УДК 004.855:004.738.5

Д. О. Литвиненко
О. М. КОЗАЧКО

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ФІШИНГОВИХ САЙТІВ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет, Україна

Анотація

Робота присвячена створенню інформаційної технології передбачення фішингових сайтів методами машинного навчання для підвищення рівня кібербезпеки користувачів. Використано методи машинного навчання та аналізу даних у середовищі Kaggle з використанням Python. На основі датасету "Web Page Phishing Dataset" побудовано низку моделей, а саме: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Trees, XGBoost Model, LightGBM Model, ExtraTrees Model та K-Nearest Neighbors Model (KNN). Серед цих моделей обрано оптимальну за балансом між точністю передбачення та швидкістю ідентифікації. Розроблену технологію можна впровадити в браузерні розширення або антивірусне програмне забезпечення для автоматизованого моніторингу та захисту користувачів від фішингових загроз.

Ключові слова: інформаційна технологія, фішинг, машинне навчання, кібербезпека, передбачення.

Abstract

This work is devoted to the creation of information technology for predicting phishing sites using machine learning methods to improve user cybersecurity. Machine learning and data analysis methods were used in the Kaggle environment using Python. Based on the "Web Page Phishing Dataset," a number of models were built, namely: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Trees, XGBoost Model, LightGBM Model, ExtraTrees Model, and K-Nearest Neighbors Model (KNN). Among these models, the one with the optimal balance between prediction accuracy and identification speed was selected. The developed technology can be implemented in browser extensions or antivirus software for automated monitoring and protection of users from phishing threats.

Keywords: information technology, phishing, machine learning, cybersecurity, prediction.

Вступ

У сучасному цифровому середовищі, що постійно ускладнюється та зростає, незважаючи на загальне підвищення рівня обізнаності користувачів про кіберзагрози, зловмисники постійно вдосконалюють методи соціальної інженерії та створюють дедалі переконливіші підроблені веб-ресурси, що суттєво ускладнює їх своєчасне виявлення.

Ключовим фактором, який використовують кіберзлочинці, залишається людський фактор. Недбалість, неуважність або недостатня поінформованість роблять користувачів вразливими до фішингових атак. Навіть найсучасніші системи безпеки не завжди можуть повністю запобігти помилкам користувачів, які стають жертвами шахраїв, розкриваючи конфіденційні дані — паролі, фінансові відомості та особисту інформацію.

Фішинг – це поширена форма інтернет-шахрайства, за якої зловмисники, використовуючи підроблені веб-сайти, електронні листи та повідомлення, обманом змушують жертв розкрити конфіденційну інформацію. З огляду на зростаючу поширеність і складність цих загроз, розробка ефективних методів автоматичного виявлення фішингових сайтів є критично важливою для забезпечення надійного рівня безпеки в Інтернеті [1].

Саме тому в роботі пропонується застосувати методи машинного навчання для створення інформаційної технології, яка здатна ідентифікувати фішингові ресурси з високою точністю, мінімізуючи ризик, спричинений людським фактором.

Метою роботи є поліпшення точності інформаційної технології для виявлення та передбачення фішингових веб-сайтів за допомогою методів машинного навчання шляхом.

Результати дослідження

На першому етапі дослідження було проведено ґрунтовний аналіз сучасних методів передбачення та виявлення фішингових сайтів. Вивчення наукових джерел та існуючих комерційних рішень підтвердило, що найбільш ефективними підходами для протидії фішинговим загрозам, які постійно еволюціонують, є методи машинного навчання. Ці методи забезпечують високу точність класифікації та здатність до

ефективного узагальнення даних, що є необхідним для боротьби з новими, замаскованими фішинговими ресурсами. На основі проведеного аналізу було обґрунтовано вибір ансамблевих моделей та класифікаторів як базових для детального порівняння та подальшої реалізації інформаційної технології [2].

Для реалізації розробки, навчання та тестування моделей було обрано хмарне середовище Kaggle, яке надає необхідні обчислювальні ресурси для швидкого та ефективного проведення експериментів. Програмну частину інформаційної технології було створено з використанням мови Python та потужних бібліотек, таких як Pandas, NumPy, Scikit-learn, XGBoost, LightGBM, що дозволило забезпечити повний цикл робіт: від підготовки даних і інженерії ознак до навчання, оптимізації та валідації прогностичних моделей.

