МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»

Кафедра математичного моделювання та статистики

Освітньо-професійна програма	**Економічна кібернетика**
Галузь знань	**05 Соціальні та поведінкові науки**
Спеціальність	**051 Економіка**

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему «**Оцінювання кредитних ризиків методами машинного навчання**»
(назва теми)

здобувача Яценко Ірини Вікторівни
(ПІБ, підпис)

Науковий керівник: кандидат економічних наук,
доцент, Лук'янець Т. В.
(науковий ступінь, учене звання, ПІБ)

_____
(підпис)

Робота допущена до захисту перед екзаменаційною комісією з атестації здобувачів вищої освіти (ЕК)

Завідувач кафедри кандидат фізико-математичних наук,
професор Великоіваненко Г. І.

_____
(підпис)

Київ 2023

ЗМІСТ

ВСТУП

У сучасному світі, банки та інші фінансові установи змушені працювати в умовах невизначеності і ризику. Одним з найбільш вагомих ризиків є кредитний ризик, який виникає у разі невиплати позики боржником. Це може призвести до значних втрат для банку та відчутного погіршення його фінансового стану та приведення до серйозних фінансових втрат. Традиційні методи оцінювання кредитних ризиків базуються на статистичних методах та експертній оцінці, але в останні роки все більше уваги приділяється використанню методів машинного навчання.

Оскільки кредитний ризик – це один з найважливіших аспектів фінансового менеджменту, то з огляду на зростання кількості кредитних операцій та комплексність ринку фінансових послуг, оцінювання кредитних ризиків є важливим завданням для банківської системи.

Як було зазначено одним з способів, якому все більше приділяють увагу, це застосування методів машинного навчання. Машинне навчання – це спосіб розв’язання задачі шляхом знаходження закономірностей в даних. Використання машинного навчання може зробити процес оцінювання кредитного ризику більш точним і швидким, знизити ризик помилкового прийняття рішення та покращити загальну якість кредитного портфеля банку чи фінансової установи.

Актуальність теми дослідження. Тема дослідження залишається актуальною в сучасному фінансовому середовищі. Кредитні ризики відображають потенційні втрати, які можуть виникнути внаслідок невиконання позичальниками своїх фінансових зобов’язань. Оцінка цих ризиків є важливою задачею для банків, фінансових установ і кредиторів загалом.

Машинне навчання, як підгалузь штучного інтелекту, має потенціал вирішувати складні завдання, пов’язані з оцінюванням кредитних ризиків. Завдяки методам машинного навчання можна аналізувати великі обсяги даних про позичальників і виявляти складні зв’язки та патерни, які вказують на

потенційні ризики. Це може допомогти у розробленні більш точних і об’єктивних рішень щодо надання кредиту і визначення ставок відсотків.

Крім того, фінансова сфера постійно розвивається, з’являються нові фінансові продукти, а також змінюються умови функціонування. Це створює потребу в оновленні і вдосконаленні методів оцінювання кредитних ризиків. Машинне навчання може бути корисним інструментом для аналізу нових фінансових даних і адаптації моделей оцінювання ризиків до умов, що змінюються.

Загалом, використання методів машинного навчання в оцінюванні кредитних ризиків є актуальною темою дослідження, оскільки вона спрямована на покращення процесів кредитування, зниження ризиків фінансових втрат і покращення прийняття рішень у фінансових установах.

Метою дослідження є аналіз кредитних ризиків на основі сучасних методів машинного навчання, порівняння результатів з традиційними методами оцінювання ризиків.

Завданнями дослідження є:

- визначити та дослідити сутність кредитних ризиків, основні методи управління ними;
- проаналізувати сучасні методи оцінювання кредитних ризиків;
- обрати вхідний масив даних для дослідження;
- вибрати методи й моделі машинного навчання для аналізу кредитного ризику;
- виконати аналіз результатів дослідження;
- порівняти якість отриманих результатів за обраними моделями та обрати найкращу для оцінювання кредитних ризиків.

Об’єктом дослідження є кредитні ризики у фінансовій сфері.

Предметом дослідження є методи машинного навчання, що застосовуються для оцінки кредитних ризиків.

Аналіз останніх досліджень і публікацій. Оцінювання та управління кредитними ризиками досліджується в роботах як українських, так і зарубіжних

вчених: Вітлінський В. В., Великоіваненко Г. І., Ніколаєнко Ю. В., Поляруш І. М., Брігхем Є., Версаль Н. І., Васильченко З. М., Роуз П. С., Валенцева Н.І..

У роботі розглянуто наступні методи машинного навчання, такі як дерева рішень, логістична регресія, нейронні мережі та метод опорних векторів. Для кожного методу проведено експерименти з використанням різних параметрів та виконано порівняння їх результатів.

В результаті цієї роботи було досліджено модель кредитного ризику різними методами машинного навчання. Це дослідження допомогло виявити кращий метод машинного навчання, за яким будуть достатньо точно оцінювати кредитний ризик клієнтів.

Отже, використання методів машинного навчання для оцінювання кредитних ризиків є корисним для фінансових установ, оскільки дозволяє знизити ризик неправильного прийняття рішення та покращити кредитний портфель.

РОЗДІЛ 1

ТЕОРЕТИЧНІ АСПЕКТИ УПРАВЛІННЯ КРЕДИТНИМИ РИЗИКАМИ

1.1 Класифікація кредитних ризиків

Кредитний ризик – це ймовірність фінансових втрат через неспроможність позичальника повернути кредит. По суті, кредитний ризик – це ризик того, що кредитор може не отримати основну суму боргу та відсотки, а тобто свій прибуток. Що в наступному призведе до переривання грошових потоків та збільшення витрат на стягнення заборгованості. Кредитори можуть зменшити кредитний ризик, аналізуючи фактори кредитоспроможності позичальника, такі як поточне боргове навантаження та дохід.

Хоча неможливо точно знати, хто саме не виконає своїх зобов'язань, належна оцінка та управління кредитним ризиком може зменшити серйозність збитків. Відсоткові платежі від позичальника або емітента боргового зобов'язання є винагородою кредитора або інвестора за прийняття кредитного ризику.

Об'єктом кредитного ризику вважають фізичну або юридичну особу, яка бажає отримати кредит або вже його отримала. Іншими словами можна сказати, що це позичальник.

Суб'єктом кредитного ризику є організація, що надає кредитні послуги. Це може бути як і банк, так і фінансова установа. Тобто іншими словами це кредитор.

Джерело кредитного ризику є фактори або події, які можуть призвести до невиконання позичальником своїх фінансових зобов'язань та створити небезпеку для кредитора або фінансової установи [1].

Оцінка кредитного ризику залежить від багатьох факторів, таких як кредитна історія, платіжна здатність, наявність гарантій та інші ризикові чинники. Іноді кредитний ризик може бути пов’язаний з ризиком банкрутства, затримкою платежів, зміною відсоткової ставки та іншими факторами, які можуть вплинути на спроможність позичальника повернути позику кредитору.

Коли кредитори пропонують іпотеку, кредитні картки або інші види позик, існує ризик того, що позичальник може не повернути кредит. Аналогічно, якщо компанія пропонує кредит клієнту, існує ризик того, що клієнт може не сплатити рахунки.

Кредитні ризики розраховуються на основі загальної здатності позичальника погасити кредит відповідно до його початкових умов. Щоб оцінити кредитний ризик за споживчим кредитом, часто аналізують наступні дані клієнта: кредитну історію, платоспроможність, капітал, умови кредиту та відповідну заставу.

Багато компаній створили цілі відділи, що відповідають за оцінку кредитних ризиків своїх поточних і потенційних клієнтів. А сучасні технології надали бізнесу можливість швидко аналізувати дані, що використовуються для оцінки профілю ризику клієнта.

Ризик є досить складним явищем. Його можна охарактеризувати абсолютно за допомогою різних факторів. Це дослідження може бути пов’язано з походженням, місцем розташування, масштабом, характеристиками прояву ризику та іншими обставинами ризиків. Але ми розглянемо з огляду на те, які результати можемо отримати. І тут ми розділяємо їх ще на дві досить великі групи ризиків, а саме: чисті та спекулятивні ризики [2].

Чисті ризики – це ризики, які пов’язані з невизначеністю і не мають можливості для отримання вигоди чи прибутку. Ці ризики можуть мати негативні наслідки для сторін, в нашому випадку як і для кредитора, так і для позичальника. Спекулятивні ризики – це ризики, пов’язані з високою невизначеністю і можливістю отримання або ж високих винагород, або ж високих збитків в результаті участі в спекулятивній діяльності.

Чисті та спекулятивні ризики зазвичай вимірюються у фінансовому відношенні. Тож надалі розглянемо різновид спекулятивних ризиків, а саме фінансовий ризик, що несе за собою фінансові втрати підприємства. Виникає він через нестабільність фінансового ринку та втрати на ньому за рахунок змін цін акцій, валют, відсоткових ставок. Його нестабільність та невизначеність може призвести до великих збитків, а в деяких випадках навіть до банкрутства підприємства.

Загалом розподіл фінансових ризиків набуває такого вигляду [3]:

- *ризик ринкової ціни* – це ризик, пов'язаний зі змінами в ринковій ціні активу. Цей ризик може бути зв'язаний з різними активами, такими як акції, облігації, товари та інші.
- *ризик кредитування* – це ризик, пов'язаний з можливістю невиконання зобов'язань з боку позичальника. Цей ризик може бути знижений за допомогою відповідного оцінювання кредитоспроможності позичальників.
- *ризик ліквідності* – це ризик, пов'язаний зі здатністю продати активи на ринку. Цей ризик може бути зменшений за допомогою збалансованого плану ліквідації активів.
- *ризик валюти* – це ризик, пов'язаний зі змінами в курсі валют. Цей ризик може бути зменшений за допомогою відповідних заходів ризик-менеджменту.
- *ризик політичної стабільності* – це ризик, пов'язаний зі змінами в політичній ситуації, що може вплинути на фінансові ринки. Цей ризик може бути зменшений шляхом диверсифікації інвестицій у різні країни.
- *ризик зміни регуляторних умов* – це ризик, пов'язаний зі змінами в законодавстві або регуляторних умовах, що може вплинути на фінансову діяльність. Цей ризик може бути зменшений шляхом вивчення законодавства та регуляторних умов та планування дій відповідним відділом.

Тобто цей ризик має право на існування за проведення різного виду діяльності у біржовій та фінансовій сфері, а також в операціях з цінними

паперами. За рахунок даного ризику є велика ймовірність понести збитки та збанкрутіти.

З даного розподілу можна виокремити фінансові та кредитні категорії ризиків, хоч серед вчених та дослідників даної сфери, єдиної так і немає щодо такого розподілу. Проте часто можна зустріти в літературі, яка б вона не була, вітчизняна чи іноземна, «кредитні ризики» входять до складу поняття «фінансові ризики» та є їхньою частиною.

Та все ж кредитний ризик є найбільший та найвагомішим ризиком банків та фінансових установ. Виникає він, коли позичальники або ж контрагенти не виконують своїх договірних зобов'язань. Тобто виникає ймовірність неповного або повного невиконання зобов'язань, чи то погашення кредиту, чи то іншого виду боргового зобов'язання з боку позичальника.

Незважаючи на те, що банки не можуть бути повністю захищені від кредитного ризику через характер їхньої бізнес-моделі, вони можуть зменшити свій ризик кількома способами. Оскільки погіршення ситуації у позичальника або певній галузі є часто непередбачуваним явищем, банки зменшують свої ризики шляхом диверсифікації кредитного портфеля.

Таким чином, під час кредитного спаду банки мають меншу ймовірність надмірного впливу на категорію з великими збитками. А щоб зменшити свій ризик, вони можуть кредитувати людей із хорошою кредитною історією, проводити операції з високоякісними контрагентами або застосовувати застави чи інші гаранти. Та для всього цього перш за все потрібно провести аналіз кредитоспроможності позичальників.

