Universitatea Tehnică a Moldovei

EVALUAREA ALGORITMILOR DE ÎNVĂȚARE AUTOMATĂ ÎN DETECTAREA TRANZACȚIILOR FRAUDULOASE

EVALUATION OF MACHINE LEARNING ALGORITHMS IN FRAUDULENT TRANSACTION DETECTION

Student: **Verebceanu Mirela,**
gr. IS-211M

Coordonator: **Zaharia Gabriel,**
asist. univ.

Chişinău, 2023

Rezumat

Teza de masterat cu titlul ”Evaluarea algoritmilor de învățare automată în detectarea tranzacțiilor frauduloase”, a fost elaborată de studenta grupei IS-211M, Verebceanu Mirela, Universitatea Tehnică a Moldovei. Aceasta este scrisă în limba engleză și conține 52 pagini, 4 tabele, 16 figuri, 9 listări de cod și 23 de referințe. Teza este alcătuită dintr-o listă de figuri, o listă de listări de cod, introducere, patru capitole, concluzie și bibliografie.

Tranzacțiile cu cardul, utilizând cardul fizic, aplicațiile mobile sau descktop sunt în creștere, mai ales în ultimii ani. Prin urmare, s-a detectat o creștere și în crimele financiare, cum sunt fraudarea tranzacțiilor. Pe lângă aceasta, domeniul inteligenței artificiale se devoltă continuu și mai multe modele de învățare automată pot fi utilizate pentru detectarea tranzacțiilor frauduloase. O problemă cu care se pot confrunta aceste modele este setul de date cu informații despre tranzacții, deoarece consține cu mult mai multe tranzacții legitime decât fraude.

Obiectivele acestei lucrări au ca scop analiza și descrierea fraudelor tranzacționale, colectarea și pregătirea setului de date, identificarea algoritmilor de învățare automată potrivit și dezvoltarea acestora, compararea modelelor utilizând metricile de evaluare corespunzatoare.

Pentru realizarea cercetării am studiat articole din domeniul financiar și învățare automată. Am analizat trei seturi de date despre tranzacții, am colectat datele și le-am comparat. Datele au fost pregătite pentru a lucra cu modelele de învățare automată. Am cercetat și evaluat patru modele de înavățare automată.

Rezultatele obținute sunt reprezentate în cele patru capitole. În primul capitol se descrie domeniul de cercetare, tipurile de fraude, tehnicile de detectare a unei fraude. Capitolul doi este despre analiza și vizualizarea datelor, iar capitolul trei depsre modelele de învățare automată studiate. În ultimul capitol este efectuată evaluarea celor patru algoritmi.

Acest document este destinat cititorilor specializați în domeniul tehnic.

Cuvinte-cheie: tranzacție, fraudă, date, algoritmi, metrici de evaluare.

Abstract

The master's thesis entitled ”Evaluation of machine learning algorithms in fraudulent transaction detection” was developed by the student of the IS-201M group, Mirela Verebceanu, Technical University of Moldova. It is written in English and contains 52 pages, 4 tables, 16 figures, 9 listings, and 23 references. The thesis consists of a list of figures, a list of listings, an introduction, four chapters, a conclusion, and a list of references.

Card transactions, using the physical card, mobile or desktop applications are increasing, especially in recent years. Therefore, an increase was also detected in financial crimes, such as transaction fraud. In addition, the field of artificial intelligence is still developing and more machine learning models can be used to detect fraudulent transactions. One problem these models may face is the transaction information data set, as it contains far more legitimate transactions than fraud.

The objectives of this paper aim to analysis and description of transactional frauds, collection and preparation of data sets, identification of suitable automatic learning algorithms and their development, comparison of models using appropriate evaluation metrics.

To conduct the research, I studied articles from the financial field and machine learning area. Have analyzed three sets of transaction data, collected the data and compared them. The data was prepared for the automatic learning models. Have researched and evaluated four automatic learning models.

The obtained results are represented in the four chapters. The first chapter describes the research field, fraud types, fraud detection techniques. The second chapter is about data analysis and visualization, and the third chapter is about the studied automatic learning models. In the last chapter, the evaluation of the four algorithms is performed.

This document is intended for readers with technical background.

Key words: transaction, fraud, data, algorithm, evaluation metrics.

Contents

INTRODUCTION

In the era of technologies, world is changing and evolving extremely fast. The financial area also changed a lot. COVID19 was one more shift to online platforms and led to implement more IT solution in every industry. Therefore, financial crimes have increased significantly. This problem caused monetary losses all around the world.

