Financial Fraud Detection

Özge Güney - 2100365

Department of Computer Engineering, Engineering Faculty, Bahçeşehir University,
ozge.guney2@bahcesehir.edu.tr

Abstract

Today, artificial intelligence is being developed for many purposes. Developing technology and increasing e-commerce volume make it easier for fraudsters. Therefore, fraud prevention and detection is more important than ever. In the early stages, the data used for fraud detection is often highly structured, such as transaction logs. However, these manual and rule-based approaches have now become costly and ineffective. Also, rule-based methods cannot detect new patterns in data because they only follow a previously created scenario. The small set of rules outlined by man is no longer sufficient to meet the demand. Unlike legacy rule-based methods, machine learning algorithms process raw data such as emails or texts. Then, these algorithms learn from the input and get smarter. Today, many machine learning-based methods have been developed. In this article, I present data-driven machine learning-based methods to deal with the problems of old-school rule-based methods. The proposed study uses and evaluates the accuracy results of different classification models such as Naive Bayesian Classifier, K-Nearest Neighbors, Decision Trees and Support Vector Machine. I tested the algorithms on a synthetic dataset that was scaled down to 1/4 of the original dataset and created only for Kaggle.

Keywords: Fraud Detection, K-Nearest Neighbors, Decision Trees, Support Vector Machine.

1. Introduction

Financial fraud can be defined as a deliberate action of cheat involving financial transactions for individual gain[1]. Fraud detection, too, is a set of activities performed to prevent money from being obtained because of false claims. The use of e-commerce is increasing day by day and increasing e-commerce volume attracts scammers. Therefore, preventing and detecting fraud is becoming more important than ever. All banking institutions should be aware of the potential risks in their organization as well as sources of fraud. While technology has allowed for easier methods of digital payments and transactions, it has also created opportunities for fraud. In other words, increased digitization has also exposed businesses to online frauds and scams. As a result, companies began to focus on ways to reduce security vulnerabilities in payment systems.

The global fraud detection and prevention market is growing. With AI platform and innovation of new technologies, banks can reduce the risk of fraud. AI can successfully help detect and prevent fraud in banking. With artificial intelligence and machine learning techniques, banks can use data to predict and prevent future fraudulent events.

Manual and rule-based approaches have now become costly and ineffective[2]. Also, rule-based methods cannot detect new patterns in data because they only follow a previously created scenario. The small set of rules outlined by man is no longer sufficient to meet the demand. In this article, I present data-driven machine learning-based methods to deal with the problems of old-school rule-based methods. The proposed study uses and evaluates the accuracy results of different classification models such as Naive Bayesian Classifier, K-Nearest Neighbors, Decision Trees and Support Vector Machine.

2. Fraud Detection Methods

2.1. Rule-based Methods

A set of rules can be used to filter fraud transactions. However, these manual and rule-based approaches have now become particularly costly and ineffective. Moreover, rule-based methods cannot detect new patterns in data because they only follow a predetermined scenario. The small set of rules outlined by man is no longer sufficient to meet demand.. Firstly, I look at my dataset and get the type of transactions in Figure 1. As seen in Figure 2, based on the existing rules, while 5% of the transactions are fraudulent, it takes into account of the 2% of the total amount.