Коротко розглянемо ще структуру кредитного ризику [4]. Приватні банки, компанії з активами, в тому списку й страхові компанії, мають чітко та якісно аналізувати можливі для них кредитні ризики (рис. 2.1).

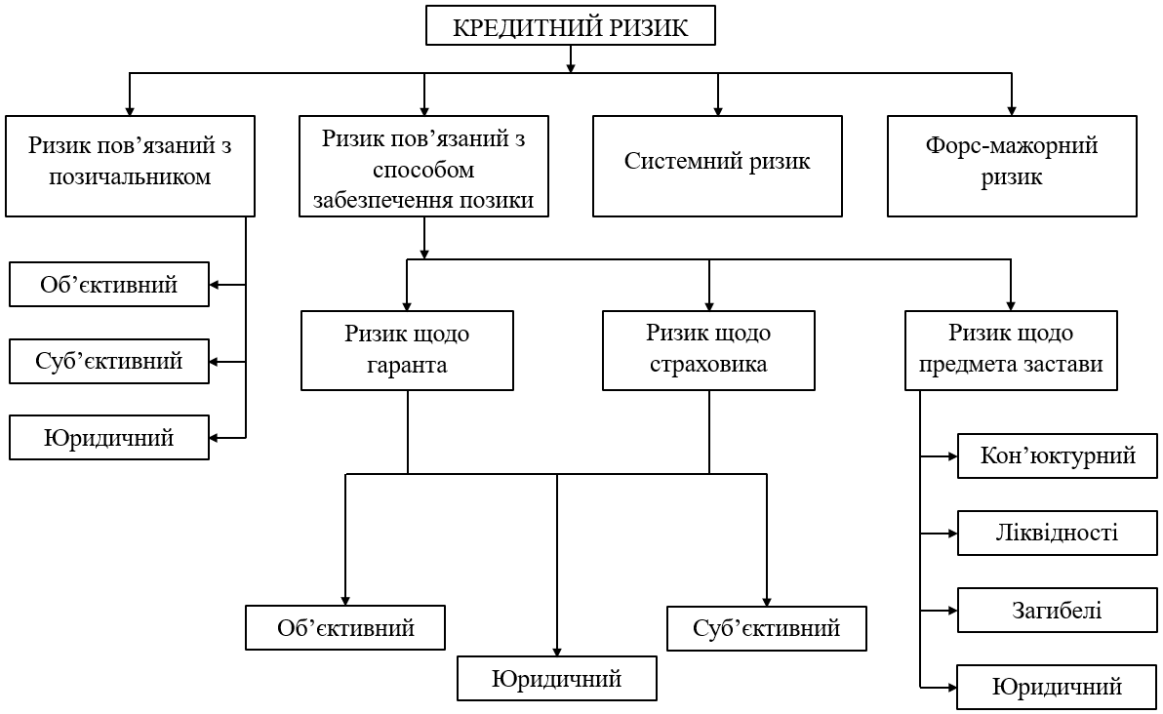

Рисунок 1.1 – Структура кредитного ризику [4]

Тож виділимо основну класифікацію кредитних ризиків. Як бачимо вони діляться на такі групи як:

- ризик пов'язаний з позичальником;
- ризик пов'язаний з способом забезпечення позики;
- системний ризик;
- форс-мажорний ризик.

В свою чергу деякі з них також діляться на певні групи. Так от як ризик пов'язаний з позичальником ділиться на об'єктивний, суб'єктивний та юридичний.

Суб'єктивний ризик – це ризик невиплати кредиту, який пов'язаний з внутрішніми факторами позичальника. Оцінюється він на основі даних позичальника, а саме його кредитної історії, фінансових показників, платоспроможності та інших факторів, які можуть впливати на здатність позичальника повернути кредит та відсотки за нього.

Об'єктивний ризик – це ризик невиплати кредиту, який пов'язаний навпаки з зовнішніми факторами. Такими як макроекономічна ситуація, стан ринку, політична нестабільність, ризики валютних коливань, ризики процентної ставки та інші фактори, що залежать від зовнішнього середовища.

Юридичний ризик може бути пов’язаний з помилками, які були допущені при видачі кредиту, такими як неправильний аналіз фінансової звітності або недостатньою перевіркою документів позичальника. Також він може бути пов’язаний з недобросовісністю позичальника, а саме у зв’язку з юридичними факторами, такими як порушення законодавства, несумісність правових форм, відсутність документів та інші проблеми, пов’язані з юридичними питаннями.

Ризик пов’язаний з способом забезпечення позики має трішки інший поділ на групи. Він ділиться на

- ризик щодо гаранта;
- ризик щодо страховика;
- ризик щодо предмета застави.

Перші дві групи також мають поділ на групи та вони є такими ж самими як і в ризиках пов’язаних з позичальником. А от ризик щодо предмета застави має такі групи як: кон’юктурний ризик, ризик ліквідності, ризик загибелі та вже відомий юридичний ризик.

Кон’юктурний ризик – пов’язаний з можливим змінами в економічному середовищі, які можуть вплинути на спроможність позичальника повернути кредитні кошти згідно з умовами договору. Цей тип ризику пов’язаний з підвищенням ризику дефолту позичальника внаслідок негативних змін в економіці, таких як зменшення попиту на товари або послуги, падіння цін на ринку, зниження виробництва тощо.

Ризик ліквідності – пов’язаний з тим, що позичальник може не мати достатньої ліквідності для того, щоб повернути кредитні кошти в строк або здійснити платежі по відсоткам за кредитним договором.

Ризик загибелі – пов’язаний з тим, що позичальник не зможе повернути кредит через смерть або непередбачувані обставини, такі як стихійні лиха, війни, терористичні акти та інше.

1.2 Методи оцінювання кредитних ризиків

До методів оцінювання кредитних ризиків можна віднести: кредитний скоринг; аналіз фінансової звітності; коефіцієнтний метод; метод комплексного аналізу. Опишемо їх нижче.

Кредитний скоринг. Скоринг є методом оцінювання кредитного ризику, який базується на статистичних моделях [5]. В основі цього методу лежить аналіз різноманітних параметрів, які характеризують позичальника. Наприклад, його дохід, стаж роботи, кредитна історія та інші фактори, які можуть впливати на здатність позичальника повернути кредит.

Використання кредитного скорингу фінансовими установами дає можливість прогнозувати, наскільки ймовірним є повернення кредиту клієнтом. Результати моделювання дозволяють оцінювати ризики неповернення кредитів та одночасно не відмовляти потенційним клієнтам, які мають високу ймовірність повернення коштів. В порівнянні з даними Бюро кредитних історій, скорингова оцінка на основі Big Data є більш актуальною, оскільки інформація з Бюро може бути застарілою та містити інформацію про період, коли клієнт був некредитоспроможним, або може бути взагалі відсутньою, якщо клієнт не мав кредитної історії.

Параметри оцінюються за допомогою математичних алгоритмів, які дозволяють встановити ймовірність того, що позичальник зможе повернути кредит. Чим вище отриманий кредитний скоринговий бал, тим більші шанси на успішне повернення кредиту має позичальник.

Найпопулярнішими моделями кредитного скорингу є FICO та VantageScore [5].

Моделі кредитного рейтингу можуть дещо відрізнятися за тим, як вони оцінюють кредит. Система оцінки кредитоспроможності Fair Isaac Corporation, відома як рейтинг FICO [7], є найпоширенішою системою оцінки

кредитоспроможності у фінансовій галузі, яку використовують понад 90% провідних кредиторів.

Кредитний рейтинг FICO – це число від 300 до 850, причому 850 є найвищим можливим балом. Кредитні показники для малих підприємств, наприклад, FICO Small Business Scoring Service (SBSS), коливаються від нуля до 300.

На кредитний рейтинг впливають п’ять категорій:

- Історія платежів (35%)
- Суми заборгованості (30%)
- Тривалість кредитної історії (15%)
- Новий кредит (10%)
- Сума кредитів (10%)

Оцінка від 800 до 850 вважається винятковою. Середня оцінка FICO становить 714, а приблизно 21% людей мають оцінку 800 або більше. У середньому менше 1% позичальників із винятковою кредитною оцінкою серйозно прострочують свої платежі.

Однак наступною популярною моделлю для оцінки кредитоспроможності – VantageScore. Вона була створена фундаментальними компаніями кредитної звітності, а саме TransUnion, Experian і Equifax [8].

Цей метод дещо відрізняється від FICO Score за деякими параметрами, такими як підхід до збору даних, оцінки кредитного ризику і формат звіту. Крім того, він має шкалу оцінок від 300 до 850, яка є ідентичною зі шкалою FICO Score.

Основна перевага VantageScore полягає в тому, що він використовує новітні моделі аналізу даних та статистичні методи, що дозволяє оцінити кредитний ризик більш точно і об’єктивно.

Аналіз фінансової звітності. Аналіз фінансової звітності також є одним з методів оцінювання кредитного ризику. Цей метод використовується для оцінки фінансової стабільності позичальника та його здатності повернути кошти.

При аналізі фінансової звітності оцінюються фінансові показники, такі як прибуток, активи, заборгованість, капітал та інші. Застосування методу аналізу

фінансової звітності дозволяє визначити ризик несплати кредиту, ризик банкрутства та загальний ризик інвестування в компанію.

Для проведення аналізу фінансової звітності використовуються різні показники та методи, такі як аналіз балансу, аналіз рахунку результатів, рентабельність активів, показники ліквідності та інші. Використання цих методів дозволяє отримати повний інформаційний звіт про фінансовий стан позичальника, який в подальшому допоможе визначити ризики, пов’язані з видачою кредиту [9].

Аналіз фінансової звітності є важливим етапом процесу кредитування та дозволяє банкам та іншим фінансовим установам приймати рішення про видачу кредиту та визначення умов його повернення.

Коефіцієнтний метод. Це метод оцінювання кредитного ризику є одним з найпоширеніших методів оцінювання кредитного ризику. Цей метод полягає у використанні різних коефіцієнтів та показників, що допомагають визначити можливість позичальника погасити кредитні зобов’язання в строк [10].

До основних коефіцієнтів належать такі:

1. Коефіцієнт ліквідності (current ratio) – відображає співвідношення поточних активів та поточних зобов’язань; чим вище значення, тим більша ймовірність того, що позичальник зможе вчасно погасити свої зобов’язання.

2. Коефіцієнт заборгованості (debt ratio) – відображає співвідношення загальної заборгованості позичальника до його загального капіталу; чим нижче значення, тим менша ймовірність того, що позичальник не зможе погасити свої зобов’язання.

3. Коефіцієнт прибутковості активів (return on assets) – відображає співвідношення прибутку позичальника до його активів; чим вище значення, тим більша ймовірність того, що позичальник зможе погасити свої зобов’язання.

4. Коефіцієнт покриття процентів (interest coverage ratio) – відображає співвідношення чистого прибутку до обсягу процентних платежів; чим вище значення, тим більша ймовірність того, що позичальник зможе погасити свої зобов’язання.

5. Коефіцієнт фінансової стійкості – відображає загальну фінансову стійкість позичальника; чим вище значення, тим менший кредитний ризик.

Метод комплексного аналізу. Метод комплексного аналізу є одним з методів оцінювання кредитного ризику, який дозволяє проводити більш об'єктивну оцінку фінансового стану позичальника [11]. Цей метод полягає в аналізі різних показників, що відображають фінансовий стан позичальника, і враховуються при визначенні кредитного ризику.

Для проведення комплексного аналізу зазвичай використовують такі фінансові показники, як загальний обсяг активів, прибуток, рентабельність, фінансові показники ризику, а також дивіденди, вартість акцій тощо.