One of the most eloquent issues in e-commerce is credit card fraud. Card fraud represents a form of identity theft, an unauthorized usage of the credit card information. Since technological progress, new avenues for credit card fraud have been created. Nowadays perpetrators do not need physical cards to make unauthorized purchases.[1]

Worldwide financial losses caused by fraudulent activities are worth tens of billions of dollars. The newest European financial organization ECB report analyses trends in card fraud in 2019 which are based on card payment schemes running within the euro area. The analysis concentrates on data for 2019, that are put into the context of a five-year time frame from 2015 to 2019. Card payment schemes running within the euro area report data reduced by Single Euro Payments Area SEPA country, covering almost the entire card market. Card fraud are considered (i) fraudulent transactions with physical cards (card-present fraud), like cash withdrawals with counterfeit or stolen cards; and (ii) fraudulent transactions conducted remotely (card-not-present fraud), for instance where criminals conduct online payments with card information acquired through phishing or data breaches. The whole value of fraudulent transactions using cards issued within SEPA and bought worldwide amounted to €1.87 billion in 2019. For cards issued within the euro area, the total value of fraudulent card transactions amounted to €1.03 billion.[2]

REFERENCES

[1] Cornell Law School. *Legal Information Institute*. https://www.law.cornell.edu/wex/credit_card_fraud. Last accessed on: 11/10/2022.

[2] European Central Bank. *7th report on card fraud.* https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202110~cac4c418e8.en.html. Last accessed on: 11/10/2022.

[3] VITALINA SIMERET, CHII. *Tot mai mult,i moldoveni ˆıs,i achită cumpărăturile cu cardul sau prin e-comert,*. https://diez.md/. Last accessed on: 11/10/2022.

[4] SRII SRINIVASAN. *How Banks Conduct Transaction Fraud Investigations*. https://www.chargebackgurus.com/blog/transaction-fraud-investigations. Last accessed on: 11/10/2022.

[5] Infosys BPM. *How can machine learning help in credit card fraud detection?* https://www.infosysbpm.com/blogs/bpm-analytics/machine-learning-for-credit-card-fraud-detection.html. Last accessed on: 11/10/2022.

[6] YANN-A Ë L LE BORGNE et al. *Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook*. 2022. URL: https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook.

[7] E. A. LOPEZ-ROJAS, A. ELMIR, and S. AXELSSON. *PaySim: A financial mobile money simulator for fraud detection*. https://www.kaggle.com/datasets/ealaxi/paysim1. Last accessed on: 11/10/2022.

[8] *IEEE-CIS Fraud Detection*. https://www.kaggle.com/competitions/ieee-fraud-detection. Last accessed on: 15/11/2022.

[9] *Standard Scaler*. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Last accessed on: 09/11/2022.

[10] BENAI KUMAR. *10 Techniques to deal with Imbalanced Classes in Machine Learning*. https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/. Last accessed on: 09/11/2022.

[11] *Imbalanced learn*. https://imbalanced-learn.org/stable/references/. Last accessed on: 09/11/2022.

[12] SWASTIK SATPATHY. *Overcoming Class Imbalance using SMOTE Techniques*. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/. Last accessed on: 09/11/2022.

[13] *Machine Learning Models*. https://www.databricks.com/glossary/machine-learning-models. Last accessed on: 09/11/2022.

[14] JR. DAVID W. HOSMER, STANLEY LEMESHOW, and RODNEY X. STURDIVANT. *Applied Logistic Regression.* , 2013 John Wiley Sons, Inc. Published 2013 by John Wiley Sons, Inc.

[15] MIRKO STOJILJKOVIĆ. *Logistic Regression in Python*, https://realpython.com/logistic-regression-python/. Last accessed on: 09/11/2022.

[16] *XGBoost Documentation*. https://xgboost.readthedocs.io/en/stable/. Last accessed on: 09/11/2022.

[17] R SRUTHI. *Understanding Random Forest*. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. Last accessed on: 15/11/2022.

[18] *Random Forest Classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Last accessed on: 15/11/2022.

[19] PUSHKAR MANDOT. *What is LightGBM, How to implement it? How to fine tune the parameters?* https://medium.com/@pushkarmandot/. Last accessed on: 18/12/2022.

[20] *Model Tuning*. https://www.dominodatalab.com/data-science-dictionary/. Last accessed on: 18/12/2022.

[21] *Evaluating ML Models*. https://docs.aws.amazon.com/machine-learning/latest/dg/evaluating_models.html. Last accessed on: 09/11/2022.

[22] JEREMY JORDAN. *Evaluating a machine learning model*. https://www.jeremyjordan.me/evaluating-a-machine-learning-model/. Last accessed on: 09/11/2022.

[23] CARMEN CHAN. *What is a ROC Curve and How to Interpret It*. https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/. Last accessed on: 15/11/2022.