КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2(14), 2021
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2021.14.5067
УДК 004.056

Цирканюк Діана Андріївна
магістрантка кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський університет імені Бориса Грінченка, Київ, Україна
ORCID ID: 0000-0002-9422-8617
d.tsyrkaniuk@gmail.com

Соколов Володимир Юрійович
к.т.н., доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський університет імені Бориса Грінченка, Київ, Україна
ORCID ID: 0000-0002-9349-7946
v.sokolov@kubg.edu.ua

Мазур Наталія Петрівна
к.п.н., доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський університет імені Бориса Грінченка, Київ, Україна
ORCID ID: 0000-0001-7671-8287
n.mazur@kubg.edu.ua

Козачок Валерій Анатолійович
к.т.н., доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський університет імені Бориса Грінченка, Київ, Україна
ORCID ID: 0000-0003-0072-2567
v.kozachok@kubg.edu.ua

Астапеня Володимир Михайлович
к.т.н., доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський університет імені Бориса Грінченка, Київ, Україна
ORCID ID: 0000-0003-0124-216X
v.astapenia@kubg.edu.ua

МЕТОД ПОБУДОВИ ПРОФІЛІВ КОРИСТУВАЧА МАРКЕТПЛЕЙСУ І ЗЛОВМИСНИКА

Анотація. Кількість і складність кіберзлочинів постійно зростає. З'являються нові різновиди атак і конкурентної боротьби. Кількість систем зростає швидше, ніж навчаються нові спеціалісти з кібербезпеки, тому все складніше стає відслідковувати вручну в режимі реального часу дії користувачів. Особливо активно розвивається електронна торгівля. Не всі ретейлери мають достатній ресурс для підтримки власних інтернет-крамниць, тому вони вимушені співпрацювати з посередниками. Роль посередників все частіше виконують спеціальні торгівельні площадки зі своїми електронними каталогами (вітринами), сервісами оплати і логістики, контролем якості – маркетплейси. У статті розглянута проблема захисту персональних даних користувачів маркетплейсу. Метою статті є розробка математичної моделі поведінки для підвищення захисту персональних даних користувача для протидії фроду (антифроду). Профілювання може бути побудоване за двома напрямками: профілювання легітимного користувача і зловмисника (питання прибутковості та скорінгу виходять за межі даного дослідження). Профілювання користувача побудоване на типовій поведінці, сумах і кількості товарів, швидкості наповнення електронного візочка, кількість відмов і повернень тощо Досліджено основні алгоритми побудови поведінкового профілю користувачів та застосовано метод виявлення порушника шляхом порівняння його дій з діями середньостатистичного користувача. Запропоновано власну модель профілювання поведінки

ISSN 2663 - 4023

користувачів на основі мови програмування Python та бібліотеки Scikit-learn методом випадкового лісу, лінійної регресії й дерева рішень, використано метрику застосовуючи матрицю помилок, проведено оцінку алгоритмів. У результаті порівняння оцінки даних алгоритмів трьох методів, метод лінійної регресії показав найкращі результати: A – 98,60%, P – 0,01%, R – 0,54%, F – 0,33%. Правильно визначено 2% порушників, що відповідно позитивно впливає на захист персональних даних.

Ключові слова: маркетплейс, профіль користувача, модель користувача, дерево рішень, профілювання поведінки.

ВСТУП

Розслідування кіберзлочинів набуває все більше актуальності, оскільки останнім часом спостерігається більше зростання випадків кіберзлочинів. Згідно останніх даних платформи даних Opendatabot, в останні декілька років в Україні збільшилась кількість інформаційних злочинів у 2,5 рази [1]. В статті розглянута проблема забезпечення доступності та цілісності безпроводових абонентів у стільникових та інших безпроводових корпоративних мережах. Маркетплейс – це сервіс який надає можливість різним продавцям укладати акт купівлі-продажу товарів та послуг. Перевагою маркетплейсу в порівняні з онлайн магазином є те, що користувачі можуть порівнювати між собою один товар у різних продавців.

Метою статті є підвищення захисту персональних даних користувача та протидія фроду (антифрод), шляхом розробки математичної моделі профілювання поведінки. Для досягнення поставленої мети, у статті: досліджено існуючі методи, проведено порівняння алгоритмів профілювання користувачів, розроблено власну модель за допомогою методів лінійної регресії та дерева рішень.

Запропонована модель може бути використана для побудови системи профілювання поведінки користувачів з метою підвищення рівня захищеності персональних даних користувачів будь-якої організації чи компанії.

ОГЛЯД ОСНОВНИХ МЕТОДІВ ПРОФІЛЮВАННЯ КІБЕРЗЛОЧИНІВ

Профайлінг – технологія складання психологічних портретів, що є інструментом розслідування інформаційних злочинів та дає можливість класифікувати кіберзлочинців, встановити їх мотиви, навички. Профілювання використовується як інструмент для розслідування кіберзлочинів. Методика з'явилася у 70-ті роки як розробка ізраїльської армії: військові використали його, щоб обчислити у групі людей потенційних терористів.

Кримінальне профілювання є ключовим інструментом, який використовується для звуження підозрюваних та оцінки ймовірності вчинення підозрюваним злочину. Кримінальне профілювання – це науковий прийом для оцінки та аналізу злочину та визначення поведінкових характеристик особи, яка вчинила злочин [2]. Профіль складається з набору характеристик, що притаманні особам, які вчинили певний вид злочину [3]. Профілювання кіберзлочинців є ефективним, коли при розробці профілів використовується стандартна методологія, яка ґрунтується на обґрунтованих припущеннях. Переважають два типи кримінального профілювання: індуктивне та дедуктивне. Діагностика, яка дозволяє у відносно короткий термін визначити психотип людини, її базові цінності та риси характеру.

Метод індуктивного профілювання (рис. 1) передбачає використання бази даних (БД), яка містить інформацію про порушників. Аналізуючи отримані дані потрібно встановити співвідношення та визначити характеристики притаманні для статистично великої кількості порушників, що вчинили конкретний вид злочину [3].

Рис. 1. Методика індуктивного профілювання

Метод індуктивного профілювання використовує статистичний аналіз, методи аналізу даних для розробки моделей для виявлення шаблонів і включає в себе перевірку даних для визначення моделей, які відповідають відомим профілям шахрайства [4]. Збирається інформація про кіберзлочинців, які вчинили конкретний вид злочину. Ця інформація надходить з офіційних розслідувань, спостережень за поведінкою відомих злочинців, клінічних та інших інтерв'ю із злочинцями, даних, що є у різних БД. Шляхом обробки отриманих даних та пошуку кореляцій досягаються характеристики, які вважаються загальними для статистично достатньої кількості злочинців, які вчинили певний вид злочину [5].

Застосування індуктивного методу призводить до отримання характеристик, які є конкретними, а скоріш узагальненими. У свою чергу дедуктивний метод профілювання (рис. 2) передбачає аналіз криміналістичних доказів та профілювання жертви для визначення її мотиву, характеристик зловмисника. Аналізує криміналістичні докази, використовує принципи віктимології та свій досвід для визначення кримінальних характеристик [4].

Рис. 2. Методика дедуктивного профілювання

Кримінальний профіль кіберзлочинця включає набір даних індуктивного і дедуктивного профілів злочинця.

Індуктивний профіль кіберзлочинця поєднує:

– збір статистичних даних, пов'язаних із певними поведінковими моделями;

– демографічні характеристики злочинця.

Дедуктивний профіль кіберзлочинця включає наступний набір даних

– сукупність отриманих доказів;

– докази, виявлені на місці злочину та сліди;

– віктимологічні аспекти; опис особи злочинця [6].

Індукційний метод починається з моніторингу та збору інформації, на основі якої реалізується теоретична модель, яка застосовується на практиці. Дедуктивний метод починається з припущення, яке конкретизується в конкретні характеристики (рис. 3).

Рис. 3. Відмінності між індуктивним та дедуктивним методами профілювання

Використовувати лише на індуктивне профілювання не достатньо для ефективного профілювання. З іншого боку, покладаючись лише на дедуктивне профілювання, дослідники не звертають уваги на поточні тенденції, такі як популярні методи атак, ймовірні цілі та жертви [4]. Тому для профілювання необхідно використовувати гібридну методологію.

ПРОФІЛЮВАННЯ КІБЕРЗЛОЧИНЦЯ

Методологія профілювання кіберзлочинців, як показано на рис. 4, має бути інтерактивним процесом для точності та ефективності. Дослідження мають повторний характер. Попередні розслідування виявляють основні деталі. У міру того, як розслідування переходить від базової стадії до просунутої, збирається більше інформації, яка може допомогти у виявленні додаткових мотивів, які не були врахуванні.

Рис. 4. Методологія опису профілю кіберзлочинця

Злочинне профілювання – це мистецтво і наука розробки опису характеристик злочинця (фізичних, інтелектуальних та емоційних) на основі інформації зібраної на місці злочину. Завдяки порівнянню поведінки кабірезлочинця з поведінкою валідного користувача з поведінкою маркетплейсу побачивши аномалії в діях можна виявити кіберзлочинця.

При складанні кримінального профілю кіберзлочинця виявляються специфічні загальні риси та ознаки. Слід погодитися з тим, що профілювання є більш ефективним у разі кількох, а не одного злочину [7]. Традиційно виділяють три підходи до кримінального профілювання:

– підхід до розслідування справ;

– підхід клінічного лікаря;

– науковий статистичний підхід [8].

Кожен із цих підходів має розбіжності у сфері знань, основі яких передбачається пояснювати злочинне поведінка.

Одним із напрямків профілювання є психологічна оцінка певних властивостей. Профіль кіберзлочинця поєднує у собі опис поведінки та якостей людини, що створюється без знання особистості злочинця.

Кримінальне профілювання передбачає ідентифікацію невідомого злочинця за допомогою кількох методик:

– аналіз місця злочину;

– визначення особливостей кримінального правопорушення;

– характеристика особи злочинця [9].

Профіль кіберзлочинця можна описати, включивши такі ключові елементи, як:

1. Характеристики особистості, які притаманні конкретній людині, і які схиляють особу до кіберзлочину. Риси особистості визначаються як широкий індивідуально-психологічний профіль, який описує міжособистісні та загальні індивідуальні відмінності поведінки і почуттів. Риси проявляються в індивідуальній діяльності в різних ситуаціях та час [10].

2. Злочинний професіоналізм – це риси особистості, які сприяють безпечному та ефективному вчиненню кіберзлочинності. Він включає чотири обов'язкові ознаки: специфічні особистісні якості; знання та вміння; безстрашність, сміливість і впевненість у собі; ефективність і життєздатність дій; вчинення кримінального правопорушення та досягнення певної мети [11].

3. Технічні знання, пов'язані зі спеціальними знаннями та технічними навичками роботи зі складними програмами та пристроями, що дозволяють здійснювати кіберзлочинність [11].

4. Соціальними характеристиками є демографічні особливості, соціально-економічний статус, соціально-психологічні та моральні якості. Основними елементами є стать, вік, національність, соціально-економічний статус.

5. Характеристика мотивації. У кримінології під мотивацією розуміють сукупність мотивів дій, у яких кожен із мотивів визначає елемент мотивації і існує як у свідомості, так і в підсвідомості. Мотиви розвиваються і формуються під впливом людських емоцій і почуттів. Мотиви бувають внутрішні – вибрані людиною та зовнішні – керовані іншими [12].

Мотив – це провідна і сприяюча функція діяльності (внутрішнє психічне заохочення), яка, створюючи предмет діяльності, спрямовує діяльність людини [11].

Більшість кіберзлочинів за своєю природою є серійними, оскільки злочинець привикає до своєї поведінки і здійснює безліч правопорушень. З цього можна зробити висновок про характер та способи дії.

Хоча в реальному житті це не так просто, кримінальне профілювання є важливим інструментом, який може дати розслідуванням багато підказок щодо особи, яка вчиняє конкретний злочин або серію злочинів. Важливо розуміти, що навіть детально описані

ISSN 2663 - 4023

профілі дають лише загальний опис про особу, яка вчинила злочин. Профіль не вказує на контрактного підозрюваного [1].

МОДЕЛЮВАННЯ ПОВЕДІНКИ КОРИСТУВАЧА МАРКЕТПЛЕЙСУ

Завдяки порівнянню поведінки порушника з поведінкою середньостатистичного користувача маркетплейсу можна виявити зловмисника. Тому формування поведінкового профілю є найважливішою частино моделі користувача. Всі дії користувача можна класифікувати та сформувати загальну модель поведінки з попередньо визначеною послідовністю дій.

Кожен користувач маркетплейсу зважаючи на свою мету, формує своє замовлення, що займає певний час. Під час здійснення вибору, пошуку товару, введення своїх персональних даних, виконуючи всі дії формується його модель поведінки. Аналізуючи дії, що виконує користувач при формуванні замовлення, можемо зробити висновок, що більшість користувачів має однаковий алгоритм дій (рис. 5).

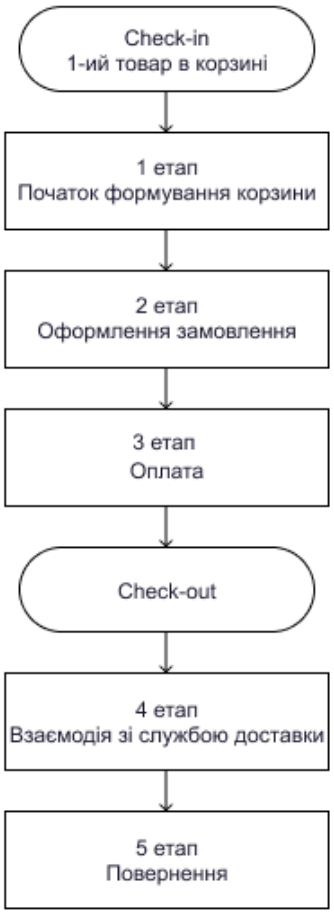

Рис 5. Алгоритм дій користувачів маркетплейсу

Звідси робимо висновок про те що поведінка користувача є нестандартною (рис. 6), якщо ми бачимо, що за ним не спостерігається загальна послідовність дій та час витрачений на них. Тому розглянемо кожен етап більш детально для формування профілю нормальної поведінки користувача.

Рис. 6. Модель поведінки користувача

Про користувача, що здійснив оформлення замовлення можна сформувати модель, що складається з двох профілів:

– загальна інформація про користувача маркетплейсу;

– опис поведінки користувача.

Для кожного користувача, що має аномальну чи злочину поведінку, можна узагальнити типові маркери дій.

До загальної інформації про користувача, можна віднеси інформацію, яка не змінюється з часом і є унікальною для мережі (рис. 7 і табл. 1):

– ідентифікатор замовлення;

– ПІБ замовника;

– ПІБ отримувача;

– індикатори в мережі: поштова адреса, обліковий запис, номер телефону;

– адреса доставки.

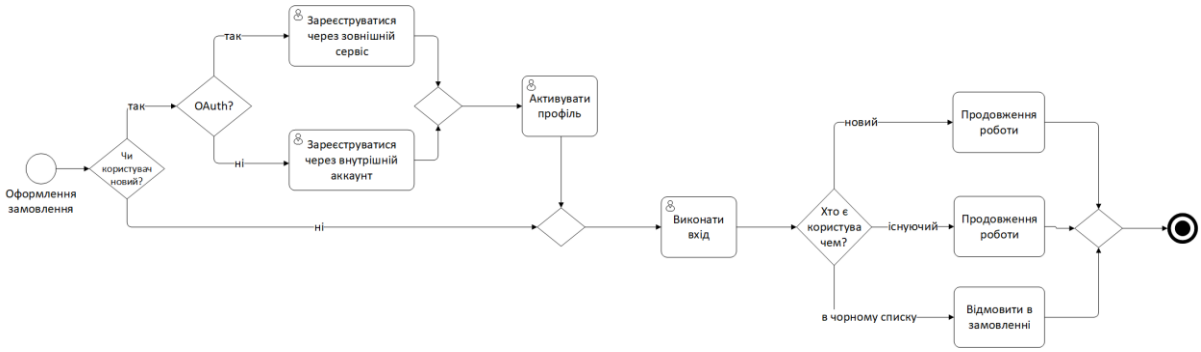

Рис. 7. Оформлення замовлення

Поведінка користувача:

– браузер, що використовує користувач;

– метод оплати (рис. 8);

– канал формування замовлення (R3/2);

– підтвердження замовлення;

– час на формування замовлення (R3/4);

– дані про причину відмови від замовлення.

ISSN 2663 - 4023

Таблиця 1

Ролі користувачів та опис поведінки

Індекс ролі	Роль	Типовий опис	Індекс маркеру	Маркери
R1	Неадекватний користувач	Користувач або бот, який свідомо виконує велику кількість компульсивних та нелогічних операцій з метою завдати шкоди	R1/1	Формує замовлення не сезонних товарів у великій кількості
			R2/2	Формує замовлення товарів з різних категорій, не пов'язаних між собою
R2	Порушник	Особа, яка має на меті здійснити злочині дії (в т.ч. і для свої вигоди)	R2/1	Неуспішна оплата товару
R3	Фродер	Особ, яка не свідомо здійснює набір певних операцій при формуванні замовлення, логічну послідовність яких не може бути встановлено	R3/1	Висока активність вночі.
			R3/2	Велика кількість кліків з одного IP чи ID.
			R3/3	Різне місце замовлення в одного користувача.
			R3/4	Мінімальний час між початком формування замовлення і його виконанням.
			R3/5	Шаблона поведінка, однаково витрачений час підчас переходів на сайті.

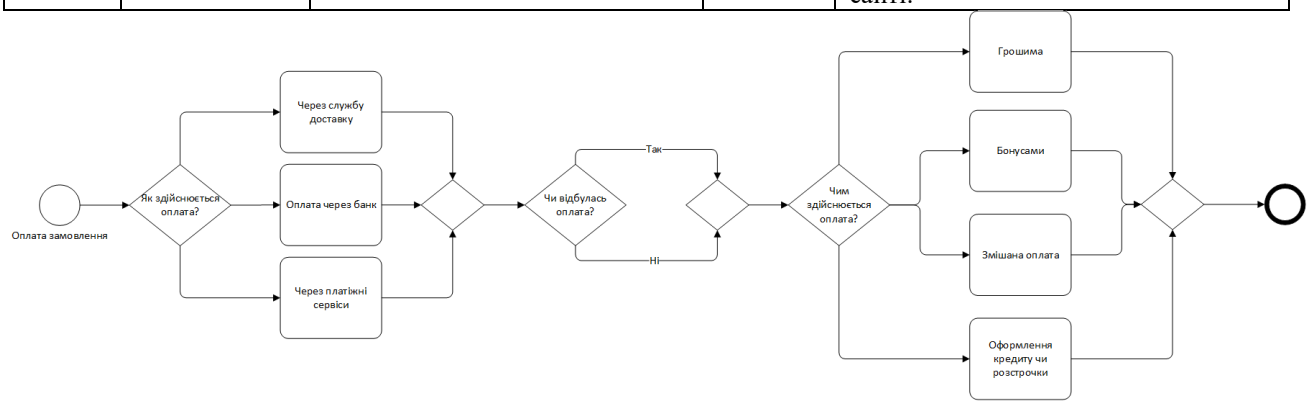

Рис. 8. Оплата замовлення

Наприклад, підчас взаємодії зі службою доставки. У разі відмови аналізуючи час витрачений на доставку, відгуки про продавця, якість товару, можна розрахувати потенційну ймовірність відмови та сформувати параметри за яких виконання замовлення є вдалим (рис. 9).