Кожен показник оцінюється за своєю вагою, що відображає його значимість в оцінці кредитного ризику. Для кожного показника також встановлюється межа безпеки, яка вказує на те, що якщо значення показника нижче цієї межі, то ризик видачі кредиту збільшується.

У результаті аналізу кожен показник отримує свою оцінку, яка потім використовується для розрахунку загальної оцінки фінансового стану позичальника. Ця оцінка дозволяє визначити кредитний ризик і прийняти рішення про видачу кредиту або відмову у ньому.

1.3 Методи управління кредитними ризиками

Управління кредитним ризиком означає вимірювання певних ризиків та застосування заходів щодо їх усунення, пов'язаних із позиченою сумою, а також знання про резерви банку, які будуть використані в будь-який момент часу. Управління ризиками тут передбачає сприяння прийняттю правильних рішень банківськими та фінансовими установами [12].

Управління кредитним ризиком включає перевірку низки факторів, що гарантують повернення суми позичальником та надійність його як клієнта вцілому. Насамперед очікується, що кредитори мають ретельно розглядати заявки позичальників на кредит. Крім того, вони повинні гарантувати, що позичальники зможуть здійснювати щомісячні платежі в майбутньому.

Кредитори мають перевіряти фінансову ситуацію, кредитну історію та кредитний бал позичальників. В подальшому це формує довіру до позичальників, яким вони все ж погодили заявку на кредит. З іншого ж боку, якщо ситуація буде обернена, заявку відхиляють та фіксують клієнта як неплатоспроможного.

У контексті управління кредитним ризиком можна виділити дві категорії: ризик на рівні конкретного позичальника (або окремої позики) та ризик на рівні кредитного портфеля (рис. 1.2.).

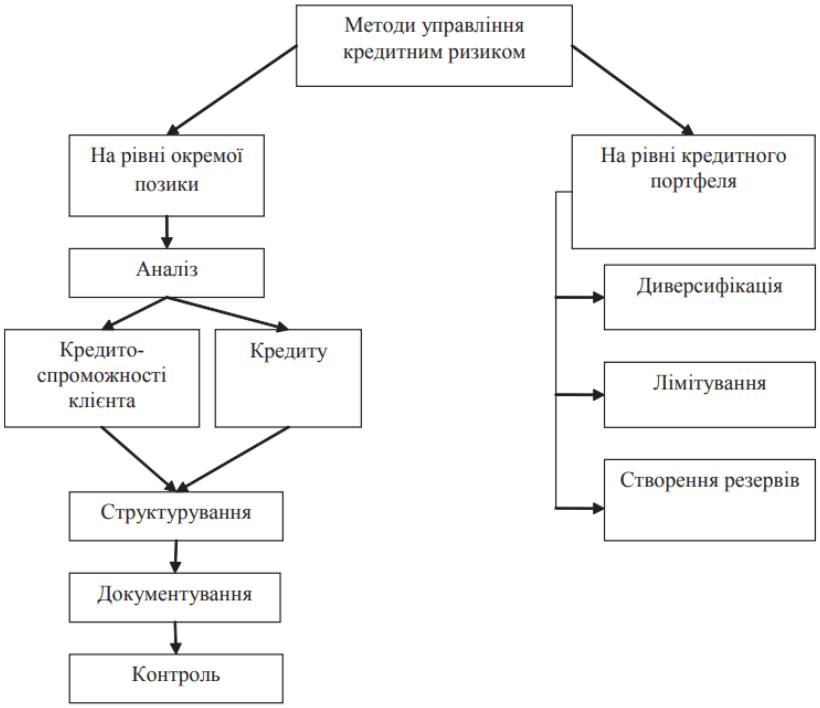

Рисунок 1.2. – Схема класифікації методів управління кредитними ризиками [12]

На рівні окремого клієнта. Під час роботи з клієнтом, в результаті якої можливе укладання угоди, менеджер (верифікатор) має ретельно проаналізувати його кредитоспроможність по вибраному кредиту та знайти фактори, які можуть зумовити його непогашення або запропонувати інші умови (кредит), який він зможе осилити.

Кредитоспроможність передбачає можливість позичальника повернути позичені кошти в найближчому майбутньому, і це відрізняється від платоспроможності. Рамки кредитоспроможності залежать від кількох чинників, таких як розмір кредиту, термін його повернення, діяльність позичальника та його забезпечення. Для отримання інформації про можливого позичальника менеджери можуть використовувати різноманітні джерела, але основними є фінансові звіти клієнта.

Аналіз реального стану потенційного позичальника, його перспектив та здатності погасити кредит, в даному випадку буде проведений шляхом розрахунку різних коефіцієнтів на основі даних звітів. Для цього також досліджується реноме позичальника, доброчесність та порядність, взаємини з іншими установами, досвід та знання галузі, у якій він працює, особисте благополуччя позичальника, ринкова вартість бізнесу та інші фактори.

При оцінці кредитоспроможності, банки та інші фінансові установи повинні дотримуватися свого внутрішнього регулювання та зважати на нормативні вимоги Національного банку України, зокрема Рекомендації щодо фінансового стану позичальників та Положення про порядок формування і використання резерву на можливі втрати позиками комерційних банків. Згідно з цими документами, Національний банк України розрізняє три категорії позичальників: юридичні особи (крім комерційних банків); комерційні банки; та фізичні особи.

Українські комерційні банки найчастіше використовують кілька форм забезпечення виплат позичальником своїх зобов’язань перед банком, зокрема поручительство третьої особи, заставу майна, стягнення пені і штрафів, страхування відповідальності позичальника за неповернення кредиту та ризику його непогашення. Цивільний кодекс України визначає всю необхідну правову основу для цих форм застави.

На рівні кредитного портфеля:

1) Диверсифікація. Диверсифікація є стратегією управління ризиками , яка поєднує широкий спектр інвестицій у портфель. Диверсифікація кредитного

портфеля – це метод управління кредитним ризиком, що полягає у розподілі кредитних ресурсів між різними позичальниками, галузями, регіонами та іншими факторами, що можуть впливати на кредитний ризик [13]. Цей метод зменшує ризик збитків в разі невиплати позик, оскільки ризики розподіляються між різними позичальниками.

Диверсифікація може бути реалізована через різні види кредитних продуктів, такі як кредитні лінії, кредитні картки, іпотечні кредити та інші. Для розподілу кредитних ресурсів можуть використовуватися різні методи, такі як географічна диверсифікація, розподіл за рівнем доходу, розподіл за видами підприємств та інші.

Диверсифікація кредитного портфеля є важливим інструментом для зменшення кредитного ризику. Вона дозволяє зменшити концентрацію ризику та збільшити ефективність управління кредитним портфелем. Однак, важливо також враховувати, що диверсифікація може підвищити витрати на управління портфелем, а також може знизити прибутковість портфеля у разі обмеженої кількості перспективних кредитних можливостей.

Обґрунтування цього метода полягає в тому, що портфель, створений із різних видів активів, у середньому принесе вищі довгострокові прибутки та знизить ризик будь-якого окремого володіння чи цінного паперу.

2) Лімітування (обмеження ризику) – це загальний і широко використовуваний метод управління ризиками та портфелем. Він позначає один або кілька числових порогів, визначених у зв’язку з конкретними ризиками, такими як кредитний ризик, ринковий ризик або ризик ліквідності [14].

Ліміти внутрішньої політики щодо ризиків спрямовані на те, щоб рівень ризику, який приймає організація, був нижчим за прийнятний.

Загальна організація лімітів ризику зазвичай здійснюється в контексті системи лімітів, яка може бути частиною ширшої системи апетиту до ризику. Ліміти ризику найбільш широко використовуються в управлінні ринковим ризиком, кредитним ризиком і ризиком ліквідності, де кількісні показники, що характеризують схильність до ризику, можуть бути встановлені з певною мірою

визначеності. Ліміти можуть встановлюватися на різних рівнях ієрархії, залежно від типу ризику: наприклад, загальні, галузеві, регіональні, за бізнес-лініями, ліміти на одного боржника. В ідеалі це має бути узгоджена система лімітів.

Іноді вводять поняття ліміту на основі ризику, щоб відрізнити системи лімітів, які використовують додаткові (потенційно залежні від моделі ризику) параметри, від тих, які ґрунтуються безпосередньо на спостережуваних показниках.

Ліміт ризику визначається через:

- застосовної метрики ризику (наприклад, умовна сума, вартість під ризиком тощо)
- граничне значення, виражене через метрику ризику (наприклад, 10 млн., 1% від загального ризику тощо)
- сферу його застосування (посада, відносини, відділ, підрозділ тощо)
- особа, відповідальна за дотримання ліміту (трейдер, портфельний менеджер, бізнес-лінія тощо)

Типи лімітування:

- ліміти можуть бути м’якими або жорсткими залежно від внутрішньої процедури, якої дотримуються у випадках перевищення ліміту. М’які ліміти мають на меті забезпечити структурований спосіб ескалації та затвердження порушення ліміту.
- ліміти можуть бути абсолютними або відносними за своєю природою. Абсолютні ліміти встановлюють грошову суму (наприклад, 10 млн. грн.), тоді як відносні ліміти встановлюють відсоток від відповідного портфеля (наприклад, 5% від суми ризику або капіталу тощо).
- ліміти можуть бути єдиними або індивідуальними. Уніфіковані ліміти застосовуються однаково до портфеля або портфелів. Індивідуальні ліміти є більш деталізованими. Наприклад, це може бути конкретний ліміт, встановлений для кожного сектору або навіть для кожного контрагента.

У контексті управління кредитним ризиком ліміти ризику можна класифікувати на такі типи:

- ліміти на експозицію, які обмежують індивідуальну або загальну суму кредитного ризику з точки зору, наприклад, залишків за кредитами, умовних сум кредитних деривативів тощо;
- ліміти строку, що обмежують індивідуальний або середній строк погашення кредитного портфеля;
- ліміти на основі кредитного рейтингу, що обмежують індивідуальний або середній рейтинг;
- ліміти на основі очікуваних кредитних збитків, що враховують як ймовірність дефолту, так і збитки в разі дефолту за позиціями;
- ліміти на основі волатильності, вартості під ризиком та капіталу, що обмежують внесок ризику в неочікувані збитки, споживання капіталу під ризиком або регулятивного капіталу.

Лімітування кредитного портфеля є ефективним методом управління кредитним ризиком, але його застосування може зменшити можливості збільшення прибутку за рахунок збільшення обсягів кредитування. Крім того, для ефективного застосування цього методу потрібна достатня кількість даних про кредитний портфель та позичальників, що може бути складним для менших кредитних установ.

3) Створення резервів кредитного портфеля – це ще один з методів управління кредитним ризиком, що передбачає виділення коштів на спеціальний рахунок для покриття можливих збитків в результаті неповернення позичок або інших кредитних зобов’язань [16].

Створення резервів дає банкам можливість зменшити вплив невиплат на їх прибутковість та забезпечити фінансову стабільність. Крім того, це є одним з вимог регуляторних органів щодо забезпечення кредитної безпеки.

Резерви можуть бути створені на рівні окремих кредитних операцій, груп кредитів або загального кредитного портфеля. Розмір резервів зазвичай

визначається відсотком від вартості кредитів, що належать до певної категорії ризику.

Для створення резервів можуть використовуватися різні методи, включаючи метод прямих збитків, метод коефіцієнтів покриття ризику, метод на основі кредитного скорингу та інші. Рівень резервів може бути періодично переглянутий і змінений залежно від зміни ситуації на ринку та в кредитному портфелі банку.

Це забезпечення зазвичай використовується для покриття різних видів втрат за кредитами, таких як непрацюючі кредити, банкрутство клієнта та переглянуті кредити, за якими сплачуються нижчі, ніж попередньо оцінені платежі.

Банки зобов’язані звітувати про потенційні неповернення кредитів і витрати, щоб переконатися, що вони представляють точну оцінку свого загального фінансового стану.

Резерви на покриття збитків за кредитами – це стандартне бухгалтерське коригування резервів банку на покриття збитків за кредитами, включене у фінансову звітність банків. Резерви на покриття збитків за кредитами постійно створюються для врахування мінливих прогнозів щодо збитків від кредитних продуктів банку. Незважаючи на те, що стандарти кредитування значно покращилися, банки все ще стикаються з простроченнями виплати кредитів і неплатежами.

РОЗДІЛ 2

МАТЕМАТИЧНІ МЕТОДИ ТА МОДЕЛІ ОЦІНЮВАННЯ КРЕДИТНИХ РИЗИКІВ

2.1 Основні методи машинного навчання для оцінки кредитних ризиків

Методи машинного навчання.

Логістична регресія. Логістична регресія є однією з найпоширеніших моделей машинного навчання для прогнозування ймовірності настання події на основі даних про попередні події. Цей метод широко використовується в задачах класифікації, де необхідно визначити, до якого класу належить об'єкт [18].

Формула логістичної регресії (2.1) має такий вигляд:

$$P(y = 1|x) = \frac{1}{1+e^{-\theta^T x}} \quad (2.1)$$

де P(y=1|x) – ймовірність того, що залежна змінна y приймає значення 1 при заданих значеннях незалежних змінних x;

x – вектор незалежних змінних;

θ – вектор параметрів моделі, які необхідно знайти шляхом навчання;

T – оператор транспонування;

e – число Ейлера, основа натурального логарифма.

Для більш ніж двох класів, логістична регресія застосовується з використанням підходу «один проти всіх» (one-vs-all). Тобто, для кожного класу будується своя логістична регресія, яка відрізняється вектором параметрів. Класифікація нового об'єкта проводиться шляхом визначення ймовірності належності до кожного класу і вибору класу з найбільшою ймовірністю.

Цей тип статистичної моделі часто використовується для класифікації та прогнозної аналітики. Логістична регресія оцінює ймовірність події, наприклад повернення кредиту або ж дефолт, на основі заданого набору даних незалежних змінних. Оскільки результат є ймовірністю, залежна змінна обмежена між 0 і 1.

У рамках машинного навчання логістична регресія належить до сімейства керованих моделей машинного навчання. Вона також вважається дискримінаційною моделлю, що означає, що вона намагається розрізняти класи.

Дерева рішень. Дерева рішень – це метод машинного навчання, що використовується для розв'язання задач класифікації та регресії. Вони представляють собою структуру дерева, де кожен вузол відповідає питанню про значення однієї з ознак, а ребра відповідають відповідям на це питання. Кожне листовий вузол дерева представляє класифікаційне рішення або числове значення у випадку регресії [20].

Формула для дерева рішень не має однозначної форми, оскільки цей метод може використовувати різні алгоритми побудови дерева (CART, ID3, C4.5). Однак, загальний принцип побудови дерева полягає в пошуку найбільш інформативної ознаки, яка дозволяє максимально розділити вибірку на чисті класи (у випадку класифікації) або зменшити дисперсію (у випадку регресії).

Шлях для побудови дерева рішень може бути записаним наступним чином:

1. Оберіть найкращий розділ, який максимізує інформаційний приріст чи мінімізує ентропію.
2. Розділіть набір даних на дві підмножини відповідно до значення вибраної умови.
3. Для кожної підмножини повторіть крок 1, поки не буде досягнуто кінцевої точки, наприклад, коли досягнуто задану глибину дерева або коли всі об'єкти у підмножині належать до одного класу.
4. Вузол дерева, який містить всі об'єкти одного класу, вважається листом дерева і є кінцевим рішенням.
5. При класифікації нового об'єкта, використовується створене дерево рішень, рекурсивно перевіряючи умови на шляху до листа, що містить рішення.

Нейронні мережі. Нейронні мережі – це моделі машинного навчання, які намагаються імітувати роботу людського мозку. Вони складаються з нейронів, які з'єднані між собою ваговими зв'язками. Кожен нейрон обробляє вхідні дані та передає їх до наступного нейрона у мережі. У нейронних мережах

використовуються алгоритми зворотного поширення помилок, які навчають модель, виправляючи помилки у вагових зв’язках [21].

Ці моделі машинного навчання дозволяють розпізнавати складні закономірності та залежності у великих вибірках даних та вирішувати завдання класифікації, регресії, сегментації та генерації зображень.

Формально, нейронна мережа складається зі шарів нейронів, де кожен шар передає вихідні дані до наступного шару. Перший шар мережі називається вхідним шаром, а останній – вихідним шаром. Мережа містить вагові коефіцієнти, які використовуються для зваженого сумування вхідних даних.

Методи заповнення пропусків даних

Метод k-найближчих сусідів. Метод k-найближчих сусідів можна використовувати для заповнення пропусків в даних.

Цей метод полягає у виборі k-найближчих сусідів до об’єкту, який має пропущені значення. Якщо пропущене значення є кількома функціями, то можна обчислити середнє арифметичне або медіану значень, залежно від типу даних [22].

Алгоритм заповнення пропущених значень за допомогою методу k-NN можна описати наступними кроками.

- Обирається параметр k - кількість найближчих сусідів, яку потрібно врахувати при заповненні пропущених значень.
- Обчислюється відстань між об’єктом з пропущеним значенням та усіма іншими об’єктами в даних.
- Вибирається k-найближчих сусідів до об’єкта з пропущеним значенням.
- Обчислюється середнє арифметичне або медіана значень знайдених k-найближчих сусідів.
- Заповнюється пропущене значення середнім арифметичним або медіаною.

Цей метод може бути ефективним, якщо дані не містять багато пропущених значень або якщо кількість пропущених значень невелика. Однак, якщо даних

багато, то використання методу k-найближчих сусідів може призвести до великої кількості обчислень, що збільшить час обробки даних.

Метод випадкового лісу. Метод випадкового лісу також можна використовувати для заповнення пропущених значень в даних.

Випадковий ліс – це ансамбль рішучих дерев, де кожне дерево натреноване на підмножині даних із випадковими підмножинами ознак [23].

Для заповнення пропущених значень використовується метод, що називається «out-of-bag» (OOB) – це підхід до оцінки точності моделі, де для кожного прикладу з навчального набору даних випадковий ліс використовує ті дерева, які не містять цей приклад, для прогнозування цієї змінної.

Алгоритм заповнення пропущених значень за допомогою методу випадкового лісу можна описати наступними кроками.

- Прибрати стовпці з пропущеними значеннями від інших стовпців.
- Визначити множину змінних, що не містять пропущених значень, які будуть використані для тренування випадкового лісу.
- Розділити навчальний набір на дві частини: одну з частин будемо використовувати для тренування випадкового лісу, а іншу для використання методу OOB для заповнення пропущених значень.
- Тренуємо випадковий ліс на першій частині даних.
- Використовуємо OOB для заповнення пропущених значень у другій частині даних.
- Повторюємо кроки 4 і 5 декілька разів для отримання середнього значення заповнених пропущених значень.

Цей метод може бути ефективним, якщо в даних є багато пропущених значень, але також може займати більше часу на обробку даних, оскільки випадковий ліс є більш складною моделлю, ніж метод k-найближчих сусідів.

Методи фільтрації даних

Метод фільтрації даних є одним із методів очищення даних, який використовується для видалення шуму та зменшення кількості помилок у даних. Цей метод полягає у використанні різноманітних статистичних та математичних

методів для зменшення або видалення значень, які не належать до корисної інформації.

Існує декілька методів фільтрації даних, де кожен з них підходить до вирішення конкретного типу проблем у даних.

Розглянемо основні методи фільтрації даних.

*Експоненціальний метод (згладжування)***.** Цей метод використовується для видалення шуму в часових рядах. Він полягає у згладжуванні даних за допомогою фільтрів, які видаляють низькочастотні та високочастотні складові.

Експоненціальне згладжування є загальноприйнятою статистичною технікою для прогнозування часових рядів. Зазвичай даний метод використовується для прогнозування даних часових рядів на основі попередніх припущень користувача, таких як сезонність або систематичні тенденції [24].

Метод полягає у розрахунку середнього значення даних з попередніх періодів, а потім у ваговому коефіцієнті, що враховує додаткову вагу останніх значень. Цей коефіцієнт називається параметром згладжування або фактором згладжування і може мати значення від 0 до 1. Чим більше значення параметра згладжування, тим більшу вагу отримує останнє значення, але його точність може зменшуватися. З іншого боку, чим менше параметр згладжування, тим більшу вагу отримують попередні значення, але прогнози майбутніх значень можуть бути менш точними.

Формула експоненціального згладжування (2.2):

$$S(t) = \alpha * Y(t) + (1 - \alpha) * S(t - 1) \qquad (2.2)$$

де S(t) – прогнозоване значення в момент часу t
Y(t) – спостережуване значення в момент часу t
S(t-1) – прогнозоване значення в попередній момент часу
α – параметр згладжування, що враховує вагу останнього спостережуваного значення.

Також є й інші методи фільтрації даних. Ось деякі з них.

Метод медіани. Цей метод використовується для видалення випадкових помилок в даних. Він полягає у заміні значень, що виходять за межі діапазону, на медіанне значення даних.

Метод заміщення. Цей метод використовується для заповнення пропущених значень в даних. Він полягає у виборі значень з найбільш близьких записів до запису з пропущеним значенням та використанні цих значень для заповнення пропуску.

Метод видалення. Цей метод використовується для видалення даних, які не належать до корисної інформації. Наприклад, якщо в даних є декілька записів з однаковими значеннями, то один з цих записів може бути видалений, оскільки він не надає додаткової інформації.

Метод статистичних тестів. Цей метод використовується для видалення неправдоподібних даних.

2.2 Опис моделей прогнозування

Авторегресійна модель. Авторегресійна модель (AR) є одним з методів прогнозування часових рядів. Цей метод базується на ідеї, що майбутні значення часового ряду залежать від попередніх значень ряду [25].

У авторегресійній моделі кожне значення ряду представлене як лінійна комбінація попередніх значень ряду, а також стохастичного складового (шуму). Часто для розрахунку коефіцієнтів авторегресійної (2.3) моделі використовують метод найменших квадратів, який дозволяє знайти оптимальні коефіцієнти, які найкраще описують даний часовий ряд.

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t \qquad (2.3)$$

де ε_t – білий шум. Це множинна регресія з лаговими значеннями y_t в якості предикторів.

За допомогою авторегресійної моделі можна прогнозувати майбутні значення часового ряду на основі попередніх значень. Для цього необхідно використовувати попередні значення ряду, які відомі на момент прогнозування, і

розрахувати прогнозоване значення за допомогою коефіцієнтів авторегресії та стохастичного складового.

Авторегресійна модель є потужним інструментом для прогнозування часових рядів, але вона передбачає, що майбутні значення ряду залежать лише від попередніх значень. Тоді як в реальному світі існують багато інших факторів, що можуть впливати на прогноз.

Модель ковзного середнього. Модель ковзного середнього (MA) є одним з методів прогнозування часових рядів, який базується на згладжуванні шуму та різких змін в часовому ряді [26].

У моделі ковзного середнього кожне значення ряду представлене як середнє значення попередніх значень, тобто використовуються попередні значення ряду, щоб визначити середнє значення на певний момент часу. Модель може мати різну довжину ковзного середнього.

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q} \quad (2.4)$$

де ε_t – білий шум. Це множинна регресія з минулими помилками як предикторами.