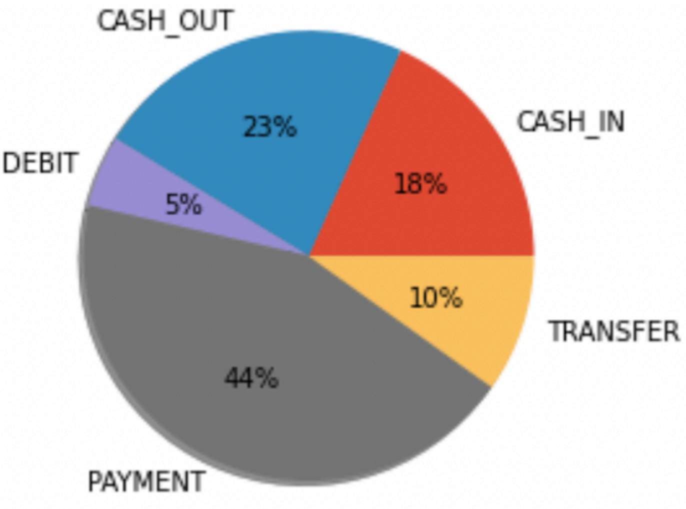

Figure 1. Type of Transactions

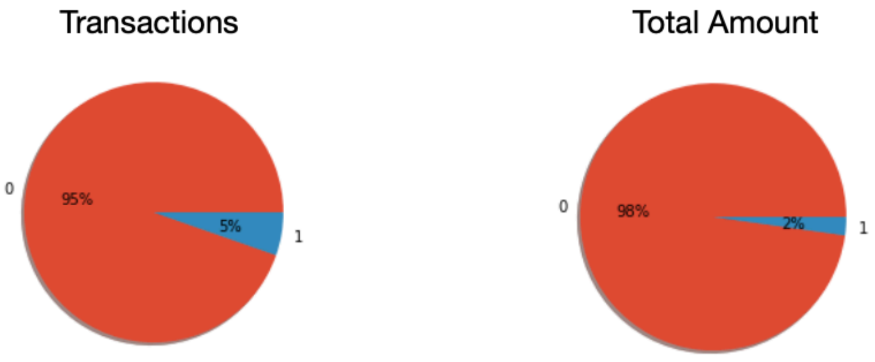

Figure 2. How much of our dataset is fraudulent?

2.2. Artificial Intelligence Techniques Used for Fraud Detection

In this section, techniques I used are summarized. I use accuracy metrics to measure the correctness of the fraud detection by using confusion matrix.

A confusion matrix is a technique for summarizing the performance of a classification algorithm. It may described as in Table 1. I use this matrix to evaluate the methods.

Table 1. Confusion Matrix

		Predicted Condition	
		Fraud	Non-Fraud
Actual Condition	Fraud	TP	FP
	Non-Fraud	FN	TN

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERROR. The formula is as in Eq. (1):

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

The simple form of the calculation for Bayes Theorem is as in Eq. (2):

$$P(A|B) = P(B|A) * P(A)/P(B) \quad (2)$$

Naive Bayesian Classifier: The naive Bayes classifier divides the data into different classes according to Bayes' Theorem, assuming that all estimators are independent of each other. It assumes that a particular property in a class is unrelated to the existence of other properties. Having correlation will go against the Naive assumption. Correlation between features in Naive Bayes simply means that if one feature "says" it's class A, then the other feature(s) will often say the same. the accuracy is 0.60 because of correlation. I didn’t used correlation method and confusion matrix is as in Figure 3.

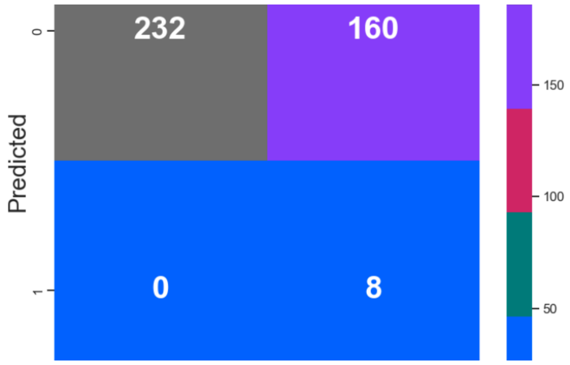

Figure 3. Naive Bayes Confusion Matrix

Support Vector Machine (SVM): In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Given a set of training examples, each marked as belonging to one of thetwo categories, an SVM training algorithm creates a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM maps training samples to points in space to maximize the width of the gap between the two categories. New samples are then mapped to the same area and predicted to belong to a category based on which side of the gap they fall on [6]. As we see in Figure 4, H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximal margin [7]. While using this method, accuracy for the given dataset is 0.98. Confusion matrix is as in Figure 5.

