University of Zagreb

Faculty of Organization and Informatics

Marko Jurišić

USER BEHAVIOR ANALYSIS FOR DETECTING COMPROMISED USER ACCOUNTS

- DOCTORAL THESIS -

Thesis advisers:
Associate. prof. Igor Tomičić
Prof. Igor Bernik

Varaždin, 2026.

Sveučilište u Zagrebu

Fakultet organizacije i informatike

Marko Jurišić

ANALIZA KORISNIČKOGA PONAŠANJA ZA DETEKCIJU KOMPROMITIRANIH KORISNIČKIH RAČUNA

- DOKTORSKA DISERTACIJA -

Mentori:
Doc. dr. sc. Igor Tomičić
Prof. dr. sc. Igor Bernik

Varaždin, 2026.

Sažetak

Porast online trgovine donio je i povećanu učestalost online prijevara. Razvoj robustnih sustava strojnog učenja za njihovu detekciju otežan je nedostatkom javno dostupnih skupova podataka, naročito onih fokusiranih na preuzimanje korisničkih računa (*account takeovers*). Ova disertacija adresira taj nedostatak kroz dva primarna cilja: razvoj novog sintetičkog skupa podataka koji sadrži slučajeve preuzimanja korisničkih računa i sustavnu usporedbu metoda i paradigmi strojnog učenja na četiri skupa podataka.

Prvi doprinos rada predstavlja TOMATO, novi skup podataka internetskog oglasnika, koji simulira različite scenarije prijevara. Drugo, evaluirana je učinkovitost naivnog Bayesa, skrivenih Markovljevih modela (*Hidden Markov Models - HMM*), LSTM neuronskih mreža te neuronskih kontroliranih diferencijalnih jednadžbi (*Neural Controlled Differential Equations* - NCDE). Modeli su testirani na skupovima podataka CERT 4.2 i 6.2, novom TOMATO skupu podataka, rezultati su potvrđeni validacijom na skupu anonimiziranih stvarnih korisničkih podataka. Pritom su korištene različite tehnike inženjerstva značajki (*feature engineering*) i različiti intervali grupiranja. Dodatno, pristup treniranja pojedinačnih modela po korisniku je uspoređen s pristupom treniranja globalnih modela na podacima cijele populacije.

Evaluirane su i dokumentirane različite metode vrednovanja (*scoring methods*), pri čemu je pokazano da najbolja metoda strojnog učenja ovisi o kombinaciji temeljnih podataka, procesa izlučivanja značajki (*feature extraction*) i same metode vrednovanja.

Značajno je napomenuti kako rezultati pokazuju da jednostavne metode mogu parirati složenim arhitekturama, ili ih čak nadmašiti, uz pretpostavku opsežnog procesa inženjerstva značajki i korištenja adekvatnih metoda vrednovanja.

Keywords: sigurnost, analiza korisničkog ponašanja, UBA, strojno učenje, naive Bayes, skriveni Markovljevi model, sintetički skup podataka, CRISP-DM, NB, HMM, LSTM, NCDE

Abstract

The increase in online commerce also brought an increased prevalence of online fraud. The development of robust machine learning detection systems is hindered by a lack of publicly available datasets, especially in the account takeover area. This thesis addresses this gap through two primary objectives: the development of a novel synthetic dataset containing account takeovers and a systematic comparison of machine learning methods and paradigms on four datasets.

First, this thesis introduces TOMATO, a new technical online marketplace dataset, simulating various fraud scenarios. Second, the performance of Naive Bayes, Hidden Markov Models, Long Short-Term Memory (LSTM) neural networks, and Neural Controlled Differential Equations (Neural CDEs) was evaluated. These models were tested on the CERT 4.2 and 6.2 datasets, the new TOMATO dataset and a validation set of real-world user data, utilizing various feature engineering techniques and grouping intervals. Additionally, per-user models were compared against global models trained on population-wide data.

Different scoring methods were evaluated and documented, showing that the best machine learning method depends on the combination of the underlying data, feature extraction process and the scoring method itself.

Notably, the results demonstrate that simple methods can perform on par with, or even outperform complex architectures, given a careful feature engineering process and using adequate scoring methods.

Keywords: security, user behavior analysis, UBA, machine learning, Naive Bayes, Hidden Markov Models, synthetic dataset, CRISP-DM, LSTM, NB, HMM, NCDE

Extended Abstract

The main objectives of this study are a) to develop a new synthetic dataset containing account takeovers, and b) to test various machine learning methods on different datasets, considering model performance and resources. To fulfill these goals, the following research questions were identified:

- **RQ1:** Which features have the most impact on detection of malicious/compromised users?
- **RQ2:** Which of the selected machine learning methods (NB, HMM, LSTM, NCDE) produces best results (F1 score, ROC/AUC) on different datasets (CERT 4.2/6.2, new synthetic dataset and real data)?
- **RQ3:** Which hybrid model performs the best for recognizing malicious/compromised users (F1, ROC/AUC)?
- **RQ4:** Which of the analyzed methods is the best considering resources needed for training and deployment?

From research objectives and questions, the following hypotheses are formed:

- **H1:** A subset of all the features allows for a reliable identification of malicious users with error rate within 2% compared to using the entire feature set.
- **H2:** One of the methods analyzed consistently outperforms the other analysed methods in terms of F1 score, ROC/AUC across different datasets.
- **H3:** A hybrid method results in a 5% improvement of F1 score and ROC/AUC compared to any singular method.

The dissertation is structured as follows:

Chapter 1 introduces the goals, research questions and hypotheses, **Chapter 2** defines the approach used - the CRISP-DM, an industry-standard, iterative process for data-centric projects with clearly defined phases and transitions, facilitating experimentation, while **Chapter 3** gives

a literature overview, from user behavior analysis to usages of the selected machine learning methods and CERT dataset to various feature engineering approaches.

Chapter 4 presents the results, with the sections within the chapter mirroring the CRISP-DM process:

- Business Understanding - presenting results of interviews with stakeholders and common attack scenarios
- Data Understanding - introducing the datasets and variants used in the research
- Data Preparation - discussing the various feature engineering approaches used
- Modeling - the central part, discussing architecture and application of mentioned machine learning models
- Evaluation - presenting and comparing the experiment results

Chapter 5 discusses the results and answers the research questions:

- RQ1: the impact of feature engineering, with results showing that the optimal feature set is highly dependent on the model architecture, without clear winner across feature sets or machine learning methods. Thus, H1 is refuted as a general rule.
- RQ2: best performing model across datasets, directly tied to H2. The results indicate that there is no single "best" model, with best model depending on the characteristics of the feature- and dataset being tested. Therefore H2 is formally refuted - there is no universally "best" model, performance is a result of a combination of the model architecture and the characteristics of the data used.
- RQ3: Hybrid model performance, directly tied to H3. Surprisingly good results of NB and LSTM on the TOMATO dataset (F1: 0.99291) made increasing the score by 5% mathematically impossible, making hybrid model training superfluous and refuting H3 as a general rule.
- RQ4: Practicality of deployment and resource consumption. Various experiment constellations show stable performance of Naive Bayes, sometimes being within a few points of much more complex architectures, while training in minutes compared to days when using

neural networks. Naive Bayes emerged as a clear winner, showing that a simple model can outperform complex architectures.

A few methodological discoveries were made: comparison of the results of a per-user model approach with a more practical and real-world applicable approach of training a single global model for all users shows that a properly tuned global model can outperform per-user models, given a strong anomalous signal in the dataset and corresponding data engineering process.

Another important discovery was that also the best scoring method is highly dependent on dataset and model characteristics, with per-user models giving best results when using a self-contained Histogram Entropy method and global model scoring better when measured using standard population-based methods such as Per-User Z-Score, Grubbs Test or Interquartile Range.

As already mentioned, feature engineering process proved to be a crucial step, with weekly data grouping giving the best results on the CERT 4.2 but producing completely unusable results on the TOMATO dataset, which performed the best when using grouping by day.

Chapter 6 concludes the dissertation, summarizes the findings and introduces plans for further research.

Keywords: security, user behavior analysis, uba, machine learning, lstm, naive bayes, hidden markov models, nb, hmm, synthetic dataset, CRISP-DM