За допомогою моделі ковзного середнього можна зменшити вплив шуму на часовий ряд та згладити різкі зміни значень в ряді. Прогнозування майбутніх значень ряду виконується шляхом розрахунку середнього значення попередніх значень.

Однак, модель ковзного середнього має свої обмеження, зокрема вона не підходить для прогнозування складних часових рядів зі складними тенденціями та залежностями між значеннями.

Авторегресійна модель з ковзним середнім. Авторегресійна модель з ковзним середнім (ARMA) – це метод прогнозування часових рядів, який поєднує в собі модель авторегресії (AR) та модель ковзного середнього (MA) [27].

ARMA модель може бути записана у вигляді:

$$y_t = c + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q} + \varepsilon_t \quad (2.5)$$

Прогнози включають як лагові значення y_t, так і лагові помилки.

Для прогнозування значень часового ряду з використанням ARMA моделі необхідно визначити значення коефіцієнтів φ та θ. Це можна зробити за

допомогою методу максимальної правдоподібності або методом найменших квадратів.

ARMA модель може бути корисна для прогнозування рядів зі складними тенденціями та шумом, однак, як і будь-яка модель прогнозування, вона має свої обмеження та певний ризик помилки прогнозування.

Сезонна модель ARIMA. Сезонна модель ARIMA (Autoregressive Integrated Moving Average) є однією з найпоширеніших моделей для прогнозування часових рядів. Вона дозволяє урахувати сезонність у даних і використовується для прогнозування тих часових рядів, які мають чітко виражений сезонний характер [28].

ARIMA складається з трьох компонентів: авторегресійної моделі (AR), інтегрованої моделі (I) та моделі ковзного середнього (MA). Кожен з компонентів може мати різні порядки (позначаються змінними p, d, q, P, D, Q відповідно), що дозволяє моделювати різні типи поведінки часових рядів.

Сезонна модель ARIMA додає до цих трьох компонентів ще один параметр – сезонність (S). Це означає, що ми враховуємо взаємозв'язок між значеннями ряду, які віддалені один від одного на S спостережень. Наприклад, якщо ми аналізуємо щомісячний часовий ряд, то S = 12 (оскільки в одному році 12 місяців), якщо щотижневий ряд – то S = 52 (оскільки в одному році 52 тижні), де m – кількість спостережень на рік (Рис. 2.1)

ARIMA	
Несезонна частина	Сезонна частина
(p, d, q)	$(P, D, Q)_m$

Рисунок 2.1. – Відмінність ARIMA моделей [38]

2.3 Критерії адекватності математичних моделей і якості прогнозів

Інформаційний критерій Акаіке (Akaike Information Criterion, AIC) – це метрика для порівняння різних моделей з однією і тією ж самою залежною

змінною. Він використовується для визначення оптимальної моделі, яка найкраще описує дані. AIC базується на ймовірності та складається з двох частин: перша частина відображає точність моделі, а друга – кількість параметрів, використовуваних в моделі. AIC є компромісом між точністю моделі та складністю [29].

Формула Akaike's Information Criterion (AIC):

$$AIC = -2\log(L) + 2(p + q + k + 1) \tag{2.6}$$

де L – це ймовірність даних;

$k = \begin{cases} 1, & \text{якщо } c \neq 0 \\ 0, & \text{якщо } c = 0 \end{cases}$

Змінений інформаційний критерій Акаіке (Corrected Akaike Information Criterion, AICc) – це виправлений варіант Акаіке критерію інформації (AIC), що враховує збільшення середнього квадратичного відхилення при обмеженій кількості спостережень [30].

Змінений критерій Акаіке використовується для вибору оптимальної моделі з кількох кандидатів. Він базується на тому ж принципі, що і AIC, але додатково враховує кількість параметрів моделі та кількість спостережень.

Формула corrected AIC:

$$AICc = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}. \tag{2.7}$$

Оптимальна модель має найменший AICc, тобто модель, яка найкраще поєднує точність та складність, враховуючи кількість спостережень. Змінений інформаційний критерій Акаіке зазвичай використовується в разі обмеженого числа спостережень для вибору оптимальної моделі.

Інформаційний критерій Байєса (Bayesian Information Criterion, BIC) – це критерій, який використовується для вибору оптимальної моделі з кількох кандидатів. Він базується на ідеї мінімізації критерію максимальної правдоподібності, з урахуванням складності моделі [31].

Формула Bayesian Information Criterion:

$$BIC = AIC + [\log(T) - 2](p + q + k + 1). \tag{2.8}$$

Оптимальна модель має найменше значення критерію BIC. У порівнянні з іншими критеріями, такими як AIC, BIC найсильніший до моделей з меншою

кількістю параметрів. Це означає, що BIC вважає більш складну модель менш бажаною, ніж AIC. Тому, якщо кількість спостережень невелика, BIC може допомогти уникнути перенавчання моделі. Однак, якщо кількість спостережень досить велика, то різниця між BIC і AIC зазвичай є невеликою, і вибір критерію залежить від конкретної задачі. Але загалом краща модель вибирається шляхом мінімізації всіх вище перерахованих критеріїв.

Критерії якості прогнозів. Існує кілька критеріїв якості прогнозування, які використовуються для оцінки точності прогнозів моделей. Одні з найбільш поширених наведено та описано далі.

Середньоквадратична помилка (Mean Squared Error, MSE). Це квадрат різниці між прогнозованою та фактичною значеннями, після чого береться середнє значення помилки. Цей критерій добре підходить для моделей, які передбачають неперервні величини [32].

Формула для обчислення MSE:

$$MSE = \left(\frac{1}{n}\right) * \sum(y_i - \hat{y}_i)^2, \quad (2.9)$$

де n – кількість спостережень;

y_i *– фактичне значення;*

$\hat{y}_i$ *– прогнозоване значення.*

Значення MSE може бути від 0 до нескінченності, і чим менше значення MSE, тим краще прогноз. MSE має одиницю виміру, квадрат від одиниці виміру фактичного значення, що може бути корисним для порівняння з іншими критеріями.

Один з недоліків MSE полягає в тому, що він відреагує на великі помилки більш сильно, ніж на менші помилки. Тому у випадку, коли є викиди або великі помилки, може бути краще використовувати інші критерії, такі як середня абсолютна помилка (MAE) або коефіцієнт детермінації (R^2).

Середня абсолютна помилка (Mean Absolute Error, MAE). Це абсолютна різниця між прогнозованою та фактичною значеннями, після чого береться середнє значення помилки. Цей критерій також добре підходить для моделей, які передбачають неперервні величини [33].

Формула для розрахунку MAE виглядає так:

$$MAE = \left(\frac{1}{n}\right) * \sum|y - \hat{y}|, \quad (2.10)$$

де n – кількість спостережень;

y – спостережувані значення;

ŷ – прогнозовані значення.

Значення MAE може бути від 0 до нескінченності, і чим менше значення MAE, тим краще прогноз. MAE є більш стійкою, ніж середньоквадратична помилка (MSE), оскільки вона не залежить від великих відхилень від середнього значення і може бути корисною у випадку, коли великі помилки прогнозування можуть бути небажаними.

Проте MAE має один недолік – вона недиференційована в нулі, що робить її менш корисною для оптимізації функції в рамках навчання моделі.

Коефіцієнт детермінації (Coefficient of Determination, R-squared). Цей критерій вимірює відсоток дисперсії у фактичних значеннях, який може бути пояснений прогнозною моделлю. Він зазвичай використовується для оцінки точності моделей, які передбачають неперервні величини [34].

Коефіцієнт детермінації можна обчислити за наступною формулою:

$$R^2 = 1 - \left(\frac{SSR}{SST}\right), \quad (2.11)$$

де SSR – сума квадратів помилок (різниця між фактичними та прогнозованими значеннями);

SST – загальна сума квадратів (різниця між фактичними значеннями та середнім значенням).

Коефіцієнт детермінації, як й інші коефіцієнти, може приймати значення від 0 до 1. Значення близьке до 1 означає, що модель добре підходить для даних і висока частка дисперсії пояснюється моделлю. Значення близьке до 0 означає, що модель погано підходить для даних і низька частка дисперсії пояснюється моделлю.

РОЗДІЛ 3

КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ КРЕДИТНОГО РИЗИКУ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

3.1 Огляд та вибір програмного забезпечення щодо оцінювання кредитних ризиків

MATLAB. MATLAB (MATrix LABoratory) – це інтерактивна програмна система для чисельних обчислень та створення технічних обчислювальних додатків. Вона використовується в наукових дослідженнях, інженерній практиці, аналізі даних та інших галузях.

MATLAB має велику кількість вбудованих функцій для роботи з матрицями, числовими обчисленнями, графічним відображенням даних, аналізом та обробкою сигналів, оптимізацією, статистичним аналізом та багато іншого. MATLAB також має вбудовану мову програмування, що дозволяє користувачам створювати власні функції та програми [35].

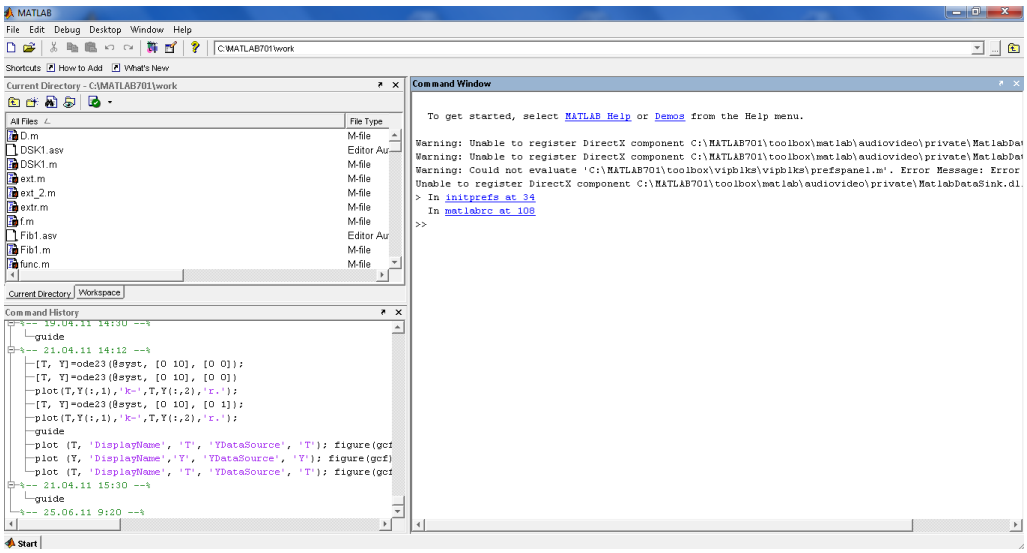

Рисунок 3.1 – Інтерфейс MATLAB

Переваги MATLAB:

- Широкий функціонал. MATLAB має вбудований набір функцій для чисельних обчислень, роботи з графікою, статистики, машинного навчання та багато іншого. Це робить його дуже потужним інструментом для аналізу даних.
- Простота використання. MATLAB має простий інтерфейс, який дозволяє легко і швидко створювати і виконувати програми.
- Широке співтовариство. MATLAB має велику спільноту користувачів, що дозволяє швидко знайти відповіді на питання, отримати поради та розв’язати проблеми.
- Великий вибір пакетів. MATLAB має великий вибір додаткових пакетів, що дозволяє користувачам розширювати функціональність програмного забезпечення.

Недоліки MATLAB:

- Вартість. MATLAB є платним програмним забезпеченням, що може бути дорогим для студентів та маленьких компаній.
- Обмеження ліцензії. MATLAB має обмеження на кількість користувачів, які можуть використовувати програмне забезпечення одночасно.
- Проблеми з пам’яттю. MATLAB може використовувати значну кількість пам’яті при обробці великих обсягів даних.
- Не оптимізовано для обробки великих обсягів даних. MATLAB може бути повільним при обробці великих обсягів даних через обмежену пам’ять та процесорну потужність.

Мова програмування R. R – це мова програмування та середовище для статистичного аналізу та візуалізації даних. Вона є безкоштовною та відкритою для використання, що дозволяє користувачам розробляти власні функції та пакети для аналізу даних. R є дуже популярним серед статистиків та дослідників у багатьох галузях, включаючи біологію, фізику, економіку та соціологію [36].

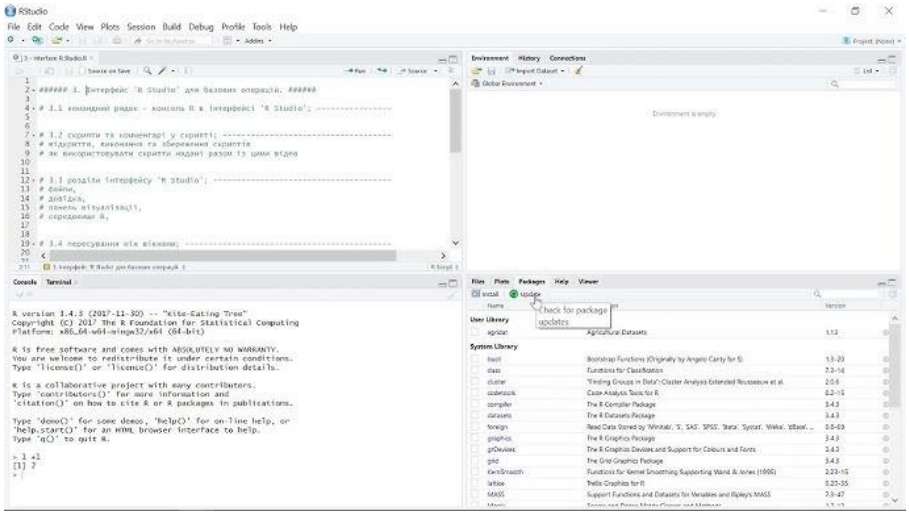

Рисунок 3.2. – Інтерфейс R

Переваги R:

- Відкрите програмне забезпечення, що дозволяє безкоштовно використовувати його для будь-яких цілей.
- Широкий спектр статистичних та машинного навчання функцій та бібліотек, які дозволяють виконувати різноманітні аналітичні та наукові дослідження.
- Наявність великої спільноти користувачів та розробників, що забезпечує підтримку, розвиток та поширення R.
- Легка інтеграція з іншими мовами програмування, зокрема з Python та C++.

Недоліки R:

- Не підходить для великих даних та обробки даних в реальному часі.
- Передбачувана швидкість роботи може бути нижчою порівняно з іншими мовами програмування, зокрема з Python.
- Інтерфейс користувача має обмежені можливості порівняно з комерційними аналітичними програмами, такими як SAS та SPSS.
- Не має додаткових функцій, які можна було б використовувати в комерційних цілях.

SAS. SAS (Statistical Analysis System) – це програмне забезпечення, призначене для проведення статистичного аналізу даних, бізнес-аналізу, біоінформатики та інших досліджень. Основна мова програмування SAS – SAS

Programming Language, який зазвичай використовується для створення, збереження, збору та відображення даних [37].

Рисунок 3.3. – Інтерфейс SAS

Основні переваги SAS:

- Велика кількість вбудованих функцій і процедур, що дозволяє ефективно працювати з даними в багатьох областях;
- Висока надійність та безпека даних, що особливо важливо в бізнес-сфері;
- Наявність спеціалізованих модулів для різних індустрій, таких як фінанси, фармація, біоінформатика;
- Простота використання при роботі з великими обсягами даних;
- Забезпечення високої швидкості обробки даних за рахунок оптимізованих процедур і функцій.

Основні недоліки SAS:

- Висока вартість ліцензії;
- Обмежена можливість зміни та розширення функціональності програмного забезпечення;
- Відсутність безкоштовної версії для некомерційного використання, що обмежує доступність для студентів та дослідників.

Python. Python – це високорівнева мова програмування загального призначення з великою кількістю бібліотек та інструментів для обробки даних, машинного навчання, візуалізації, статистичного аналізу та багато іншого [38].

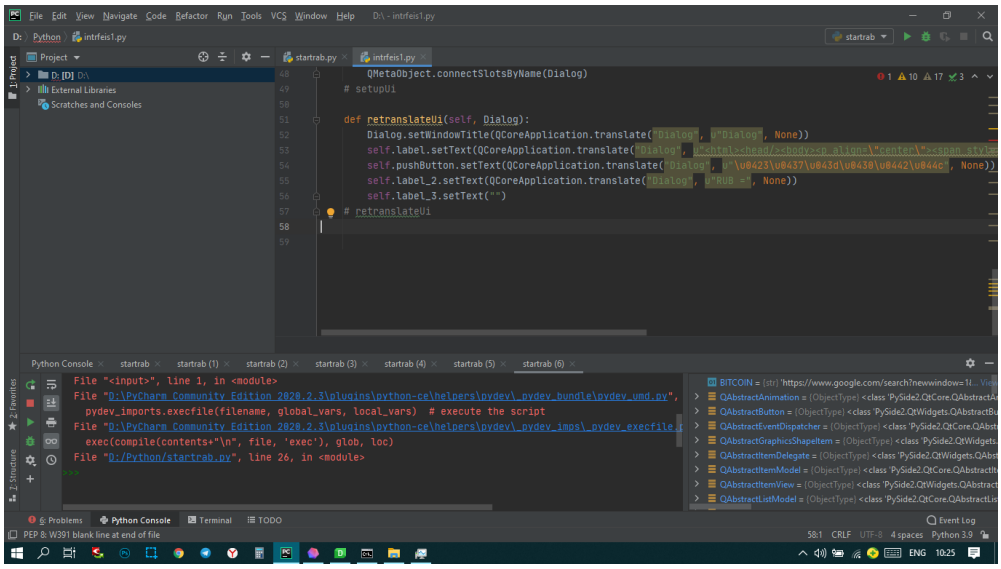

Рисунок 3.4. – Інтерфейс мови Python у PyCharm

Переваги Python:

- Легка читаність і простота синтаксису
- Велика кількість бібліотек, модулів та фреймворків для різних задач машинного навчання та наукових обчислень
- Відкритий код і велике співтовариство розробників, що постійно додає нові функції і можливості
- Підтримка багатьох платформ, таких як Windows, Linux і Mac, та можливість легко інтегрувати з іншими мовами програмування
- Широкі можливості для візуалізації даних та створення графіків

Недоліки Python:

- Порівняно повільна швидкість в порівнянні з компільованими мовами програмування, такими як C++ або Java
- Не підходить для розробки додатків в реальному часі, таких як вбудовані системи або операційні системи
- Обмежені можливості для паралельного програмування і оптимізації пам’яті
- Не має нативної підтримки для розробки мобільних додатків.

Вибір програмного забезпечення для проведення дослідження.

Перш ніж прийняти рішення щодо вибору програмного забезпечення для мого дослідження, я ретельно оглянула основні програмні інструменти, які використовуються для аналізу даних. Та після їх докладного аналізу, було прийнято рішення використовувати Python. Нижче наведені основні аргументи, які підтримують мій вибір:

Відкрите джерело. Python є вільно розповсюджуваним програмним забезпеченням з відкритим вихідним кодом. Це означає, що ми можемо безкоштовно використовувати Python і мати можливість модифікувати і розповсюджувати його за необхідності. Відкритий характер Python створює широкі можливості співпраці та спільної розробки з іншими дослідниками та програмістами.

Багата екосистема. Тут йдеться мова про те, що Python включає багато бібліотек і модулів для аналізу даних, машинного навчання, візуалізації і статистики. Наприклад, бібліотеки такі як NumPy, Pandas, Matplotlib і Scikit-learn надають потужні інструменти для маніпулювання даними та розв’язання складних завдань аналізу. До речі ці бібліотеки я саме й використовувала на практиці для свого дослідження.

Простота використання. Простий і зрозумілий синтаксис Python, робить його доступним для різних категорій користувачів, незалежно від їхнього рівня експертизи в програмуванні. Це робить його популярним вибором серед дослідників, які можуть бути менш орієнтовані на програмування, але мають потребу в аналізі даних.

Інтеграція з іншими мовами та інструментами. Python може використовуватися як основна мова програмування або як мова скриптування для інших програмних забезпечень. Він має добре розроблені інтерфейси для взаємодії з іншими мовами, такими як C++, Java і R, що дозволяє використовувати найкращі інструменти з різних середовищ.

Широке співтовариство. Велику та активну спільноту користувачів, яка постійно розвивається і підтримує різноманітні проекти звісно це про Python. А

це означає, що завжди є можливість знайти допомогу, документацію, навчальні матеріали та приклади коду, що значно полегшує роботу і дозволяє швидко реалізувати свої дослідження.

Отже, на основі відкритого джерела Python, багатофункціональної екосистеми, простоти використання, здатності до інтеграції та широкої спільноти користувачів, я прийняла рішення використовувати Python як основний інструмент для проведення свого дослідження.

3.2 Аналіз та підготовка статистичної бази

Для цього дослідження було обрано наступні дані:

ID – ID кожного клієнта (унікальне значення);

LIMIT_BAL – сума виданого кредиту;

SEX – стать (1-чоловіча; 2-жіноча)

EDUCATION – освіта (1-коледж; 2-університет; 3-школа; 4-інше; 5 та 6-невідомо)

MARRIAGE – сімейний стан (1-одружений; 2-неодружений; 3-інше)

AGE – вік у роках

PAY_0 – статус погашення за 04.2023 (-1-вчасні платежі; 1-затримка на 1 місяць; 2-затримка на 2 місяці; …; 8-затримка на 8 місяців, 9-затримка на 9 місяців і більше)

PAY_2 – статус погашення за 03.2023 (-1- вчасні платежі; 1-затримка на 1 місяць; 2-затримка на 2 місяці; …; 8-затримка на 8 місяців, 9-затримка на 9 місяців і більше)

PAY_3 – статус погашення за 02.2023 (-1- вчасні платежі; 1-затримка на 1 місяць; 2-затримка на 2 місяці; …; 8-затримка на 8 місяців, 9-затримка на 9 місяців і більше)

PAY_4 – статус погашення за 01.2023 (-1- вчасні платежі; 1-затримка на 1 місяць; 2-затримка на 2 місяці; …; 8-затримка на 8 місяців, 9-затримка на 9 місяців і більше)

PAY_5 – статус погашення за 12.2022 (-1- вчасні платежі; 1-затримка на 1 місяць; 2-затримка на 2 місяці; …; 8-затримка на 8 місяців, 9-затримка на 9 місяців і більше)

PAY_6 – статус погашення за 11.2022 (-1- вчасні платежі; 1-затримка на 1 місяць; 2-затримка на 2 місяці; …; 8-затримка на 8 місяців, 9-затримка на 9 місяців і більше)

BILL_AMT1 – сума виписки за рахунком за 04.2023

BILL_AMT2 – сума виписки за рахунком за 03.2023

BILL_AMT3 – сума виписки за рахунком за 02.2023

BILL_AMT4 – сума виписки за рахунком за 01.2023

BILL_AMT5 – сума виписки за рахунком за 12.2022

BILL_AMT6 – сума виписки за рахунком за 11.2022

PAY_AMT1 – сума середнього платежу за 04.2023

PAY_AMT2 – сума середнього платежу за 03.2023

PAY_AMT3 – сума середнього платежу за 02.2023

PAY_AMT4 – сума середнього платежу за 01.2023

PAY_AMT5 – сума середнього платежу за 12.2022

PAY_AMT6 – сума середнього платежу за 11.2022

default.payment.next.month – дефолт (1-так; 0-ні)

Для повного дослідження необхідно проаналізувати три наступні моделі та обрати найкращу з них. Тут я використаю такі моделі: логістичну регресію, метод k-найближчих сусідів та метод опорних векторів.

3.3 Оцінювання кредитних ризиків та порівняльний аналіз отриманих результатів

Аналіз вхідних даних. Перше, що потрібно зробити з вибраними даними це провести перевірку на пропущені значення. Це необхідно для забезпечення якості, точності та надійності аналізу та моделювання даних.

```
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  float64
 2   SEX                         30000 non-null  int64
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  int64
 5   AGE                         30000 non-null  int64
 6   PAY_0                       30000 non-null  int64
 7   PAY_2                       30000 non-null  int64
 8   PAY_3                       30000 non-null  int64
 9   PAY_4                       30000 non-null  int64
 10  PAY_5                       30000 non-null  int64
 11  PAY_6                       30000 non-null  int64
 12  BILL_AMT1                   30000 non-null  float64
 13  BILL_AMT2                   30000 non-null  float64
 14  BILL_AMT3                   30000 non-null  float64
 15  BILL_AMT4                   30000 non-null  float64
 16  BILL_AMT5                   30000 non-null  float64
 17  BILL_AMT6                   30000 non-null  float64
 18  PAY_AMT1                    30000 non-null  float64
 19  PAY_AMT2                    30000 non-null  float64
 20  PAY_AMT3                    30000 non-null  float64
 21  PAY_AMT4                    30000 non-null  float64
 22  PAY_AMT5                    30000 non-null  float64
 23  PAY_AMT6                    30000 non-null  float64
 24  default.payment.next.month  30000 non-null  int64
dtypes: float64(13), int64(12)
```

Рисунок 3.5. – Перевірка на пропущенні значення

Джерело: розроблено автором.

Наступним моїм кроком все ж було видалення деяких змінних. Для того щоб не перенавантажувати вибірку даних, потрібно залишити лише необхідні для аналізу та моделювання змінні. Тож зрештою було видалено стовпець «ID», який відповідає за унікальність в наборі даних.

Наступним кроком була перевірка всіх інших змінних. Так як змінні можуть бути не пустими, але містити в собі некоректну інформацію, то варто це перевірити. Таким чином, було знайдено такі викиди в стовпцях «EDUCATION» та «MARRIAGE». І в цій ситуації можна розглянути два способи вирішення

проблеми. Перший варіант полягає у видаленні невідомих значень і тим самим зменшенням вибірки. Цього можна уникнути, якщо вибрати інший варіант, який полягає у заміні невідомих значень на найбільш відповідне та логічне значення. Саме цим способом я й скористалася. Цю заміну було проведено лише в двох змінних, оскільки в інших змінних не було такої потреби й вони виглядали зовсім коректно.

На виході отримано наступні дані:

Таблиця. 3.1. – Кількість позичальників за освітою

EDUCATION	COUNT
1 – Коледж	10 930
2 – Університет	14 030
3 – Школа	4 917
4 – Інше	123

Джерело: розроблено автором.

Таблиця. 3.2. – Кількість позичальників за статусом сім'ї

MARRIAGE	COUNT
1 – Одружений	13 659
2 – Неодружений	15 964
3 – Інше	377

Джерело: розроблено автором.

Таблиця. 3.3. – Кількість позичальників за статтю

SEX	COUNT
1 – Чоловік	11 888
2 – Жінка	18 112

Джерело: розроблено автором.

Таблиця. 3.4. – Кількість позичальників за дефолтним

DEFAULT	COUNT
0 – Ні	23 364
1 – Так	6 636

Джерело: розроблено автором.

Наступним кроком після перевірки на пропуски є перевірка основної описової статистики даних. Описова статистика надає загальний огляд характеристик даних, що дозволяє зрозуміти їх розподіл, центральні тенденції та

розкид значень. Вона допомагає виявити аномалії, викиди, випадковість або зв’язки між змінними.

Основні характеристики, які перевіряються в описовій статистиці, включають:

- середнє значення – це міра центральної тенденції, яка вказує на "типове" значення в наборі даних. Розраховується як сума всіх значень поділена на кількість спостережень. Середнє значення може бути корисним для загального розуміння рівня або тренду змінної.
- медіана – це значення, яке розбиває набір даних на дві рівні частини: 50% спостережень знаходяться вище, а інші 50% - нижче. Вона використовується для оцінки центральної тенденції, особливо коли дані мають викиди або неправильний розподіл.
- мінімум і максимум – вони вказують на найменше і найбільше значення в наборі даних, допомагаючи оцінити діапазон значень.
- стандартне відхилення – це міра розкиду даних навколо середнього значення. Вона вказує на те, наскільки значення розподілені відносно середнього значення. Більше стандартне відхилення вказує на більший розкид даних, тоді як менше стандартне відхилення означає менший розкид.
- квантилі – дозволяють визначити значення, яке розбиває набір даних на певні відсоткові частини. Наприклад, 25-й квантиль (перший квартиль) розбиває дані на дві рівні частини, де 25% спостережень знаходяться вище, а решта 75% – нижче.
- загальна кількість спостережень – вона вказує на загальну кількість даних, які доступні для аналізу. Це дозволяє зрозуміти, наскільки представний є набір даних і яка кількість спостережень може бути виключена з аналізу.

Перевірка основної описової статистики даних допомагає отримати загальний уявлення про їх характеристики та зрозуміти їх особливості перед подальшим аналізом, включаючи побудову моделей, виявлення залежностей та прийняття рішень.

Тож переглянувши аналіз даної вибірки можна зробити певні висновки щодо клієнтів в даному випадку позичальників. Вибірка має досить широкий розкид даних за віком від 21 року до 79 років, з стандартним відхиленням майже у 10 років.

У висновку попереднього дослідження було вирішено відслідкувати дані на аномальні значення. Іншими словами аномальні значення, також відомі як викиди, є відхиленнями від очікуваного розподілу даних або значень, які видаються надзвичайно великими або малими порівняно з іншими спостереженнями.

Основні причини виникнення аномальних значень можуть бути помилки при введенні даних, випадкові помилки, аномальні події або реальні викиди, які можуть бути важливими для розуміння даних. Важливо відслідковувати аномальні значення, оскільки вони можуть спотворити результати аналізу та вплинути на правильність висновків і прийняття рішень.

Для цього дослідження було вибрано метод діаграми розмаху (Рис. 3.10). Вона дозволяє візуалізувати дані за допомогою їх квантилів.

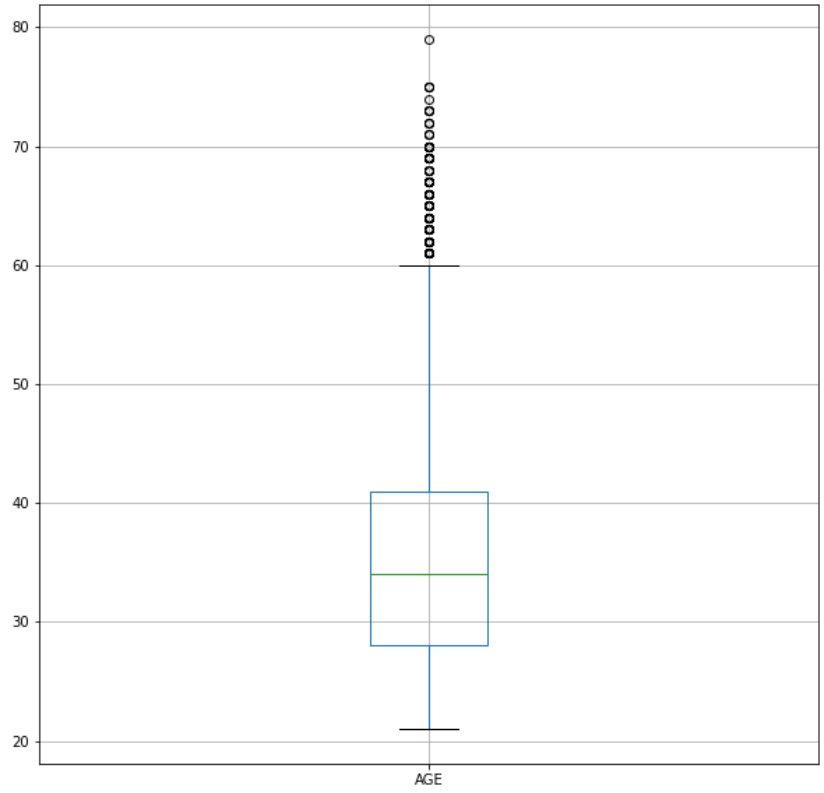

Рисунок 3.10. – Діаграма розмаху за віком позичальників

Джерело: розроблено автором.

Тож на даний момент можна спостерігати, що маємо певні викиди від 60 років до 79 років. Оскільки ми працюємо над загальним методом оцінки кредитних ризиків, то в даному випадку клієнти можуть бути різної вікової категорії і саме зараз не має сенсу видаляти ці аномальні значення. Але також для певного різноманіття цієї ситуації можна було б розділити на дві вибірки, перша була б з віковою категорією від 21 року до 60 років, інша б була з віковою категорією від 61 року до 79 років. І також можна оцінювати ці дві вибірки. Але я обрала не видозмінювати наразі нічого.

Розвідувальний аналіз та візуалізація вхідних даних. Наступним етапом дослідження даних є розвідковий аналіз, що супроводжується візуалізацією. Розвідковий аналіз даних допомагає отримати загальне розуміння даних, виявити шаблони, залежності, тенденції та потенційні висновки. Візуалізація, у свою чергу, є потужним інструментом для представлення даних у вигляді графіків, діаграм, дашбордів та інших візуальних форматів, що допомагають сприйняти й зрозуміти інформацію.

Перше на що хотілося глянути, так це на гістограму дефолту (рис. 3.11). На ній видно що приблизно від 7 000 до 8 000 позичальників, не виплатять кредит чи певну суму кредиту в наступному місяці. Тому що вони вже складають основну частину дефолтного портфеля компанії.

Враховуючи те, що близько від 23 000 позичальників будуть сплачувати позику в наступному місяці не є поганим результатом, але ми прагнемо мінімізувати саме таких проблемних клієнтів. Щоб дефолтний портфель не складався з 8 000 позичальників, а мав розмір хоча б до 4 000 позичальників.

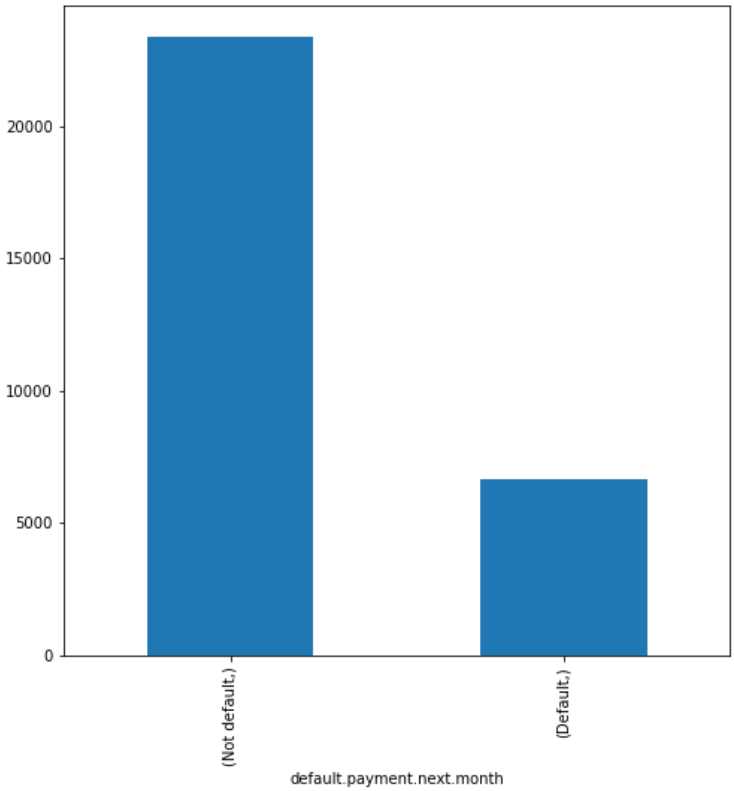

Рисунок 3.11. – Гістограма за дефолтом позичальників

Джерело: розроблено автором.

