УДК 303.09:336.717.1 JEL Classification C45, G21

ОЦІНКА ЙМОВІРНОСТІ ВИНИКНЕННЯ ШАХРАЙСТВА В ПРОЦЕСІ КРЕДИТУВАННЯ КЛІЄНТІВ БАНКУ[1]

Яровенко Г.М.,
к.е.н., доцентка, доцентка кафедри економічної кібернетики
Сумського державного університету
a.yarovenko@uabs.sumdu.edu.ua
Радько В.В.,
магістрантка кафедри економічної кібернетики
Сумського державного університету
v.radko@student.sumdu.edu.ua

Статтю присвячено актуальній темі оцінці ймовірності виникнення кредитних шахрайств у банках. Дана проблематика пов'язана із зростанням рівня діджиталізації економічних процесів та переведенням платіжних операцій у цифровий простір. Її вирішення здійснюється у восьми наукових напрямках, що підтверджено шляхом побудови та аналізу карти наукометричної бібліографії досліджень, присвячених проблемі шахрайств щодо кредитування клієнтів банків. В статті було виділено кластери наукових праць, що стосуються: процесів захисту онлайн-транзакцій; машинного, ансамблевого та інкрементного навчання для вирішення проблем кредитних шахрайств; ймовірнісних підходів; процесів виявлення аномалій у операціях, пов'язаних із відмиванням незаконних коштів у банках; процесу знаходження шахрайств у фінансовій сфері; оцінки ризиків; Data Mining. Для проведення дослідження оцінки ймовірності виникнення кредитних шахрайств у банках використано статистичні дані, які складаються з 122 змінних та 307511 записів щодо клієнтів банку. Побудова концептуальної моделі дозволила окреслити етапи здійснення моделювання, яке проводилося за допомогою сучасної мови програмування Python. Дані було очищено від пропущеної інформації та перевірено на відповідність нормального закону розподілу. В результаті отриманого набору даних було побудовано три моделі - логістична регресія, дерево рішень та нейронна мережа. Виявилося, що частка правильних прогнозів у тренувальній вибірці для логістичної регресії склала 93,09%, для дерева рішень та нейронної мережі – 100,00%, а у тестовій вибірці, відповідно, – 93,60%, 99,15%, 86,67%. Це свідчить про адекватність даних обох вибірок та високу точність прогнозування. Побудовані моделі було також перевірено на точність та якість. В результаті виявилося, що всі моделі є досить точними та якісними, але дерево рішення є найбільш точною, якісною та адекватною моделлю. Побудовані моделі є універсальними інструментами для виявлення шахрайських операцій, але вони потребують постійного моніторингу та оновлення інформації у зв'язку із появою нових ознак злочинної дії в процесі кредитування клієнтів.

Ключові слова: *кредитні шахрайства, ймовірність, Python, логістична регресія, дерево рішень, нейронна мережа.*

DOI: 10.21272/1817-9215.2021.3-17

ПОСТАНОВКА ПРОБЛЕМИ

Економічна криза, низький рівень доходів населення, зростання кількості комерційних банків, спеціалізованих фінансових компаній та стрімкий розвиток інформаційних технологій сформували умови для появи банківських шахрайств. Їх мета полягає у незаконному привласненні коштів однією особою або групою осіб. Як правило, процес скоєння банківських шахрайств є прихованим. Масовість їх здійснення може вразити фінансову безпеку банку, призвести до фінансових збитків та, як наслідок, викликати втрату довіри та репутації серед клієнтів, стати однією з причин банкрутства банківської установи.

Найбільш поширеним видом банківського шахрайства є ті, що пов'язані із кредитними операціями, тобто процесом кредитування клієнтів, а також кредитними картками клієнтів. Це відбувається завдяки спрощення процедури надання кредитів, а також створення гнучких умов для клієнтів щодо використання ними кредитних коштів та засобів платежу. Також судово-бухгалтерською експертизою фінансово-кредитних установ дедалі частіше фіксуються махінації, що стосуються незаконних

[1] Робота виконана в рамках держбюджетних науково-дослідних робіт: 0121U109559 «Національна безпека через конвергенцію систем фінансового моніторингу та кібербезпеки: інтелектуальне моделювання механізмів регулювання фінансового ринку».

кредитних операцій, до яких вдаються не тільки позичальники (юридичні та фізичні особи), але й кредитори (банки, фонди, асоціації). Останнім часом відсоток таких шахрайств зростає у порівнянні із іншими видами. Так, у 2020 році шахрайство з кредитними картками зайняло друге місце серед п'ятірки найбільш розповсюджених фінансових злочинів та становило 29,7% (шахрайства із державними пільгами, на які подано заявку, або отримано – 32,0%; різні крадіжки особистих даних – 22,9%; шахрайства із позиками для бізнесу/особистісного користування – 8,1%; податкове шахрайство – 7,3%) [1].