Для навчання моделей був використаний датасет “Web Page Phishing Dataset” [3], доступний на платформі Kaggle. На етапі передоброблення даних та розвідувального аналізу (EDA) було здійснено очищення, нормалізацію та стандартизацію вибірки. Особливу увагу приділено інженерії ознак (FE) та аналізу їхнього впливу на результат [4]. Дослідження показало, що критичний вплив на точність передбачення мають ознаки, пов'язані з архітектурою URL-адреси її довжина, кількість слешів, кількість крапок та інші (рис. 1).

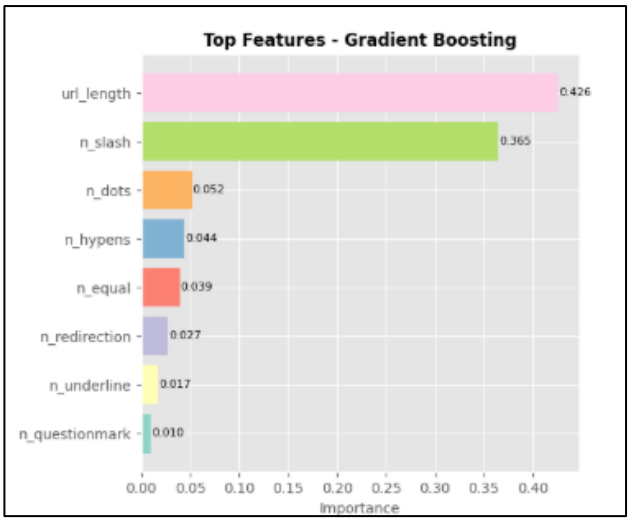

Рисунок 1 – Графік порівняння важливості ознак моделі Gradient Boosting

З метою визначення найбільш оптимального методу для вбудовування в інформаційну технологію було побудовано, навчено та протестовано широку низку моделей машинного навчання. До них увійшли як класифікаційні моделі Logistic Regression, Decision Tree, K-Nearest Neighbors Model (KNN), так і просунуті ансамблеві методи (Random Forest, Gradient Boosting Trees, XGBoost Model, LightGBM Model, ExtraTrees Model) [5]. Продемонстровано результати порівняння ефективності цих моделей на тестовій вибірці (рис. 2).

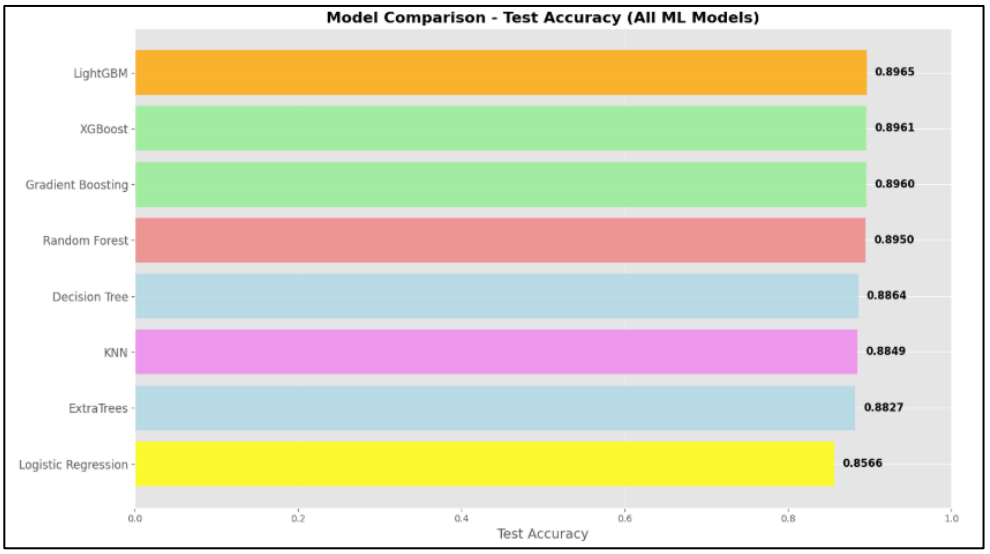

Рисунок 2 – Результати моделей на тестовій вибірці

Окрім загальної точності, також продемонстровано порівняння моделей за F1-Score метрикою, що є важливим показником балансу між точністю (Precision) та повнотою (Recall) для задач виявлення фішингових сайтів (рис. 3).

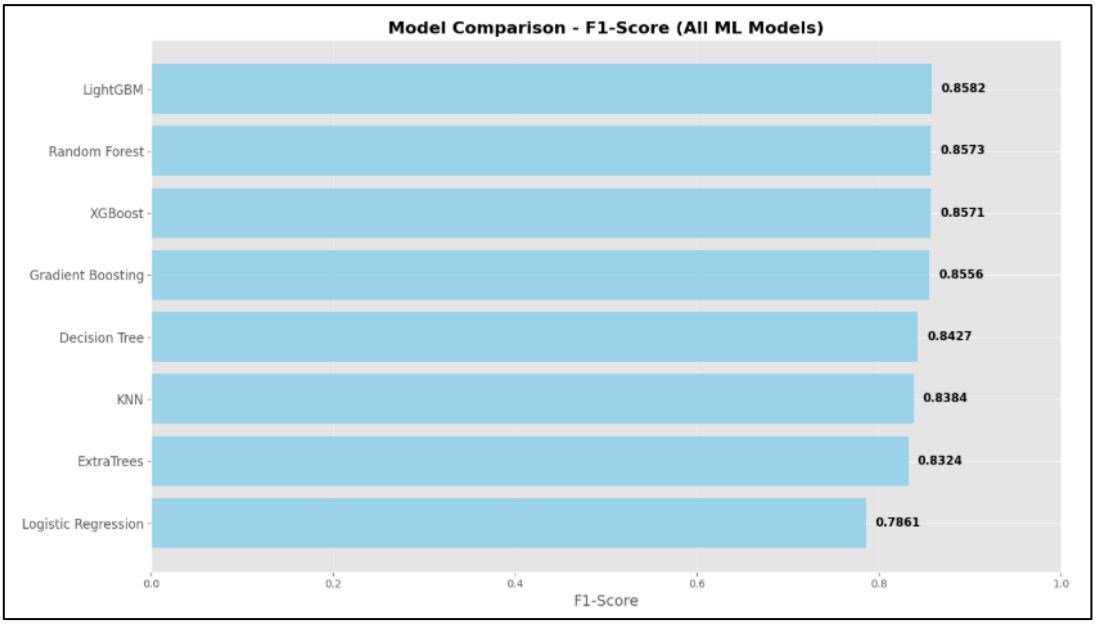

Рисунок 2 – Результати моделей за F1-Score метрикою

Найкращою моделлю виявилась LightGBM. Вона продемонструвала найвищий показник F1-Score (0.8582) серед усіх протестованих моделей. F1-Score є критично важливою метрикою для задач кібербезпеки. Це означає, що модель не тільки коректно ідентифікує фішингові сайти, але й мінімізує кількість хибнопозитивних спрацювань, що є вкрай важливим для практичного використання

Висновки

Таким чином, розроблена інформаційна технологія передбачення фішингових сайтів методами машинного навчання є ефективним інструментом для автоматизації процесів виявлення кіберзагроз та підвищення рівня захисту користувачів. Вона поєднує сучасні досягнення у галузі машинного навчання та аналізу даних, використовуючи низку потужних моделей. Розроблена технологія є стабільною у роботі та має значний науковий і комерційний потенціал для інтеграції у системи кібербезпеки в режимі реального часу.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Гуржій. С.В. Організаційно-технічний аспект протидії фішингу. Науковий журнал з проблем інформаційного права, правової інформатики, інформаційної і національної безпеки, інформації в інших галузях прав. 2023. № 3(46). DOI: https://doi.org/10.37750/2616-6798.2023.3(46).287251
2. Наука про дані: машинне навчання та інтелектуальний аналіз даних: електронний навчальний посібник комбінованого (локального та мережевого) використання [Електронний ресурс] / В. Б. Мокін, М. В. Дратований – Вінниця: ВНТУ, 2024. – 258 с. – Режим доступу: https://docs.vntu.edu.ua/card.php?id=8163
3. FERNANDO DANIEL. Web Page Phishing Dataset. 2024. [Електронний ресурс] – Режим доступу: https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset/data?select=web-page-phishing.csv
4. Литвиненко Д. Web Phishing Analysis. 2025. [Електронний ресурс] – Режим доступу: https://www.kaggle.com/code/linean/web-phishing-analysis
5. Alnemari S, Alshammari M. Detecting Phishing Domains Using Machine Learning. Applied Sciences. Applied Sciences. 2023. Vol. 13 (8). P. 4649. DOI: https://doi.org/10.3390/app13084649

Литвиненко Данило Олександрович – студент групи 2ICT-24м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: lytdanya@gmail.com

Науковий керівник: ***Козачко Олексій Миколайович*** – доц. кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця

Lytvynenko Danylo O. – student of group 2IST-24m, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: lytdanya@gmail.com

Scientific advisor: ***Kozachko Oleksiy M.*** – Associate Professor, Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia