. Finally, Missing column refers to theses features that contain too many missing values ($> 80\%$). The feature Product Image was also included because it was identified as irrelevant. Once the irrelevant features were removed, the dimension decreased to 32.

Now we proceed to analyze the different factors by calculating SU between each factor and the class and those that have no information about the class label are removed. Figure 1 shows those with $SU > 0$. In this step we removed the following features: Order item Total, Order Profit per Order, Sales per Customer, Benefit per Order, product Price, Order Item Profit Ratio, Order Item product price, Order Item Discount, Customer Zipcode, Order Item Discount Rate and Sales.

Table 1. Metrics collected at the application level on video and web-based services.

Factor	Unique	Distinct	Missing
Customer Email	✓	—	—
Customer Fname	—	—	—
Customer Lname	—	—	—
Customer Password	✓	—	—
Customer Street	—	✓	—
Latitude	—	✓	—
Longitude	—	✓	—
Order date	—	✓	—
Order Zipcode	—	—	✓
Product description	—	—	✓
Product Image	—	—	—
Product Status	✓	—	—
Shipping date	—	✓	—

**Fig. 1.** Ranking of factors according to SU measure. Each factor is represented by an id: delivery status (ds), late delivery risk (lr), type (tp), order city (oc), order state (ot), customer city (cc), order country (oo), customer state (cs), product name (pn), order region (or), order zipcode (oz), category name (cn), customer segment (ce), customer country (co), shipping mode (sm), days for shipment (dh), market (mk), Days for Shipping (di), department name (dn), order item quantity (oq).

3.2 Predictive analysis

This section presents the results of the different predictive models using all the factors. Table 2 presents the performance of the different classification algorithms used in this work. It shows the accuracy, sensitivity and specificity for the following cases:

- i Baseline dataset. In this case the number of factors is 30 and all the features that have been manually identified as irrelevant (see Table 1 have been removed.),
- ii Informative dataset. In addition to the previous case, all the features without information about the class label ($SU = 0$) have also been removed. In this case the dimensionality of the dataset is 19.
- iii Feature selection. In this case, we apply the feature selection algorithm Scatter Search to the Informative dataset to identify a small subset of predictive factors.

With baseline dataset, C4.5 is the algorithm that achieves better accuracy followed by NB and JRIP. Finally, OneR has the lowest accuracy but it is the only algorithm that correctly classifies all the fraudulent transactions (sensitivity of 1.00). Checking the informative dataset we can see that the removal of non informative features has no effect on the the predictive model of NB, C4.5 and OneR. Only for the case of JRIP the values change a little bit. Finally, with feature selection all the models achieve an accuracy of 97.9%. However, only for JRIP and OneR the sensitivity is 1.00. For the case, the factors selected by Scatter Search were Delivery Status, Order Zipcode and Customer Segment.

Table 2. Performance of the predictive models.

	Baseline			Informative			Feature selection		
	acc	sen	spe	acc	sen	spe	acc	sen	spe
NB	98.19	0.94	0.98	98.19	0.94	0.98	97.95	0.86	0.98
C4.5	99.39	0.88	0.99	99.39	0.88	0.99	97.96	0.94	0.98
JRIP	98.08	0.85	0.98	98.07	0.88	0.98	97.96	1.00	0.98
OneR	97.96	1.00	0.98	97.96	1.00	0.98	97.96	1.00	0.98

3.3 Discussion

Results achieved with JRIP and OneR using feature selection are suitable to address the problem since they are able to identify all fraudulent transactions. In addition, it is worth noting that the number of false positives (incorrectly instances labeled as fraudulent) in both models is small.

The predictive models studied in this work achieve accuracies that range from 97.96% to 99.39%. The predictive power of such models are similar to

those found by Tamim et al. [19] and outperforms those found by Constante-Nicolalde et al. [4] and Ye [21]. The preprocessing approach could be the reason of the differences found.

According to the SU-based ranking (see Figure 1), the most informative factors are Delivery Status (ds), Late delivery Risk (lr) and Type (tp). The remaining features have $SU < 0.01$. However, the factors selected by Scatter Search algorithm -Delivery Status (ds), Late Delivery Risk (lr) and Type (tp)- are not the top 3 most informative due to redundancy. So, despite the lower information that some factors may have, the information is not redundant and is necessary to take it into account to achieve a high performance.

4 Conclusions and future work

In this work we have address the detection of fraudulent transactions problem. We have applied a machine learning approach to build a good predictive model and to identify a small subset of highly predictive factors.

In order to clean the data we have a two step preprocessing pipeline. In the first step some factors were removed due to the presence of too many missing values, or because the factor had a unique value or it had too many categorical values. In the second step, we applied SU to identify those factors that contained no information about the class.

We analyzed the dataset with four different classification algorithms and results suggest that, in general, the dataset contain high-quality information and, so, the models built reach a high accuracy. However, the most interesting models are those that find all the fraudulent transactions. In this context, JRIP and OneR are the most interesting classifiers, specially the second one since it has shown to be quite robust and identify all the fraudulent transaction in the different versions of the datasets.

Finally we conclude that despite the different works that have addressed this problem with the dataset considered in this work, more research is necessary to understand the importance of the factors involved in the classification models. Most works still miss information that is important to replicate the results.