За часту шахрая виявляють вже після того, як було скоєно злочин, тому існує потреба саме у передбаченні потенційних шахрайств. Це можливо тільки в процесі оцінювання ймовірності їх виникнення в ході кредитування клієнтів банку. У контексті даної проблеми є потреба у створенні комплексу заходів для попередження кредитних шахрайств. З цією метою доцільно застосовувати математичні методи, за допомогою яких, можна створювати математичні моделі для проведення ідентифікації банківських транзакцій на предмет шахрайства, або ідентифікації потенційного клієнта банку, який може його скоїти. Використання новітніх інформаційних технологій та мов програмування дозволяє будувати моделі будь-якого рівня складності та спрощувати їх розрахунки.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Проблема виявлення та попередження шахрайств у банківській сфері є досить актуальним напрямом дослідження. Для виявлення його тенденцій проведемо бібліометричний аналіз наукових досліджень, результати яких опубліковані в міжнародних виданнях, що були проіндексовані у базі даних Scopus. З цією метою було зроблено вибірку таких джерел та шляхом застосування аналітичного програмного забезпечення VOSviewer побудовано карту наукометричної бібліографії досліджень, присвячених проблемі шахрайств щодо кредитування клієнтів банків (див. рис. 1).

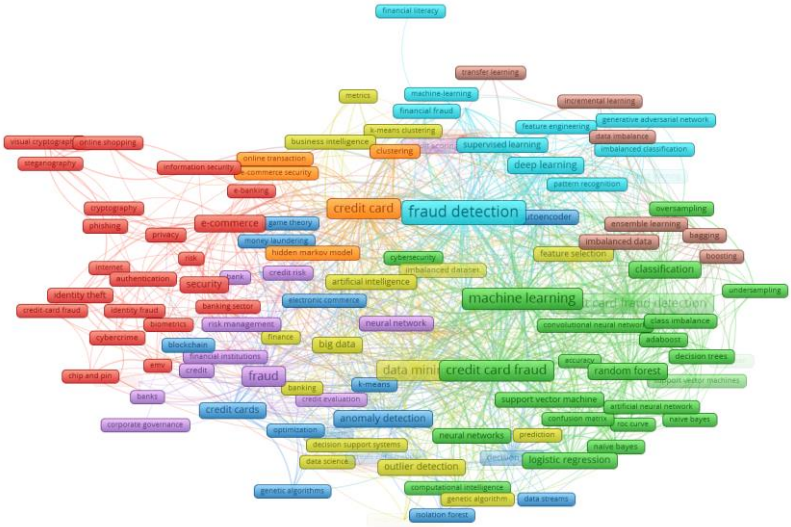

Рисунок 1 – Карта наукометричної бібліографії досліджень, присвячених проблемі шахрайств щодо кредитування клієнтів банків

Джерело дослідження: побудовано авторами на основі бази даних Scopus

Червоний кластер досліджень (див. рис. 1) охоплює публікації, які стосуються процесів захисту транзакцій електронного банкінгу, інтернет-банкінгу, електронної комерції, онлайн-магазинів, що відбуваються за допомогою засобів інформаційної безпеки – аутентифікації, ідентифікації, криптографії, стеганографії, біометрії, тощо. Тут можна виділити роботи Чжен Л., Лю Г., Ян Ч., Цзян Ч. [2], Пріша П., Нео Х.-Ф.,

Онг Т.-С., Тео Ч.-С. [3] та інші. Публікації зеленого кластеру (див. рис. 1) розкривають напрямки застосування машинного навчання для вирішення проблем кредитних шахрайств у банках. Так, застосовують такі інструменти, як кластерний аналіз, випадковий ліс, логістична регресія, нейронні мережі, методи опорних векторів, тощо. В даному напрямку працювали Діліп М.Р., Наванет А.В., Абхішек М. [4], Цуй Ю., Сон З., Ху Дж. [5] та інших. Дослідження коричневого кластеру (див. рис. 1) перетинаються із попереднім кластером, оскільки вивчаються питання ансамблевого та інкрементного навчання, бегінг та бустінг, що застосовується до незбалансованих даних. Ці аспекти досліджували Ван Р., Лю Г. [6], Собанадеві В., Раві Г. [7] та інші. Роботи помаранчевого кластеру (див. рис. 1) стосуються проблематики операцій з кредитними картками, в яких вирішуються питання за допомогою байесівського підходу та моделі Марковіца. Тут можна виділити публікації Чжоу Ю., Сон Х., Чжоу М. [8], Мішра С.П., Кумарі П. [9], тощо. Дослідження синього кластеру стосуються процесів виявлення аномалій у операціях, пов'язаних із відмиванням незаконних коштів у банках та кредитними картками, для чого дана проблема вирішується за допомогою нечіткої логіки, теорії ігор, оптимізації, великих даних та блокчейнів. Ця проблема була розкрита такими науковцями, як Рачавеліас М.Г. [10], Нана З., Сюцзянь В., Чжунцю З. [11] та інші. Публікації блакитного кластеру (див. рис. 1) відображають напрям процесу знаходження шахрайств у фінансовій сфері, для чого застосовуються розпізнавання патернів, методи глибокого навчання, навчання з вчителем та без вчителя, feature engineering, тощо. В цій сфері працювали Зоу Х. [12], Мектерович І., Каран М., Пінтар Д., Бркіч Л. [13], тощо. Напрям бузкового кольору (див. рис. 1) розкриває дослідження процесу виявлення шахрайств у банках та інших фінансових інститутах за допомогою ризиків – кредитного та операційного. Тут можна виділити роботи Джанотті Е., Даміан да Сілва Е. [14], Цзоу В., Струб Д., Венс А., Ян Дж. [15] та інших. Ключовою проблемою жовтого кластеру (див. рис. 1) є Data Mining та питання, які з ним пов'язані, а саме штучний інтелект, генетичні алгоритми, великі дані, системи прийняття рішень, бізнес-аналітика. Цим напрямом займалися Цзін Р., Тянь Х., Чжоу Г., Чжан Х., Чжен Х., Цзен Д.Д. [16], Уедраго А.-Ф., Геученн Ц., Нгуєн Ж.-Т., Тран Г. [17], тощо.