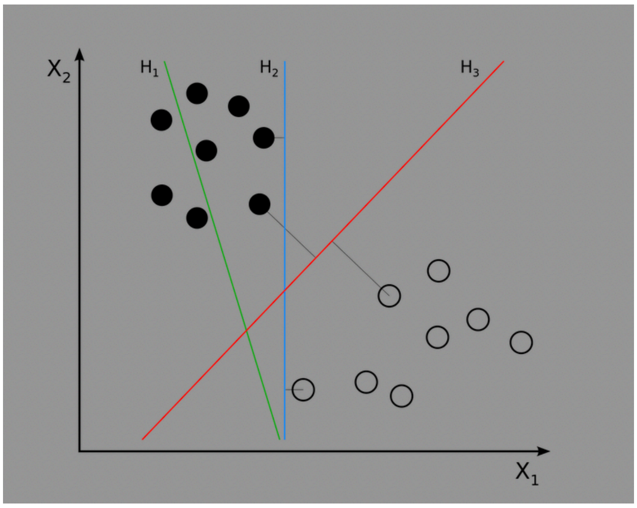

Figure 4. SVM Logic

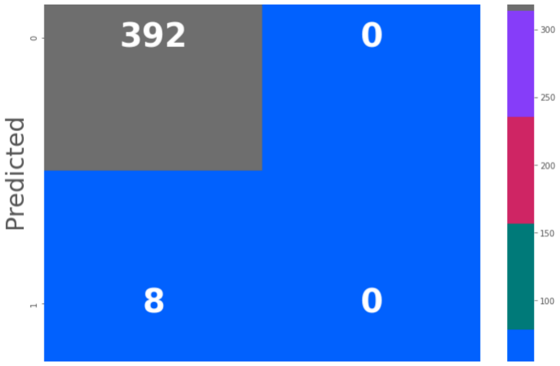

Figure 5. Confusion Matrix SVM

Decision Trees: Decision trees and their ensembles are popular methods for the machine learning tasks of classification and regression. Decision trees are widely used since they are easy to interpret, handle categorical features, extend to the multi-class classification setting, do not require feature scaling, and are able to capture non-linearities and feature interactions [8]. The tree of the dataset I used is as in Figure 6. I look at the gini values to see if the transaction is fraudulent. As we see in Figure 6, the tree depth is 3. While using this method, accuracy for the given dataset is 0.992. Confusion matrix is as in Figure 7.

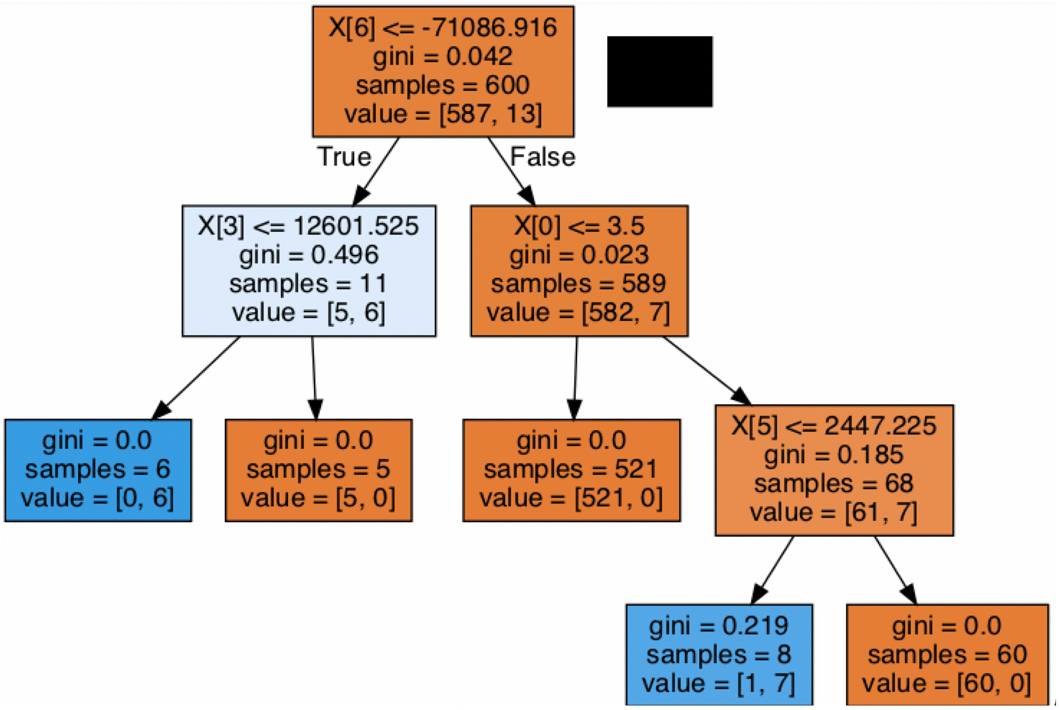

Figure 6. Decision Tree

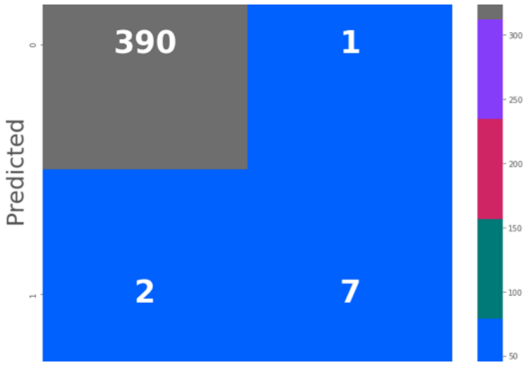

Figure 7. Confusion Matrix Decision Tree

K-Nearest Neighbors (KNN): The k-nearest neighbor classifier assigns an instance to the class most heavily represented among its neighbors. It is based on the idea that the more similar the instances, the more likely it is that they belong to the same class. We can use the same approach for classification as long as we have a reasonable similarity or distance measure [12]. Nearest mean classifier formula is as in Eq. (3) where we choose Ci if