As a future work, we want to investigate by applying explainable method to understand better the process of labeling. We also want to analyze in a deeper way the relevance and redundancy of the different factors studied.

Acknowledgments

The authors would like to thank the Spanish Ministry of Science and Innovation and Junta de Andalucía for the support under the projects PID2020-117954RB-C21, PY20-00870 and UPO-138516, respectively. The authors would also like to thank the Organismos nacionales de política de Ciencia y Técnica (ONCyT) and the Executive Committee and the CyTED Network for their support through the project P321RT0088.

References

1. G. Buddhinath and D. Derry. A simple enhancement to one rule classification. *Department of Computer Science & Software Engineering University of Melbourne, Australia*, 2006.
2. W. W. Cohen. Grammatically biased learning: Learning logic programs using an explicit antecedent description language. *Artificial Intelligence*, 68(2):303–366, 1994.
3. F. Constante, F. Silva, and A. Pereira. Dataco smart supply chain for big data analytics data, 2021.
4. F.-V. Constante-Nicolalde, P. Guerra-Terán, and J.-L. Pérez-Medina. Fraud prediction in smart supply chains using machine learning techniques. In *International Conference on Applied Technologies*, pages 145–159. Springer, 2019.
5. H. Fanai and H. Abbasimehr. A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. *Expert Systems with Applications*, 217:119562, 2023.
6. P. Fukas, J. Rebstadt, L. Menzel, and O. Thomas. Towards explainable artificial intelligence in financial fraud detection: Using shapley additive explanations to explore feature importance. In *International Conference on Advanced Information Systems Engineering*, pages 109–126. Springer, 2022.
7. M. García-Torres, R. Ruiz, and F. Divina. Evolutionary feature selection on high dimensional data using a search space reduction approach. *Engineering Applications of Artificial Intelligence*, 117:105556, 2023.
8. M. A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.
9. S. Hongjin. Analysis of risk factors in financial supply chain based on machine learning and iot technology. *Journal of Intelligent & Fuzzy Systems*, 40(4):6421–6431, 2021.
10. E. Ileberi, Y. Sun, and Z. Wang. A machine learning based credit card fraud detection using the ga algorithm for feature selection. *Journal of Big Data*, 9(1):24, 2022.
11. S. S. Kannan and N. Ramaraj. A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm. *Knowledge-Based Systems*, 23(6):580–585, 2010.
12. F. G. López, M. G. Torres, B. M. Batista, J. A. M. Pérez, and J. M. Moreno-Vega. Solving feature subset selection problem by a parallel scatter search. *European Journal of Operational Research*, 169(2):477–489, 2006.
13. I. Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. Citeseer, 2001.
14. V. F. Rodrigues, L. M. Policarpo, D. E. da Silveira, R. da Rosa Righi, C. A. da Costa, J. L. V. Barbosa, R. S. Antunes, R. Scorsatto, and T. Arcot. Fraud detection and prevention in e-commerce: A systematic literature review. *Electronic Commerce Research and Applications*, 56:101207, 2022.
15. S. Ruggieri. Efficient c4. 5 [classification algorithm]. *IEEE transactions on knowledge and data engineering*, 14(2):438–444, 2002.
16. A. Saputra et al. Fraud detection using machine learning in e-commerce. *International Journal of Advanced Computer Science and Applications*, 10(9), 2019.

17. J. Soni, P. Gangwani, S. Sirigineedi, S. Joshi, N. Prabakar, H. Upadhyay, and S. A. Kulkarni. Deep learning approach for detection of fraudulent credit card transactions. In *Artificial Intelligence in Cyber Security: Theories and Applications*, pages 125–138. Springer, 2023.
18. Y. Sun, A. K. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.
19. L. Tamym, A. N. S. Moh, L. Benyoucef, and M. D. E. Ouadghiri. Goods and activities tracking through supply chain network using machine learning models. In *IFIP International Conference on Advances in Production Management Systems*, pages 3–12. Springer, 2021.
20. L. Wang, Z. Zhang, X. Zhang, X. Zhou, P. Wang, and Y. Zheng. A deep-forest based approach for detecting fraudulent online transaction. In *Advances in computers*, volume 120, pages 1–38. Elsevier, 2021.
21. L. Ye. Research on supply chain big data management based on machine learning algorithms. In *2022 International Conference on Cloud Computing, Big Data Applications and Software Engineering (CBASE)*, pages 207–213. IEEE, 2022.
22. D. Zage, K. Glass, and R. Colbaugh. Improving supply chain security using big data. In *2013 IEEE International Conference on Intelligence and Security Informatics*, pages 254–259. IEEE, 2013.
23. G. Zhang, Z. Li, J. Huang, J. Wu, C. Zhou, J. Yang, and J. Gao. efraudcom: An e-commerce fraud detection system via competitive graph neural networks. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–29, 2022.
24. H. Zhou, G. Sun, S. Fu, X. Fan, W. Jiang, S. Hu, and L. Li. A distributed approach of big data mining for financial fraud detection in a supply chain. *Comput Mater Continua*, 64(2):1091–1105, 2020.