Не дивлячись на велику кількість досліджень щодо вирішення проблеми шахрайств, публікації в окреслених напрямах розкривають їх різні аспекти. Стосовно питання оцінювання ймовірності виникнення шахрайства щодо кредитування клієнтів банків, то воно потребує досліджень, особливо для реалій вітчизняних економічних відносин.

ПОСТАНОВКА ЗАВДАННЯ

Метою дослідження є побудова математичних моделей для оцінювання ймовірності виникнення шахрайства щодо кредитування клієнтів банків, їх реалізація і візуалізація за допомогою мови програмування Python.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ ДОСЛІДЖЕННЯ

Для проведення дослідження окресленої проблеми було узято базу даних з відкритих джерел даних, яка відображає операції кредитування клієнтів та містить основні їх характеристики. Так, набір даних сформували 122 змінні та 307511 спостережень. До незалежних змінних ввійшли такі показники, як: вік, стать, наявність нерухомості, тип нерухомості, наявність рухомого майна, сімейний статус, кількість дітей, тип зайнятості, рівень освіти та ін. Цільовою виступає бінарна змінна, яка є індикатором ймовірного шахрайства в процесі кредитування, а саме: 0 – виявлено клієнта – ймовірного шахрая; 1 – не виявлено клієнта – ймовірного шахрая.

З метою подальшої побудови математичних моделей розроблено концептуальну модель оцінювання ймовірності виявлення ознак шахрайства в процесі кредитування клієнтів банку, яка відображає основні етапи роботи з масивом вхідних даних та представлена на рисунку 2.

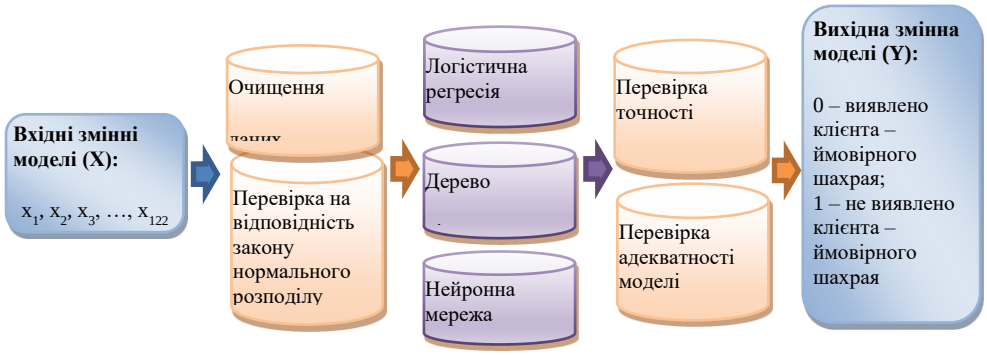

Рисунок 2 – Концептуальна модель оцінювання ймовірності виявлення ознак шахрайства в процесі кредитування клієнтів банку

Джерело: складено авторами самостійно

Виходячи з інформації концептуальної моделі (рис. 2), процес моделювання передбачає:

– попередню обробку набору даних, а саме їх очищення від пропусків та перевірка на відповідність закону нормального розподілу;

– побудову математичних моделей для оцінки ймовірності виникнення шахрайства у процесі кредитування клієнтів банку: логістичну регресію; дерево рішень; нейронну мережу;

– верифікацію побудованих моделей, тобто перевірка їх точності та адекватності.

Оскільки дослідження стосується оцінки ймовірності виникнення шахрайства, то для вирішення даної проблеми найбільш ефективними є методи інтелектуального аналізу даних Їх переваги та недоліки представлені в таблиці 1.

Таблиця 1 – Переваги та недоліки математичних моделей

Назва моделі	Переваги	Недоліки
Логістична регресія	– має один з найпростіших алгоритмів; – є легкою у виконанні та інтерпретації; – є простою у оновленні нових даних; – є ефективнішою за лінійну регресію	– алгоритм чутливий до викидів; – необхідна мінімальне значення або відсутність мультиколінеарності між незалежними змінами
Дерево рішень	– вимагає менше зусиль та часу на підготовку даних; – є дуже легким у поясненні; – не вимагає нормалізації даних; – дозволяє мати пропущені дані	– обчислення можуть бути набагато складнішими за інші алгоритми; – передбачає багато часу для навчання моделі; – є недостатнім для прогнозування безперервних значень
Нейронна мережа	– досить стійка до шуму в навчальних даних; – помилки в навчальному наборі не впливають на результат; – використовується для швидкої оцінки функції	– вимагає паралельної обробки даних; – складнощі з відображенням; – результативні значення не є оптимальними

Логістична регресія – це статистичний регресійний метод, що застосовується у випадку, коли залежна зміна являється бінарною, тобто може набувати значення 0 або 1. Логістична регресія є прогностичним аналізом та використовується для опису даних та пояснення взаємозв'язку між одним залежним фактором (змінною) та однією або декількома незалежними. Її математичний вираз можна представити формулою (1):

$$P(\hat{y} = 1) = \frac{1}{1 + exp^{-\hat{y}}} = \frac{1}{1 + exp^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}, \quad (1)$$

де: $x_1, x_2, \ldots, x_n$ – множина незалежних змінних-факторів, які впливають на результатний показник;

$\beta_0, \beta_1, \ldots, \beta_n$ – множина параметрів регресії, які необхідно оцінити в процесі побудови моделі логістичної регресії;
$\hat{y}$ – залежна змінна-фактор, значення якої прогнозується в процесі моделювання;
$P(\hat{y} = 1)$ – ймовірність виникнення випадку, при якому значення результативної змінної дорівнює 1.

Для побудови логістичної регресії використовувалася мова програмування Python. Для її реалізації дані були очищені від пропущених значень та перевірені на відповідність закону нормального розподілу. Далі набір вхідних даних було розділено на дві вибірки – тестову та тренувальну. Після здійснення даної процедури було оцінено точність опису залежної змінної незалежними, результат чого представлений на рисунку 3.

```
print("Training set score: %f" % LogReg.score(X_train, Y_train))
print("Test set score: %f" % LogReg.score(X_train, Y_train))
```

Training set score: 0.930932

Test set score: 0.936000

Рисунок 3 – Точність опису залежної змінної
Джерело: складено авторами самостійно

Результати показують, що частка правильних прогнозів у тренувальній вибірці становить 93,09%, а у тестовій – 93,60%, що свідчить про адекватність даних обох вибірок та високу точність прогнозування.

Після отримання придатного для моделювання набору, було побудовано модель логістичної регресії. Оскільки для отримання адекватної моделі необхідно, щоб її параметри були статистично значущими, то було проведено їх оцінку із використанням значення p-value та довірчих інтервалів. Статистично незначущі фактори було усунуто з моделі. Процедура повторювалася доти, доки було отримано модель із усіма статистично значущими параметрами. Так, було проведено 9 ітерацій. Результати логістичної регресії представлені на рисунку 4.

```
Optimization terminated successfully.
         Current function value: 0.627703
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:                40416
Model:                          Logit   Df Residuals:                    40409
Method:                           MLE   Df Model:                            6
Date:                Fri, 11 Jun 2021   Pseudo R-squ.:                 -1.539
Time:                        11:24:35   Log-Likelihood:                -25369.
converged:                       True   LL-Null:                       -9992.8
Covariance Type:            nonrobust   LLR p-value:                     1.000
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
NAME_TYPE_SUITE_Other_A                  -4.1769      1.008     -4.145      0.000      -6.152      -2.202
NAME_FAMILY_STATUS_Widow                 -3.4096      0.293    -11.622      0.000      -3.985      -2.835
Occupation type_Accountants              -2.8988      0.090    -32.302      0.000      -3.075      -2.723
Organization Type_Electricity            -5.2689      1.003     -5.255      0.000      -7.234      -3.304
Organization Type_Industry: type 3       -3.3706      0.322    -10.482      0.000      -4.001      -2.740
Organization Type_Military               -3.0647      0.184    -16.679      0.000      -3.425      -2.705
Organization Type_School                 -2.6126      0.128    -20.489      0.000      -2.862      -2.363
=========================================================================================================
```

Рисунок 4 – Результати оцінки та відбору параметрів логістичної регресії
Джерело: складено авторами самостійно

На рисунку 4 можна побачити, що найбільш значущими виявилися 7 параметрів логістичної регресії. Також було отримано від'ємне значення коефіцієнту детермінації. Оскільки в даному випадку розраховувався псевдо-R2, то він не має корисності, тому що не обмежений знизу. Тому для перевірки моделі на адекватність доцільно використати ROC-криву. Результати її побудови представлені на рисунку 5.

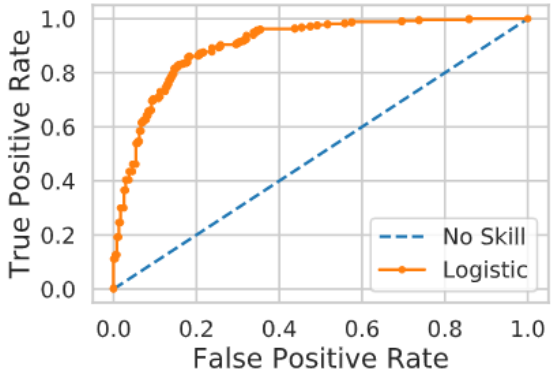

Рисунок 5 – ROC-крива для логістичної регресії
Джерело: складено авторами самостійно

На рисунку 5 показано, що логістична модель має істинне позитивне значення, яке наближається до верхнього лівого кута (до 1), тобто ROC-крива показує високу залежність кількості правильно класифікованих позитивних прикладів від кількості неправильно класифікованих негативних прикладів. Результати логістичної регресії є придатними для оцінки ймовірності виникнення шахрайств в процесі кредитування. Отриманні значення запишемо у вигляді математичної моделі логістичної регресії – формули (2):

$$P = \frac{1}{1 + E^{-2.579 - 4.1769x_1 - 3.4096x_2 - 2.8988x_3 - 5.2689x_4 - 3.3706x_5 - 3.0647x_6 - 2.6126x_7}} \qquad (2)$$

Прогнозні оцінки експоненти цієї моделі вказують на те, що коли буде змінюватися значення незалежних змінних $x_1, x_2, \dots, x_7$ на 1, ймовірність шахрайської операції буде зростати або зменшуватися у кількість разів, що відповідає визначеному значенню параметра. Наприклад, при зміні значення сімейного статусу (x_2), ймовірність шахрайства знизиться у 3,4096 разів.

Наступний метод математичного моделювання – дерево рішень. Спочатку для його побудови береться весь набір даних, що представляється кореневою вершиною. Потім визначаються варіанти розбивки даних на гілки, що відповідають кореневому вузлу. Дані гілки утворюють дерево, повернене корою вниз. Способи розбивки множини даних називають вирішальним правилом, яке відбувається за формулою (3):

$$a_{ik} = \begin{cases} 1, & s_i = r_k; \\ 0, & s_i \neq r_k, \end{cases} \qquad (3)$$

де: $a_{ik} = 1$, якщо умова s_i для правила r_k виконується;

$a_{ik} = 0$, якщо умова s_i для правила r_k не виконується;

$S\{s_i\}, i = \overline{1, l}$ – множина умов, що описують параметри обраної предметної області.

Дане правило фактично являє собою алгоритм «якщо, …то..» та ділить множину записів на дві частини [18].

Перевірка точності наборів даних виявила, що точність тестового набору дорівнює 0,9915, а тренувального – 1,0. Це означає, що точність прогнозування майже дорівнює 100%. Дерево рішень показало кращий результат ніж логістична регресія у значеннях точності опису залежної змінної незалежними змінними. Після цього проведемо побудову дерева рішень, а також здійснимо його навчання з метою отримання найкращої комбінації даних. Його результати представлені на рисунку 6.

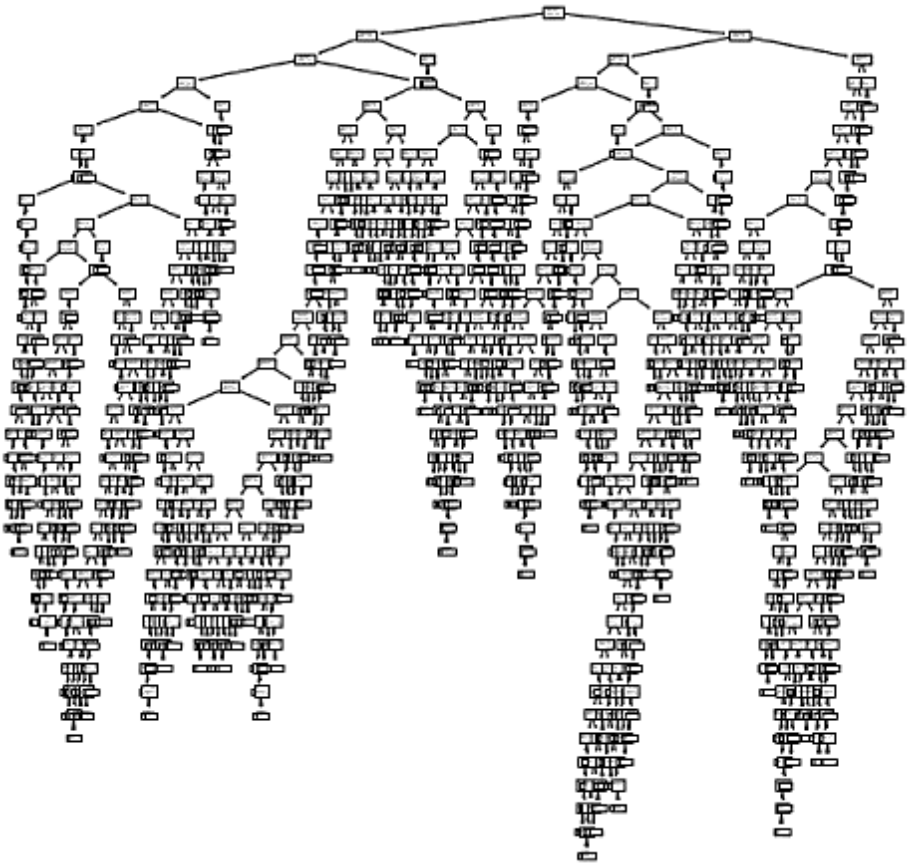

Рисунок 6 – Дерево рішень
Джерело: складено авторами самостійно

Побудувавши дерево рішень можна зробити висновок, що воно має досить складну для інтерпретації структуру, хоча має високу точність прогнозування цільової змінної, ніж логістична регресія.

Третьою моделлю є нейронна мережа, яка представляє собою алгоритм, що поєднує в собі біологічні принципи та вдосконалену статистику для вирішення задач у різних сферах. Нейронна мережа приймає базову модель нейронних аналогів, пов'язаних між собою різними способами.

Проведена перевірка точності вибірок виявила, що нейронна модель забезпечує гарну точність, що є вищою за базову (66%): для тренувального набору 100,00 %, для тестового – 86,67%. Дана модель може бути удосконалена за допомогою зміни аргументів в процесі здійснення оцінювання параметрів нейронної мережі або шляхом перехресної перевірки.

В результаті побудови та проведення навчання нейронної мережі було отримано модель, графічна інтерпретація якої представлена на рисунку 7.

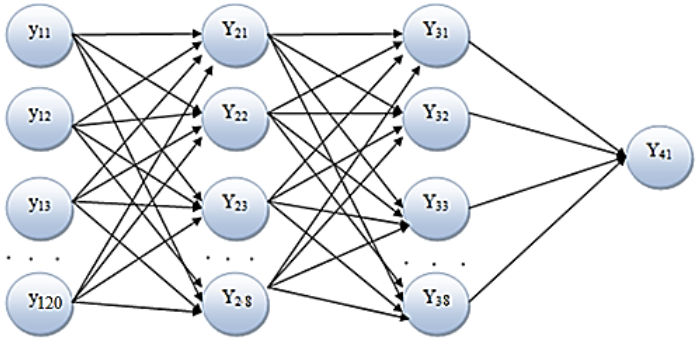

Рисунок 7 – Графічна інтерпретація отриманої нейронної моделі
Джерело: складено авторами самостійно

На рисунку 7 можна побачити, що мережа має три приховані шари. В першому буде 120 вихідних змінних, у другому – 28, у третьому – 38. На виході отримаємо фінальне рівняння. Фрагмент математичної інтерпретації нейронної мережі представлений формулами (4)-(11):

$$y_{11} = -0.1161x_1 - 0.5501x_2 + 0.3569x_3 + \cdots + 0.2065x_8 \quad (4)$$
$$y_{12} = -0.6962x_1 + 0.3309x_2 + 0.2792x_3 + \cdots + 0.2065x_8 \quad (5)$$
$$\cdots$$
$$y_{120} = 0.3499x_1 - 0.4575x_2 + 0.0761x_3 + \cdots + 0.3483x_8 \quad (6)$$
$$y_{21} = -0.3926x_1 - 0.6920x_2 + \cdots - 0.1789x_8 \quad (7)$$
$$\cdots$$
$$y_{28} = 0.5144x_1 - 0.0822x_2 + \cdots - 0.2014x_8 \quad (8)$$
$$y_{31} = -4.4977x_1 + 9.19e^{-01}x_2 + 0.27 + \cdots - 3.5448e^{-01}x_8 \quad (9)$$
$$\cdots$$
$$y_{38} = 4.44e^{-02}x_1 - 3.4965e^{-02} + \cdots + 3.4229e^{01}x_8 \quad (10)$$
$$y_{41} = 0.2948x_1 - 0.92x_2 + \cdots - 0.6388x_8 \quad (11)$$

Після побудови трьох моделей проведемо їх оцінку за точністю та якістю опису моделей тренувальним та тестовим набором даних (див. табл.2).

Таблиця 2 – Порівняння точності та якості побудованих моделей

Назва моделі	Точність, %		MSE	
	Тестові дані	Тренувальні дані	Тестові дані	Тренувальні дані
Дерево рішень	99,15	100,00	0,008	0,000
Логістична регресія	96,60	93,00	0,064	0,069
Нейронна мережа	86,70	100,00	0,089	0,004

Отримані результати точності моделей (див. табл. 2), дозволяють зробити висновок, що дерево рішень найкраще моделює ймовірність шахрайства в процесі кредитування банківських клієнтів, оскільки її точність і для тестового, і для тренувального практичного дорівнює 100%. Відповідно, значення середньоквадратичної похибки (MSE) є дуже малим та наближається до 0. Логістична регресія та нейронна мережа є майже рівноцінними, оскільки: логістична регресія має вищу точність для тестових даних, ніж нейронна мережа, а нейронна мережа, навпаки, має вищу точність для тренувального набору даних. Що стосується середньоквадратичної похибки, то її значення для обох моделей є також малим та наближається до 0. В цілому, усі три моделі дають гарні результати, тому їх можна застосовувати для оцінювання ймовірності виявлення шахрайства у процесі кредитування клієнтів банків.

ВИСНОВКИ

Проблема шахрайств у фінансовому секторі на сьогоднішній день є досить актуальною, що пояснюється впливом різних факторів. Тому на практиці існує потреба не тільки у виявленні таких випадків, але й у попередженні їх настання. Це можливо здійснити тільки із використанням сучасних інформаційних технологій та математичних методів. У даному дослідженні ця проблема вирішувалася за допомогою побудови трьох математичних моделей, що належать до класу інтелектуального аналізу даних, а саме логістичної регресії, дерева рішень та нейронної мережі. Відповідні розрахунки було проведено із використанням сучасної мови програмування Python. Дослідження передбачало формування такого набору даних, який включав не тільки кількісні характеристики клієнтів банку (дохід, депозити), але й якісні параметри – рівень освіти, тип зайнятості, сімейний стан, тип житла та ін. В результаті було отримано математичну модель логістичної регресії, яка показала досить високі значення частки правильних прогнозів у тренувальній вибірці (93,09%) та тестовій (93,60%). Тобто обидві вибірки є адекватними та демонструють високу точність прогнозування. В результаті проведеного відбору найбільш значущих параметрів було побудовано логістичну регресію із використанням семи змінних. Її гарну якість підтвердила ROC-крива. Також було побудовано дерево рішень, модель якого продемонструвала точність наборів даних вищу, ніж для логістичної моделі (для тестового набору – 0,9915, а для тренувального – 1,0). Не дивлячись на її кращі результати, модель виявилася дуже складною для інтерпретації. Нейронна мережа показала гарні показники точності вибірок: для тренувального набору 100,00 %, для тестового – 86,67%. На останньому етапі було проведено розрахунок точності та якості моделей, в результаті чого найкращі результати продемонструвала модель дерева рішень, а нейронна мережа та логістична регресії також показали гарні результати, хоча й дещо нижчі, ніж для дерева рішень. Щоб мати постійне уявлення про ймовірні шахрайства результати повинні регулярно доповнюватися, оновлюватися для використання їх у фінансових установах, що дозволить вчасно реагувати на злочинні дії та попереджати їх виникнення у процесі надання кредиту.

SUMMARY

Yarovenko H., Radko V. **Assessment of the probability of fraud in the process of lending to the bank's customers**

The article is devoted to the current topic of assessing the likelihood of credit fraud in banks. This issue is related to the growth of economic processes digitalization and the transfer of payment transactions to the digital space. Its solution is carried out in eight scientific areas, confirmed by the construction and analysis of a map of scientometric bibliography of research on the problem of fraud in lending to bank customers. The article highlights clusters of scientific papers related to processes of protection of online transactions, machine, ensemble and incremental training to solve the problems of credit fraud, probabilistic approaches, techniques of detecting anomalies in operations related to money laundering in banks, the process of finding fraud in the financial sector, risk assessments, Data Mining. The data set from 122 variables and 307,511 records of the bank's customers were used to conduct a study to assess the likelihood of credit fraud in banks. The construction of the conceptual model made it possible to outline the stages of modelling, which was carried out using the modern Python programming language. The data was cleared of missing information and checked for compliance with the normal distribution law. As a result of the obtained data set, three models were built - logistic regression, decision tree and neural network. It turned out that the share of correct predictions in the training sample for logistic regression was 93.09%, for the decision tree and neural network - 100.00%, and in the test sample, respectively - 93.60%, 99.15%, 86, 67%. It indicates the adequacy of the data of both pieces and the high accuracy of forecasting. The constructed models were also tested for accuracy and quality. As a result, it turned out that all models are pretty accurate and high quality, but the decision tree is the most accurate, high quality and adequate model. Built-in models are universal tools for detecting fraudulent transactions, but they require constant monitoring and updating of information in connection with the emergence of new signs of criminal activity in the process of lending to customers.

Keywords: *credit fraud, probability, Python, logistic regression, decision tree, neural network.*

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. How victims' information is misused. *Insurance Information Institute* : website. URL: https://www.iii.org/table-archive/20279.

2. Zheng L., Liu G., Yan C., Jiang C. Transaction fraud detection based on total order relation and behavior diversity. *IEEE Transactions on Computational Social Systems*. 2018. Vol. 5, no. 3. P. 796–806. DOI: 10.1109/TCSS.2018.2856910.

3. Prisha P., Neo H.-F., Ong T.-S., Teo C.-C. E-Commerce security and identity integrity: The future of virtual shopping. *Advanced Science Letters*. 2017. Vol. 23, no. 8. P. 7849–7852. DOI: 10.1166/asl.2017.9592.

4. Dileep M.R., Navaneeth A.V., Abhishek M. A novel approach for credit card fraud detection using decision tree and random forest algorithms. In *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*. 2021. P. 1025–10284. DOI: 10.1109/ICICV50876.2021.9388431.

5. Cui Y., Song Z., Hu J. Research on credit card fraud classification based on GA-SVM. In *Proceedings - 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering, AEMCSE 2021*. 2021. P. 1076–1080. DOI: 10.1109/AEMCSE51986.2021.00220.

6. Wang R., Liu G. Ensemble Method for Credit Card Fraud Detection. In *Proceedings - 2021 4th International Conference on Intelligent Autonomous Systems, ICoIAS 2021*. 2021. P. 246–252. DOI: 10.1109/ICoIAS53694.2021.00051.

7. Sobanadevi V., Ravi G. Handling data imbalance using a heterogeneous bagging-based stacked ensemble (hbse) for credit card fraud detection. *Advances in Intelligent Systems and Computing*. 2021. Vol. 1167. P. 517–525. DOI: 10.1007/978-981-15-5285-4_51.

8. Zhou Y., Song X., Zhou M. Supply Chain Fraud Prediction Based on XGBoost Method. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2021*. 2021. P. 539–542. DOI: 10.1109/ICBAIE52039.2021.9389949.

9. Mishra S.P., Kumari P. Analysis of techniques for credit card fraud detection: A data mining perspective. *Advances in Intelligent Systems and Computing*. 2020. Vol. 1030. P. 89–98. DOI: 10.1007/978-981-13-9330-3_9.

10. Rachavelias M.G. Online financial crimes and fraud committed with electronic means of payment – a general approach and case studies in Greece. *ERA Forum*. 2019. Vol. 19, no. 3. P. 339–355. DOI: 10.1007/s12027-018-0519-2.

11. Sadgali I., Sael N., Benabbou F. Human behavior scoring in credit card fraud detection. *IAES International Journal of Artificial Intelligence*. 2021. Vol. 10, no. 3. P. 698–706. DOI: 10.11591/IJAI.V10.I3.PP698-706.

12. Zou H. Analysis of Best Sampling Strategy in Credit Card Fraud Detection Using Machine Learning. In *ACM International Conference Proceeding Series*. 2021. P. 40–44. DOI: 10.1145/3460179.3460186.

13. Mekterović I., Karan M., Pintar D., Brkić L. Credit card fraud detection in card-not-present transactions: Where to invest? *Applied Sciences (Switzerland)*. 2021. Vol. 11, no. 151. Article number 6766. DOI: 10.3390/app11156766.

14. Gianotti E., Damião da Silva E. Strategic management of credit card fraud: stakeholder mapping of a card issuer. *Journal of Financial Crime*. 2021. Vol. 28, no. 1. P. 156–169. DOI: 10.1108/JFC-06-2020-0121.

15. Zou W., Straub D., Vance A., Yan J. The differential role of alternative data in SME-focused fintech lending. In *International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global, ICIS*. 2021. Code 167844.

16. Jing R., Tian H., Zhou G., Zhang X., Zheng X., Zeng D.D. A GNN-based few-shot learning model on the credit card fraud detection. In *Proceedings 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence, DTPI 2021*. 2021. P. 320–323. DOI: 10.1109/DTPI52967.2021.9540093.

17. Ouedraogo A.-F., Heuchenne C., Nguyen Q.-T., Tran H. Data-Driven Approach for Credit Card Fraud Detection with Autoencoder and One-Class Classification Techniques. *IFIP Advances in Information and Communication Technology*. 2021. Vol. 630 IFIP. P. 31–38. DOI: 10.1007/978-3-030-85874-2_4.

18. Обгрунтування господарських рішень та оцінка ризиків : навчальний посібник / М. Д. Балджи та ін. Одеса : ОНЕУ, 2013. 670 с.

REFERENCES

1. How victims' information is misused. *Insurance Information Institute* : website. URL: https://www.iii.org/table-archive/20279.

2. Zheng L., Liu G., Yan C., Jiang C. (2018). Transaction fraud detection based on total order relation and behavior diversity. *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 796–806. DOI: 10.1109/TCSS.2018.2856910.

3. Prisha P., Neo H.-F., Ong T.-S., Teo C.-C. (2017). E-Commerce security and identity integrity: The future of virtual shopping. *Advanced Science Letters*, vol. 23, no. 8, pp. 7849–7852. DOI: 10.1166/asl.2017.9592.

4. Dileep M.R., Navaneeth A.V., Abhishek M. (2021). A novel approach for credit card fraud detection using decision tree and random forest algorithms. In *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*. P. 1025–10284. DOI: 10.1109/ICICV50876.2021.9388431.

5. Cui Y., Song Z., Hu J. (2021). Research on credit card fraud classification based on GA-SVM. In *Proceedings - 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering, AEMCSE 2021*. P. 1076–1080. DOI: 10.1109/AEMCSE51986.2021.00220.

6. Wang R., Liu G. (2021). Ensemble Method for Credit Card Fraud Detection. In *Proceedings - 2021 4th International Conference on Intelligent Autonomous Systems, ICoIAS 2021*. P. 246–252. DOI: 10.1109/ICoIAS53694.2021.00051.

7. Sobanadevi V., Ravi G. (2021). Handling data imbalance using a heterogeneous bagging-based stacked ensemble (hbse) for credit card fraud detection. *Advances in Intelligent Systems and Computing*, vol. 1167, pp. 517–525. DOI: 10.1007/978-981-15-5285-4_51.

8. Zhou Y., Song X., Zhou M. (2021). Supply Chain Fraud Prediction Based on XGBoost Method. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2021*. P. 539–542. DOI: 10.1109/ICBAIE52039.2021.9389949.

9. Mishra S.P., Kumari P. (2020). Analysis of techniques for credit card fraud detection: A data mining perspective. *Advances in Intelligent Systems and Computing*, vol. 1030, pp. 89–98. DOI: 10.1007/978-981-13-9330-3_9.

10. Rachavelias M.G. (2019) Online financial crimes and fraud committed with electronic means of payment – a general approach and case studies in Greece. *ERA Forum*, vol. 19, no. 3, pp. 339–355. DOI: 10.1007/s12027-018-0519-2.

11. Sadgali I., Sael N., Benabbou F. (2021). Human behavior scoring in credit card fraud detection. *IAES International Journal of Artificial Intelligence*, vol. 10, no. 3, pp. 698–706. DOI: 10.11591/IJAI.V10.I3.PP698-706.

12. Zou H. (2021). Analysis of Best Sampling Strategy in Credit Card Fraud Detection Using Machine Learning. In *ACM International Conference Proceeding Series*. P. 40–44. DOI: 10.1145/3460179.3460186.

13. Mekterović I., Karan M., Pintar D., Brkić L. (2021). Credit card fraud detection in card-not-present transactions: Where to invest? *Applied Sciences (Switzerland)*, vol. 11, no. 151, article number 6766. DOI: 10.3390/app11156766.

14. Gianotti E., Damião da Silva E. (2021). Strategic management of credit card fraud: stakeholder mapping of a card issuer. *Journal of Financial Crime*, vol. 28, no. 1, pp. 156–169. DOI: 10.1108/JFC-06-2020-0121.

15. Zou W., Straub D., Vance A., Yan J. (2021). The differential role of alternative data in SME-focused fintech lending. In *International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global, ICIS*. Code 167844.

16. Jing R., Tian H., Zhou G., Zhang X., Zheng X., Zeng D.D. (2021). A GNN-based few-shot learning model on the credit card fraud detection. In *Proceedings 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence, DTPI 2021*. P. 320–323. DOI: 10.1109/DTPI52967.2021.9540093.

17. Ouedraogo A.-F., Heuchenne C., Nguyen Q.-T., Tran H. (2021). Data-Driven Approach for Credit Card Fraud Detection with Autoencoder and One-Class Classification Techniques. *IFIP Advances in Information and Communication Technology*, vol. 630 IFIP, pp. 31–38. DOI: 10.1007/978-3-030-85874-2_4.

18. Obgruntuvannia hospodarskykh rishen ta otsinka ryzykiv : navchalnyi posibnyk / M.D. Baldzhy ta in. [Substantiation of business decisions and risk assessment: a textbook]. Odessa. [in Ukrainian]