$$D(x, m_i) = \min_{j=1} D(x, m_j) \qquad (3)$$

In the case of hyper-spheric Gaussians where dimensions are independent and all are in the same scale, the distance measure is the Euclidean as in Eq. (4) [12]:

$$D(x, mi) = \| x - mi \| \tag{4}$$

As in Figure 8, The test sample (green dot) should be classified either to blue squares or to red triangles. If k = 3 (solid line circle) it is assigned to the red triangles because there are 2 triangles and only 1 square inside the inner circle. If k = 5 (dashed line circle) it is assigned to the blue squares (3 squares vs. 2 triangles inside the outer circle) [13].

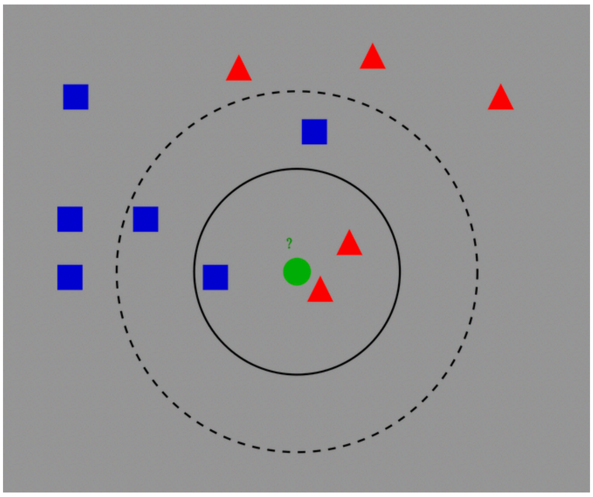

Figure 8. Example of KNN classification

In process of KNN, we classify any incoming transaction by calculating of nearest point to new incoming transaction. Then if the nearest neighbor be fraudulent, then the transaction indicates as a fraud. Larger K values can help to reduce the effect of noisy data set [10]. In my project, I choose K value 3 by using elbow method. While using this method, accuracy for the given dataset is 0.985. Confusion matrix is as in Figure 9.

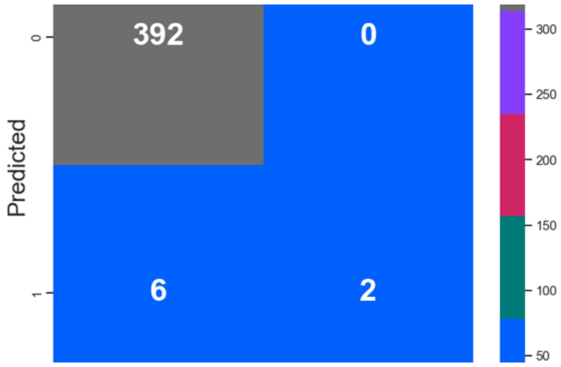

Figure 9. Confusion matrix KNN

3. Conclusion

Naive Bayes is the worst method in this application. Because, this method is excellent for the dataset which have independent features. Among these four methods, the most accurate one is Decision Tree Model. Its accuracy is 0.992. This result is very good. However, in order to get this result in real life, the rules determined by the experts on the subject must be clear. Moreover, data must be interpretable. Today, data isolation is difficult to solve. While data-driven AI techniques.have achieved excellent performance in financial fraud detection, these problems must be addressed.

References

1. https://en.wikipedia.org/wiki/Data_analysis_techniques_for_fraud_detection
2. https://www.sciencedirect.com/science/article/abs/pii/S1566253509000141
3. https://www.vynzresearch.com/ict-media/fraud-detection-and-prevention-market
4. https://www.bluegranite.com/blog/detecting-financial-fraud-with-machine-learning
5. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/#:~:text=Naive%20Bayes%20uses%20a%20similar,with%20problems%20having%20multiple%20classes
6. https://analyticsindiamag.com/understanding-the-basics-of-svm-with-example-and-python-implementation/
7. https://en.wikipedia.org/wiki/Support-vector_machine
8. https://pages.databricks.com/rs/094-YMS-629/images/financial-fraud-detection-decision-tree.html
9. https://www.sciencedirect.com/science/article/abs/pii/S0957417413003072
10. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
11. https://veribilimcisi.com/2017/07/20/k-en-yakin-komsu-k-nearest-neighborsknn/
12. Introduction to Machine Learning, Third Edition by Ethem Alpaydin, the MIT Press, 2017, pp.196-197. (for Book Reference)
13. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm