SHAP-аналіз для підвищення прозорості моделей штучного інтелекту в задачах кібербезпеки

УДК 004.8:004.056

Тетяна Бажан[1], Світлана Поперешняк[2]

Державний університет інформаційно-комунікаційних технологій,
[1]t.bazhan@duikt.edu.ua, [2] spopereshnyak@gmail.com

Сучасні системи кібербезпеки дедалі частіше використовують методи машинного навчання для виявлення аномалій, прогнозування інцидентів інформаційної безпеки, оцінювання кіберризиків та підтримки рішень у процесах реагування на загрози. Водночас високоточні моделі, зокрема ансамблеві алгоритми та нейронні мережі, часто функціонують як «чорна скринька», що ускладнює пояснення причин формування прогнозу або класифікації події як потенційно небезпечної. Для сфер, пов'язаних із захистом інформації, така проблема є особливо важливою, оскільки результати роботи моделі мають бути не лише точними, а й зрозумілими для фахівців з кібербезпеки, аналітиків SOC/CERT/CSIRT та осіб, які приймають управлінські рішення.

Актуальність дослідження зумовлена активним використанням моделей машинного навчання у сфері кібербезпеки для виявлення загроз, аналізу аномалій та прогнозування інцидентів інформаційної безпеки.

Метою дослідження є підвищення прозорості моделей штучного інтелекту в задачах кібербезпеки шляхом використання SHAP-аналізу для інтерпретації впливу ознак на результати прогнозування, виявлення аномалій та оцінювання ризиків інформаційної безпеки.

Наукова новизна дослідження полягає у застосуванні SHAP-аналізу для підвищення прозорості та інтерпретованості моделей машинного навчання у задачах кібербезпеки, що дозволяє не лише прогнозувати кіберзагрози та аномалії, а й визначати ступінь впливу окремих факторів на результати моделювання.

Для розв'язання поставленої задачі використано підхід, який дозволяє оцінити внесок кожної ознаки у результат прогнозування кіберзагроз незалежно від типу моделі. Загальна схема застосування SHAP-аналізу в задачах прогнозування представлена на рис. 1. Метод базується на теорії кооперативних ігор та забезпечує пояснення роботи моделей машинного навчання незалежно від їх архітектури [1]. SHAP дозволяє виконувати як глобальну інтерпретацію моделі, визначаючи загальний вплив факторів на результат прогнозування, так і локальну інтерпретацію окремих прогнозних значень [1, 2]. Це дає можливість виявляти ключові параметри, які найбільше впливають на виявлення кіберінцидентів та аномалій у системах інформаційної безпеки, а також аналізувати напрям і силу їх впливу.

Значення SHAP для окремої ознаки визначається як середній внесок цієї ознаки у всі можливі комбінації ознак моделі:

$$\phi_i(f,x) = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(-1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right], \quad (1)$$

де ϕ_i – SHAP-значення ознаки, F – множина всіх ознак, S– підмножина ознак, $f(x)$– прогноз моделі.

Використання SHAP-аналізу дозволяє підвищити рівень інтерпретованості моделей машинного навчання у задачах кібербезпеки, забезпечити прозорість прийняття рішень та покращити контроль за процесом прогнозування кіберзагроз і виявлення аномалій. Крім того, застосування підходів Explainable Artificial Intelligence сприяє підвищенню надійності систем інформаційної безпеки, забезпечує можливість пояснення результатів роботи моделей та підвищує обґрунтованість аналітичних висновків [2, 3].

Отримані результати підтверджують доцільність використання SHAP для аналізу складних моделей машинного навчання у задачах виявлення кіберінцидентів, прогнозування загроз та підтримки прийняття рішень у сфері захисту інформації [4].

Рис. 1. Блок-схема застосування SHAP-аналізу в задачах прогнозування

Таким чином, використання даного підходу дозволяє підвищити довіру до інтелектуальних систем, покращити інтерпретованість результатів та сприяти більш ефективному використанню технологій штучного інтелекту у практичних задачах інформаційної безпеки.

1. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017. Vol. 30. p. 4765-4774.
2. Molnar C. Interpretable Machine Learning. 2nd ed. Munich : Christoph Molnar, 2022. 318 p.
3. Arrieta A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020. Vol. 58. p. 82-115.
4. Sarker I. H. AI-based cybersecurity: A comprehensive overview, taxonomy and challenges. Journal of Cybersecurity and Privacy. 2021. Vol. 1(2). p. 154-181.