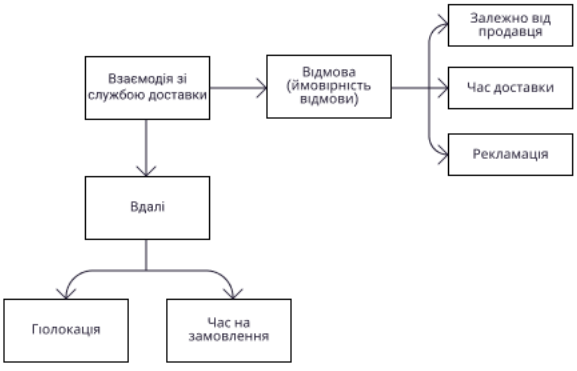

Рис. 9. Взаємодія користувача зі службою доставки

ISSN 2663 - 4023

На етапі оплати замовлення, наприклад у випадку коли оплата не відбулася (R2/1), робимо висновок про аномалію в поведінці користувача.

Кожен з етапів побудови моделі профілю користувача має важливе значення при проведені аналізу поведінки користувачів маркетплейсу:

– за допомогою загальної інформації можливо однозначно прив'язати дії до певного користувача і розглядати його окремо;

– аналізуючи поведінку користувача можна вдало порівнювати дії, що виконує користувач вже не вперше.

Формування загального профілю користувача, майже не викликає складності, у свою чергу, побудова поведінкового профілю вимагає використання різних методів класифікації.

МЕТОД ВИПАДКОВОГО ЛІСУ ТА ДЕРЕВО РІШЕНЬ

Метод «випадкового лісу» є непаратичним контрольованим методом навчання, який використовується для класифікації та регресії. Мета полягає в тому, щоб створити модель, яка передбачає значення цільової змінної, вивчаючи прості правила прийняття рішень, виведені з функцій даних.

Найефективніший метод класифікації для великої кількості даних, адже використовує одночасно декілька різних дерев рішень.

Недоліком дерева рішень є алгоритм класифікації. Вершиною дерева визначається елемент, відповідно до якого дані розділяться на групи. Наприклад, у найпростішому вигляді дерево рішень – це спосіб уявлення правил в ієрархічній, послідовній структурі [13]. Основа такої структури – відповіді True чи False на низку питань, як на рис. 10.

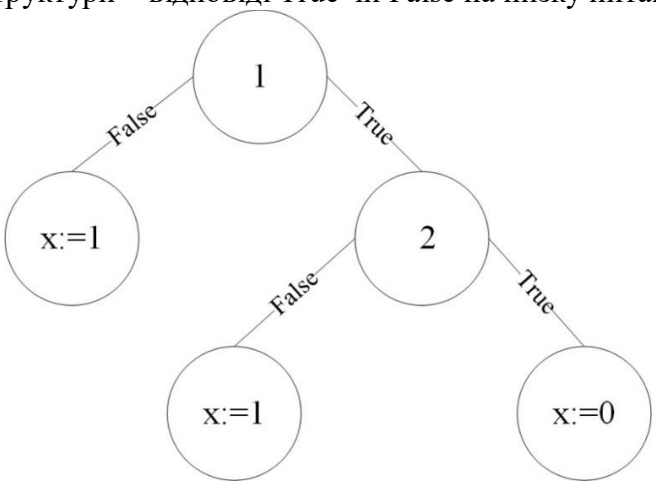

Рис. 10. Бінарне дерево рішень

Бінарні дерева є найпростішим, окремим випадком дерев рішень. В інших випадках відповідей і, відповідно, гілок дерева, що виходять з його внутрішнього вузла, може бути більше двох.

Нехай дана вибірка даних декількох користувачів описаний n–ю кількість характеристик $x_1, x_2, \dots, x_n$. Ці користувачі належать до уявних двох груп, де перша група нормальні користувачі та друга група порушники. Основною метою є можливість розрізнити нормальних користувачів від порушників (рис. 10).

ISSN 2663 - 4023

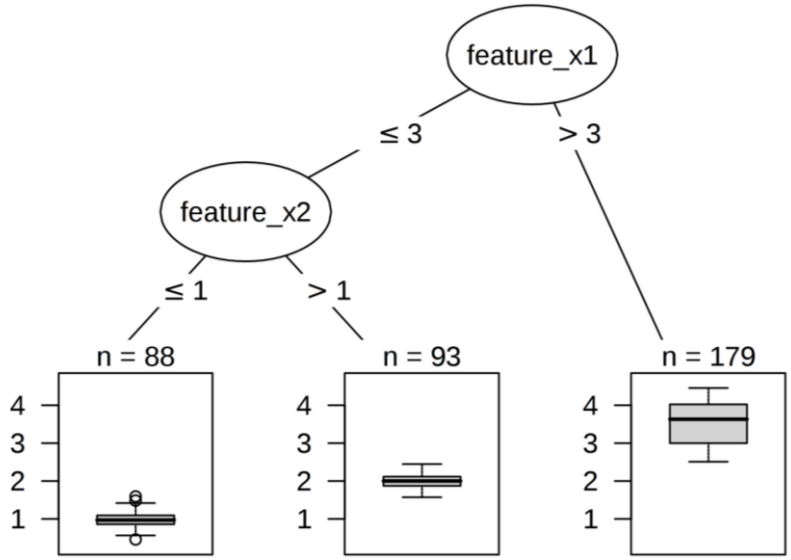

Рис. 11. Приклад побудови дерева рішень

Така модель не є результативною, оскільки її застосування можливе лише на одному прикладі даних та тому, що їх дуже складно інтерпретувати.

Незважаючи на переваги даного методу, слід пам'ятати, що для того, щоб побудувати якісну модель, необхідно розуміти природу взаємозв'язку між залежними та незалежними змінними та підготувати достатній набір даних.

Переваги дерев-рішень:

– простий для розуміння і інтерпретації;

– можна візуалізувати;

– здатний обробляти як числові, так і категоричні дані;

– можлива перевірка моделі за допомогою статистичних тестів. Це дає можливість враховувати надійність моделі;

– працює, навіть якщо припущення порушуються істинною моделлю, з якої були отримані дані.

Недоліки дерев-рішень:

– дерева рішень можуть бути нестабільними, оскільки невеликі зміни в даних можуть призвести до створення зовсім іншого дерева;

– створюють необ’єктивні дерева рішень, якщо деякі класи домінують. Тому рекомендується збалансувати набір даних перед тим, як відповідати дереву рішень.

ОЦІНКА ЯКОСТІ МОДЕЛІ КІБЕРЗЛОЧИНЦЯ

Основне завдання, при побудові моделі поведінки користувача – визначення методу. За допомогою порівняння поведінки користувача з поведінкою зловмисника можна виявити порушника. Для цього достатньо класифікувати стандартну поведінку для розпізнання відхилень. Оцінка якості потрібна, щоб мати представлення про том, наскільки добре працює отриманий алгоритм.

Для перевірки коректності роботи алгоритму, потрібно провести оцінку якості [14]. Для цього будемо застосовувати метрику:

– TP (True Positive) – звичайна поведінка користувача, вірно класифіковано;

– FP (False Positive) – аномальна поведінка порушника;

– FN (False Negative) – звичайна поведінка порушника;
– TN (True Negative) – аномальна поведінка користувача.
При цьому матриця метрики матиме вигляд наведений в табл. 2.

Таблиця 2

Позначення результатів класифікатора по відношенню до істинних значень

		Значення	
		Позитивне	Негативне
Результат класифікації	Позитивне	TP	FP
	Негативне	FN	TN

З метою оцінювання даної метрики введемо та розрахуємо наступні показники:

1. Достовірність (accuracy) – кількість коректних відповідей алгоритму класифікації поведінки

$$A = \frac{TN + TP}{FN + FP + TN + TP}. \quad (1)$$

2. Точність (precision) – показує відсоток правильний відповідей від числа вірно класифікованих

$$P = \frac{TP}{FP + TP}. \quad (2)$$

3. Повнота (recall) – відображає який відсоток звичайної поведінки виявлено алгоритмом в порівняні із загальною кількістю

$$R = \frac{TP}{FN + TP}. \quad (3)$$

4. Адекватність оцінки – розраховується як гармонійне середнє між повнотою і точністю

$$F = 2\frac{\mathrm{R} \cdot \mathrm{P}}{\mathrm{R} + P}. \quad (4)$$

В сучасних технологіях машинного навчання, основною проблемою застосування алгоритмів є дані, на яких вони застосовуються і тому саме процес обробки даних є вирішальними для подальшого аналізу. В реальному аналізі даних першим і основним етапом є збір даних. Другий, найважливіший етап, це етап відбору даних, які мають найбільший вплив на результат, який ми хочемо передбачити. Третій етап – чистка даних, наприклад відбір аномальних значень, аналіз неправильно присвоєних значень. Кожний з цих етапів має прямий вплив на якість результату сформованих профілів.

Існує ще один важливий фактор, який впливає на якість збору, обробки та зберігання даних – людський фактор. Важливо звертати увагу на те, що дані мають бути оброблені за математичними «правилами», тобто об’єктивно, а не суб’єктивно – людиною. Таким чином відсутність інформації про процес обробки даних не дає змоги одразу оцінити їх якість. Існує два основні способи покращення профілювання за допомогою аналізу вхідних даних. Кожна БД складається з характеристик або ознак, тобто категорій даних, наприклад «місто», «тип оплати» та «спосіб доставки».

Алгоритми в R та Python можуть прогнозувати різні результати для одної моделі. Тому краще в Python визначати маркування, а випробувати моделі та перевірку якості даних проводити в R [15, 16].

ОБҐРУНТУВАННЯ ВИБОРУ ВИХІДНИХ ДАНИХ ДЛЯ КЛАСИФІКАЦІЇ

Вхідними даними для застосування моделі обрано БД акаунтів користувачів маркетплейсу. В структуру БД входить множина акаунтів користувачів та множина акаунтів порушників.

Отримані дані було відфільтровано та вилучені некоректні. Проаналізовано 173 895 записів БД користувачів з них після фільтрування залишилось 2 755, що складає 1,5%.

Для аналізу та профілювання потрібно підготувати дані, оскільки від якості даних залежить якість застосованих моделей. Цей процес передбачає перетворення даних у бажану і придатну форму для подальшого використання. Обробка даних та розрахунки будуть виконуватися на основі мови програмування Python та бібліотеки Scikit-learn.

Варто звернути увагу на неперервні та категоріальні дані, що містяться в нашій БД. До неперервних даних відносяться ті що можуть приймати будь-яке значення, тобто вони мають бути дискретизовані, що є важливим для побудови моделі. Потреба в дискретизації даних, полягає в тому, що неперервні змінні можуть мати занадто зміщену асиметрію, мультимодальний розподіл, а це суперечить припущенням моделей та впливає на якість профілю.

До категоріальних змінних відносяться ті, що набувають певну обмежену кількість значень, наприклад «спосіб оплати», «тип замовлення», «метод доставки» тощо. Для реалізації запропонованих моделей встановлюємо наступні етапи проведення експерименту:

– на першому етапі попередньо відфільтровано дані;

– на другому етапі формуємо тестову множину, звантажуємо необхідні бібліотеки та залежності;

– на третьому етапі введемо метрику оцінювання якості результатів класифікації;

– на четвертому етапі перевіримо припущення про репрезентативність тестової множини.

А саме, використаємо метрику застосовуючи матрицю помилок. Припустимо, що у нас є два класи та алгоритм, що передбачає відповідність кожного об'єкта одному з класів, тоді матриця помилок класифікації буде виглядати.

Виходячи з основного завдання систем аналізу поведінки попередньо зробимо висновок, що на третьому етапі ми отримаємо два варіанти поведінки користувача

– звичайна поведінка;

– потенційно небезпечна.

В сучасних технологіях машинного навчання, основною проблемою застосування алгоритмів є дані, на яких вони застосовуються і тому саме процес обробки даних є вирішальними для подальшого аналізу. Кожна БД складається з характеристик або ознак, тобто категорій вхідних даних, так наприклад «місто», «тип оплати» та «спосіб доставки» є характеристиками даних, що розглядаються в цій роботі. Існує два основні методи покращення моделі за допомогою аналізу вхідних даних.

ТЕХНОЛОГІЯ ПОБУДОВИ МОДЕЛІ

Відфільтровуємо дані та виконаємо заміну для категоріальних змінних формату «True» та «False» на булевий тип даних 1 та 0 відповідно.

ISSN 2663 - 4023

Для реалізації моделі будемо використовувати набір даних маркетплейсу про користувачів та їх замовлення, а саме: повна вартість замовлення (без доставки), повна вартість замовлення (з доставкою), відмітка крупногабаритного товару, відмітка тестового замовлення, відмітка необхідності встановлення техніки, відмітка підтвердження замолення оператора, відмітка запиту дзвінку, відмітка перевірки адреси доставки, відмітка замовлення в «один клік» (замовлення швидко оформлюється з налаштуваннями за замовчуванням).

Вводимо метрику оцінювання якості результатів класифікації (рис. 12).

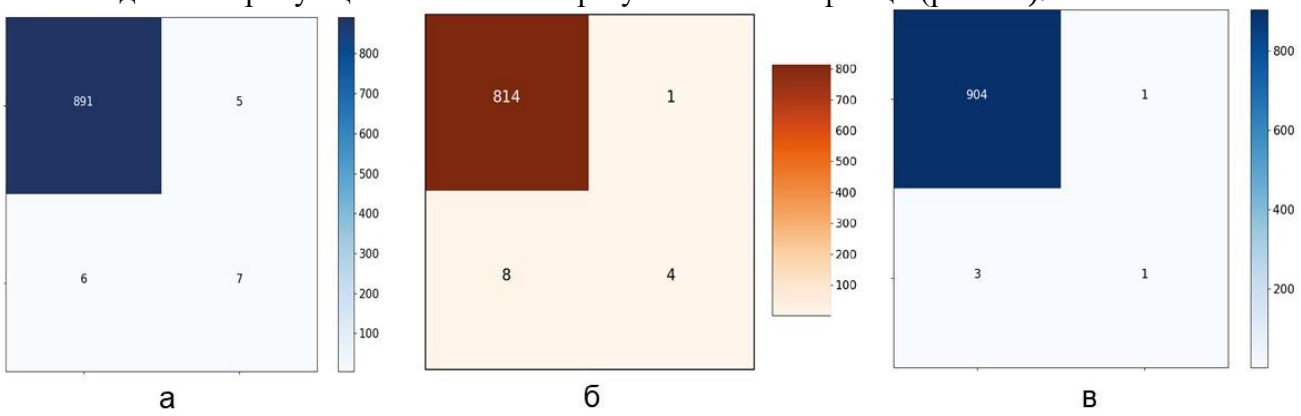

Рис. 12. Матриця плутаниці лінійної регресії (а)

плутаниці методу випадкового лісу (б); плутаниці методу дерева рішень (в)

Порівняння методів наведена у табл. 3.

Таблиця 3

Порівняння результатів застосованих моделей

Метод	Показник				
	A	*E*	*P*	*R*	*F*
Логічна регресія	0,9868	0,0132	0,5455	0,4615	0,3330
Випадковий ліс	0,9956	0,0044	0,2500	0,5000	0,1429
Дерево рішень	0,9855	0,0145	0,1111	0,2000	0,3330

Показник точність приймає однаково хороший показник для кожної з моделей, що вказує на високу якість вхідних даних. Для аналізу даних розглядається декілька моделей, проводиться їх якісний аналіз, за показниками та обирається одна модель.

Проведемо додатковий аналіз якості моделі методу випадкового лісу за допомогою візуалізації кривої концентрації (рис. 13а). З кривої концентрації можна визначити, який відсоток нормальних користувачів припадає на який відсоток даних. Для побудови кривої необхідно визначити кумулятивну частину даних та кумулятивну частину класифікованих користувачів, як нормальних.

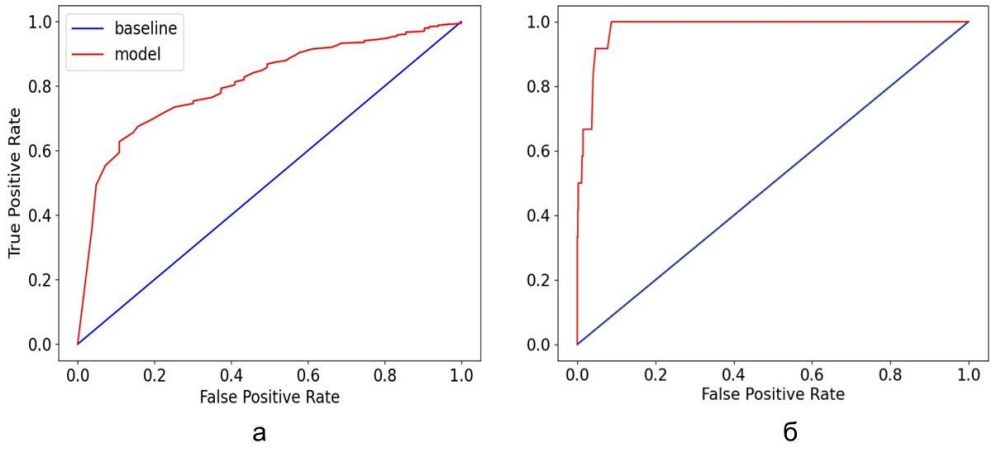

Рис. 13. Криві концентрації(а) і AUC–ROC (б)

Наступним показником якості побудованої моделі може слугувати побудова кривої ймовірності (ROC), що дає можливість виміряти якість моделі для проблем класифікації та Area Under the Curve (AUC) – площина під кривою являє собою ступінь або міру відокремюваності. Ця міра дає змогу оцінити, наскільки якісно модель здатна розрізняти класи (рис. 14б). Для визначення кривої ROC застосуємо функцію roc() в пакеті pROC. Sensitivity та Specifity задані наступними відношеннями (3).

AUC складає 98,1%. Відмінна модель має AUC близько до 1, що означає, що вона має не поганий показник відокремлюваності. Погана модель має AUC близько до 0, що означає, що вона має найгірший показник відокремлюваності. Побудуємо дерево за найбільш релевантним рішенням. Кожний з вузлів має 4 показників: клас, який розглядається (тобто 1 або 2); вірогідність події, що користувач нормальний або порушник, кількість спостережень у вузлі та ентропію (рис. 14).

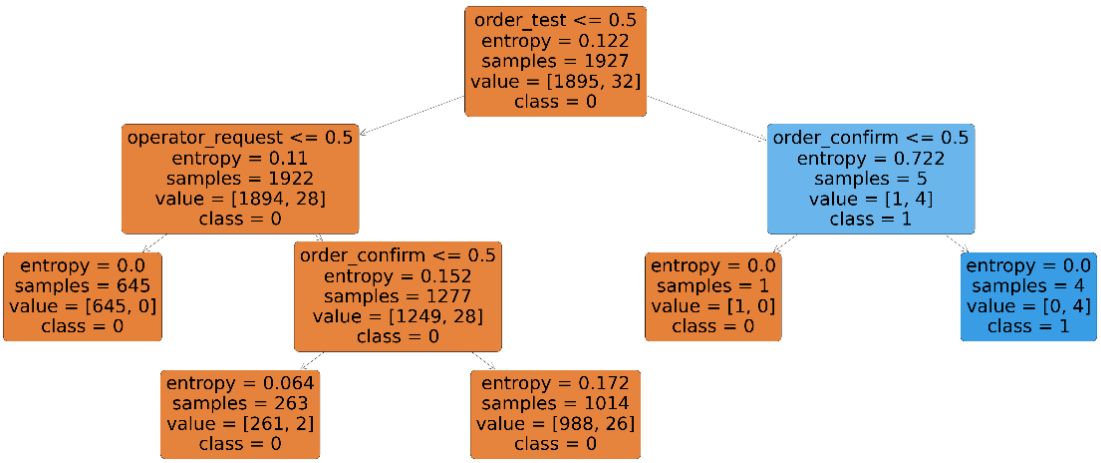

Рис. 14. Дерево за найбільш релевантним рішенням

Розглянувши перший рівень дерева, можна визначити, що 50% користувачів маркетплейсу мають нормальну поведінку. Далі вузол перевіряє чи значення «Формування замовлення в один клік» дорівнює одиниці і переходить до другого рівня. Якщо значення не дорівнює одиниці, то 11% спостережень з більш ніж двома записами є аномальними з вірогідністю 50%. Наступним кроком, у вузлі запитується, чи

«Підтвердження замовлення» дорівнює 1. Якщо дорівнює одиниці, то ми переходимо до вузла, де 6% не підтверджують замовлення і мають аномальну поведінку, як користувачі.

На цей показник не варто звертати особливої уваги до конкретної перевірки аномалії моделі, оскільки якість моделі залежить напряму від якості даних. Якість даних перевіряється після застосування моделі.

Найрозповсюдженішою проблемою вважається незбалансованість даних, коли для класифікації дані представлені не однаково. Зазвичай, коли різниці невелика, незбалансованість не викликає великих проблем, але чим більше різниця тим більше складнощів виникає для застосування моделей.

Перевіримо яка частина даних належить до першого, а яка до другого класу. Отримуємо результат, що підтверджує наше припущення, оскільки майже 98% даних складають нормальні користувачі та 2% до порушників.

Скорінг – це один із способів захисту персональних даних користувачів в основі якого закладено метод класифікації на групи. Групування може відбуватися за наступним принципом:

– користувач здійснив більш ніж 5 замовлень;

– попереднє замовлення значно перевищує середньо-статистичний кошик;

– кредитна історія тощо.

На основі критеріїв формується рейтинговий бал користувача, що вказує на його надійність, поведінку та кредитоспроможність [17, 18]. Фактично при огляді ми використовуємо ті ж дані, що і при формуванні поведінкового профілю.

Варто врахувати, можливість моніторингу поведінки користувачів порушниками з метою використання їх персональних даних, наприклад формування кредиту на користувачів з високим рейтингом чи користувачів, що здійснювали замовлення вартістю вище середнього.

Порівнюючи результати дослідів всіх методів, які представлені у табл. 3, та інших факторів можемо сказати, що метод логічної регресії має найкращі показники серед всіх досліджених алгоритмів.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Було описано основні алгоритми побудови поведінкового профілю користувачів та розглянуто метод виявлення порушника шляхом порівняння його дій з діями середньостатистичного користувача маркетплейсу.

Запропоновано власну модель профілювання поведінки користувачів на основі мови програмування Python та бібліотеки Scikit-learn методом випадкового лісу, лінійної регресії та дерева рішень, використано метрику застосовуючи матрицю помилок, проведено оцінку алгоритмів.

У результаті порівняння оцінки даних алгоритмів трьох методів, метод лінійної регресії показав найкращі результати: A – 98,60%, P – 0,01%, R – 0,54%, F – 0,33%.

У такий спосіб за допомогою порівняльного аналізу з’ясовано, що модель лінійної регресії найефективніша, оскільки ми визначаємо правильно 2% порушників, що відповідно позитивно впливає на захист персональних даних.

У подальшому заплановане дослідження механізмів покращення якості моделей. Зі сторони скорінгу розробити алгоритм не лише оцінювання користувачів, але й передбачення їхньої поведінки з метою захисту персональних даних користувачів від порушників. Розробити методичні рекомендації розв’язання таких інцидентів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Zachek, O., Dmytryk, Y. (2020). Application of Profiling to Combat Cyber Crime. Social Legal Studios 10(4), 94–100. doi:10.32518/2617-4162-2020-4-94-100.
2. Kirwan, G., Power, A. (2012). The Psychology of Cyber Crime. Advances in Digital Crime, Forensics, and Cyber Terrorism. doi:10.4018/978-1-61350-350-8.
3. Shinder, D., Tittel, E. (2002). Scene of the Cybercrime—Computer Forensics Handbook, 1st ed. Syngress Publishing.
4. Warikoo, A. (2014). Proposed Methodology for Cyber Criminal Profiling. Information Security Journal: A Global Perspective 23(4-6), 172–178. doi:10.1080/19393555.2014.931491.
5. Georgiev, V. (2019). Profiling Human Roles in Cybercrime. Information & Security: An International Journal 43(2), 145–160. doi:10.11610/isij.4313.
6. Turney, B. E. (2012). Criminal Profiling: An Introduction to Behavior Evidence Analysis. Fourth Edition (Elsevier, Oxford).
7. Conclusion. (1999). Geographic Profiling. doi:10.1201/9781420048780.ch12.
8. Muller, D. A. (2000). Criminal Profiling. Homicide Studies 4(3), 234–264. doi:10.1177/1088767900004003003.
9. Herndon, J. S., Kocsis, R. N. (2006). Criminal Profiling: Principles and Practice. Journal of Police and Criminal Psychology 22(1), 57–58. doi:10.1007/s11896-007-9005-4.
10. Rimestad, S. (2015). Seksualitāte un sociāla kontrole Latvijā 1914–1939, INETA LIPŠA, Rīga, Zinātne, 2014. ISBN 978-9984-879-65-9. Journal of Baltic Studies 46(3), 416–419. doi:10.1080/01629778.2015.1073921.
11. Kipane, A. (2019). Meaning of Profiling of Cybercriminals in the Security Context. SHS Web of Conferences. Vol. 68. P. 01009. URL: https://doi.org/10.1051/shsconf/20196801009.
12. Kshetri N. (2010). The Global Cybercrime Industry: Economic, institutional and Strategic Perspectives. Heidelberg : Springer, 2010. isbn:9783642115219.
13. Forests of Randomized Trees. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees.
14. Labintcev, E. (2017). Метрики в задачах машинного обучения. https://habr.com/ru/company/ods/blog/328372/.
15. Robinson, S. K-Nearest Neighbors Algorithm in Python and Scikit-Learn. https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/.
16. Installing Scikit-Learn. https://scikit-learn.org/stable/install.html.
17. Geetha, P., Naikodi, C., Suresh, L. (2020). K-Anonymization based Temporal Attack Risk Detection using machine learning paradigms. Journal of Circuits, Systems and Computers. doi:10.1142/S021812662150050X.
18. Protection of Personal Data. (2016). Security and Privacy in the Digital Era, 29–38. doi:10.1002/9781119347750.ch2.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (14), 2021
ISSN 2663 - 4023

Diana A. Tsyrkaniuk
Master's Student of the Department of Information and Cyber Security named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv University, Kyiv, Ukraine
ORCID ID: 0000-0002-9422-8617
d.tsyrkaniuk@gmail.com

Volodymyr Y. Sokolov
PhD, Associate Professor
Associate Professor of the Department of Information and Cyber Security named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv University, Kyiv, Ukraine
ORCID ID: 0000-0002-9349-7946
v.sokolov@kubg.edu.ua

Nataliia P. Mazur
PhD, Associate Professor
Associate Professor of the Department of Information and Cyber Security named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv University, Kyiv, Ukraine
ORCID ID: 0000-0001-7671-8287
n.mazur@kubg.edu.ua

Valerii A. Kozachok
PhD, Associate Professor
Associate Professor of the Department of Information and Cyber Security named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv University, Kyiv, Ukraine
ORCID ID: 0000-0003-0072-2567
v.kozachok@kubg.edu.ua

Volodymyr M. Astapenya
PhD, Associate Professor
Associate Professor of the Department of Information and Cyber Security named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv University, Kyiv, Ukraine
ORCID ID: 0000-0003-0124-216X
v.astapenia@kubg.edu.ua

METHOD OF MARKETPLACE LEGITIMATE USER AND ATTACKER PROFILING

Abstract. The number and complexity of cybercrime are constantly growing. New types of attacks and competition are emerging. The number of systems is growing faster than new cybersecurity professionals are learning, making it increasingly difficult to track users' actions in real-time manually. E-commerce is incredibly active. Not all retailers have enough resources to maintain their online stores, so they are forced to work with intermediaries. Unique trading platforms increasingly perform the role of intermediaries with their electronic catalogs (showcases), payment and logistics services, quality control - marketplaces. The article considers the problem of protecting the personal data of marketplace users. The article aims to develop a mathematical behavior model to increase the protection of the user's data to counter fraud (antifraud). Profiling can be built in two directions: profiling a legitimate user and an attacker (profitability and scoring issues are beyond the scope of this study). User profiling is based on typical behavior, amounts, and quantities of goods, the speed of filling the electronic cart, the number of refusals and returns, etc. A proprietary model for profiling user behavior based on the Python programming language and the Scikit-learn library using the method of random forest, linear regression, and decision tree was proposed, metrics were used using an error matrix, and algorithms were evaluated. As a result of comparing the evaluation of these algorithms of three methods, the linear regression method showed the best results: A is 98.60%, P

is 0.01%, R is 0.54%, F is 0.33%. 2% of violators have been correctly identified, which positively affects the protection of personal data.

Keywords: marketplace, user profile, user model, decision tree, behavior profiling.

REFERENCES

1. Zachek, O., Dmytryk, Y. (2020). Application of Profiling to Combat Cyber Crime. Social Legal Studios 10(4), 94–100. doi:10.32518/2617-4162-2020-4-94-100.
2. Kirwan, G., Power, A. (2012). The Psychology of Cyber Crime. Advances in Digital Crime, Forensics, and Cyber Terrorism. doi:10.4018/978-1-61350-350-8.
3. Shinder, D., Tittel, E. (2002). Scene of the Cybercrime—Computer Forensics Handbook, 1st ed. Syngress Publishing.
4. Warikoo, A. (2014). Proposed Methodology for Cyber Criminal Profiling. Information Security Journal: A Global Perspective 23(4-6), 172–178. doi:10.1080/19393555.2014.931491.
5. Georgiev, V. (2019). Profiling Human Roles in Cybercrime. Information & Security: An International Journal 43(2), 145–160. doi:10.11610/isij.4313.
6. Turney, B. E. (2012). Criminal Profiling: An Introduction to Behavior Evidence Analysis. Fourth Edition (Elsevier, Oxford).
7. Conclusion. (1999). Geographic Profiling. doi:10.1201/9781420048780.ch12.
8. Muller, D. A. (2000). Criminal Profiling. Homicide Studies 4(3), 234–264. doi:10.1177/1088767900004003003.
9. Herndon, J. S., Kocsis, R. N. (2006). Criminal Profiling: Principles and Practice. Journal of Police and Criminal Psychology 22(1), 57–58. doi:10.1007/s11896-007-9005-4.
10. Rimestad, S. (2015). Seksualitāte un sociāla kontrole Latvijā 1914–1939, INETA LIPŠA, Rīga, Zinātne, 2014. ISBN 978-9984-879-65-9. Journal of Baltic Studies 46(3), 416–419. doi:10.1080/01629778.2015.1073921.
11. Kipane, A. (2019). Meaning of Profiling of Cybercriminals in the Security Context. SHS Web of Conferences. Vol. 68. P. 01009. URL: https://doi.org/10.1051/shsconf/20196801009.
12. Kshetri N. (2010). The Global Cybercrime Industry: Economic, institutional and Strategic Perspectives. Heidelberg : Springer, 2010. isbn:9783642115219.
13. Forests of Randomized Trees. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees.
14. Labintcev, E. (2017). Metrics in Machine Learning Problems. https://habr.com/ru/company/ods/blog/328372/.
15. Robinson, S. K-Nearest Neighbors Algorithm in Python and Scikit-Learn. https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/.
16. Installing Scikit-Learn. https://scikit-learn.org/stable/install.html.
17. Geetha, P., Naikodi, C., Suresh, L. (2020). K-Anonymization based Temporal Attack Risk Detection using machine learning paradigms. Journal of Circuits, Systems and Computers. doi:10.1142/S021812662150050X.
18. Protection of Personal Data. (2016). Security and Privacy in the Digital Era, 29–38. doi:10.1002/9781119347750.ch2.