Також було цікаво переглянути чи є якась залежність між іншими змінними з переглянутою вище. Тому далі будуть описані змінні «EDUCATION» (рис. 3.12.) та «SEX» (рис. 3.13.) в розрізі дефолту наступного місяця.

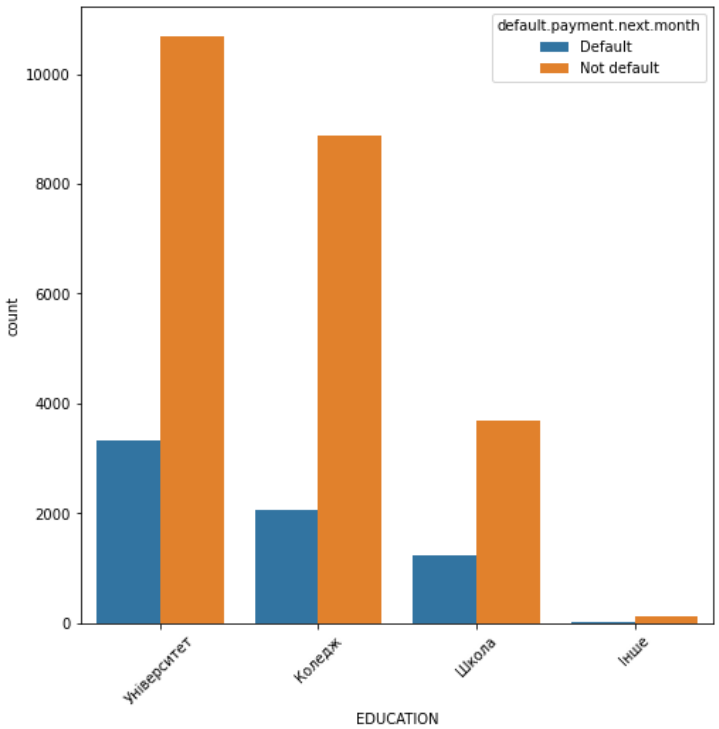

Рисунок 3.12. – Гістограма за освітою позичальників в розрізі дефолту

Джерело: розроблено автором.

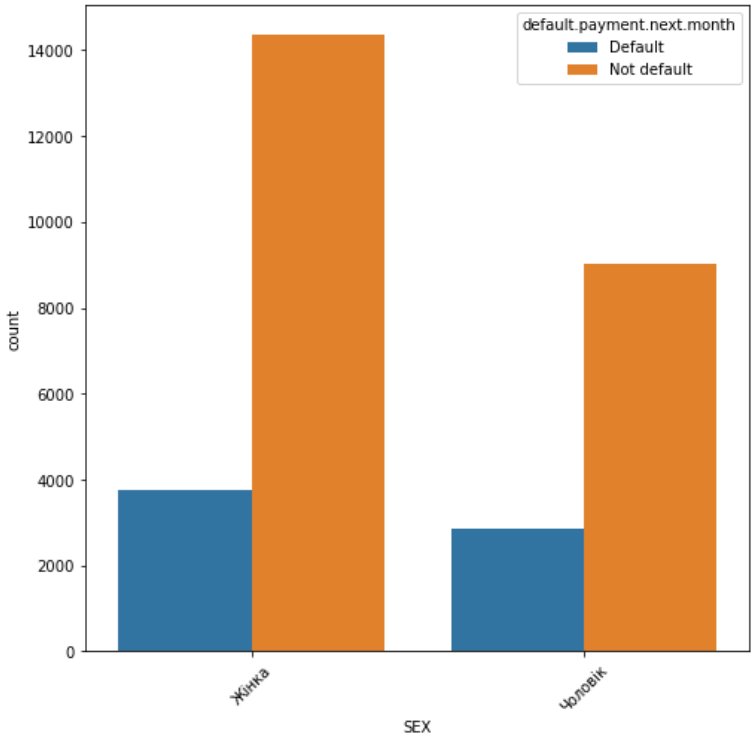

Рисунок 3.13. – Гістограма за статтю позичальників в розрізі дефолту

Джерело: розроблено автором.

Переглянувши побудовані гістограми можемо зробити висновок, що конкретних відхилень від норми немає. Та все ж оцінити дефолтних клієнтів можна. Якщо переглядати змінну «EDUCATION», то чітко видно, що найбільше дефолтних клієнтів з освітою в університеті. Але так само можна відповісти і в іншу сторону. Найбільше недефолтних клієнтів з освітою в унівеситеті. Все тому що найбільша кількість позичальників мають освіту в університеті.

Теж саме можна відповісти про змінну «SEX». Найбільше платять жінки, і також найбільше не платять жінки. І також тому що вони мають перевагу в кількості у даній вибірці.

Ще цікавим спостереженням була оцінка віку клієнта (рис.3.14). За графіком видно найбільш поширений вік, що становить 30 років. За ним також йдуть клієнти з віком: 35 років, 41 рік та 29 років.

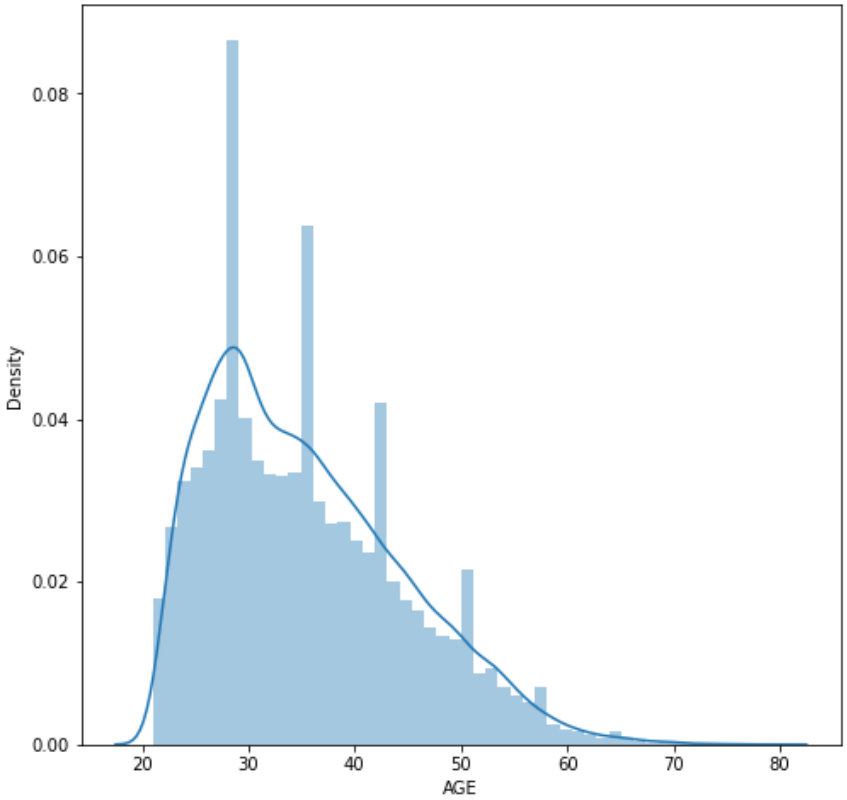

Рисунок 3.14. – Графік віку позичальника (вісь X – щільність, вісь Y – вік)

Джерело: розроблено автором.

Для наступного дослідження необхідно створити матрицю незалежних змінних та вектор залежної змінної. Це важливий крок у багатьох аналітичних методах, таких як регресія, класифікація, кластеризація та інші.

Матриця незалежних змінних складається з різних змінних, які вибрані для аналізу. Кожний стовпчик матриці представляє окрему змінну, таку як вік, стать, дохід, освіта тощо.

Вектор залежної змінної, який також іноді називають цільовою змінною або змінною відгуку, представляє змінну, яку ми хочемо передбачити, пояснити або класифікувати за допомогою аналітичних методів. Тож, якщо ми хочемо передбачити вихід клієнта у дефолтний портфель, то ціна буде нашим вектором залежної змінної, а різні характеристики клієнтів стануть змінними незалежними.

Створення матриці незалежних змінних та вектора залежної змінної допомагає організувати дані для подальшого застосування аналітичних методів.

Логістична регресія. Розглянемо логістичну регресію, яка є потужним інструментом для моделювання й передбачення категоріальних змінних, з використанням візуалізації матриці помилок.

Після створення матриці незалежних змінних та вектора залежної змінної, ми можемо застосувати логістичну регресію. Цей алгоритм будує математичну модель, яка оцінює вплив кожної змінної на ймовірність належності до певного класу.

Одним із способів оцінки ефективності моделі логістичної регресії є використання матриці помилок (рис 3.15.). Вона відображає реальні та передбачені класи спостережень.

Таблиця. 3.5. – Матриця помилок [39]

TN	FP
FN	TP

Де TP вказує на правильно передбачені позитивні класи, FP - на неправильно передбачені позитивні класи, TN - на правильно передбачені негативні класи, а FN - на неправильно передбачені негативні класи.

Візуалізація матриці помилок допомагає нам зрозуміти, наскільки добре модель працює у визначенні класів і може надати важливу інформацію для подальшого вдосконалення моделі.

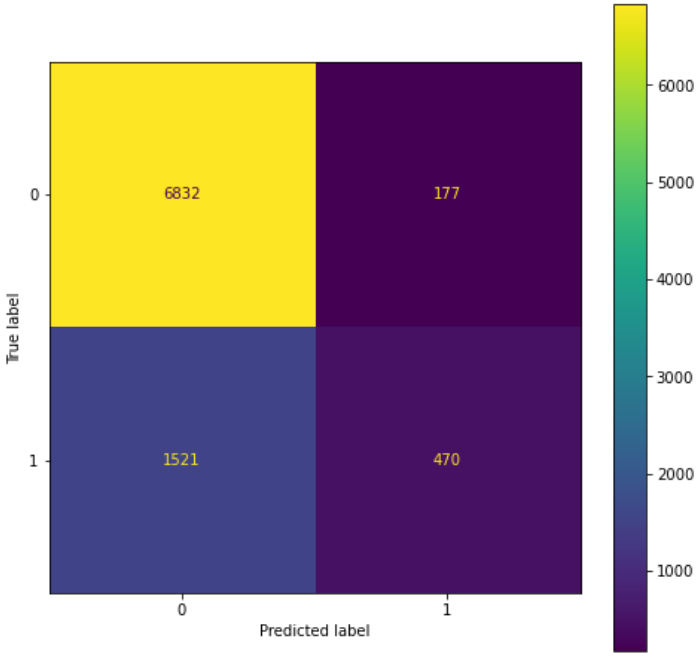

Рисунок 3.15. – Матриця помилок

Джерело: розроблено автором.

Щоб правильно оцінити продуктивність даної моделі виділимо чотири важливі критерії:

- accuracy (точність моделі) – представляє кількість правильно класифікованих екземплярів даних від загальної кількості та має наступну формулу $\frac{TN+TP}{TN+FP+TP+FN}$;
- precision (точність) – представляє показник позитивного значення у класифікації екземплярів даних та має наступну формулу: $\frac{TP}{TP+FP}$;
- recall (запам'ятовування) – представляє справжній позитивний рівень класифікації екземплярів даних та має наступну формулу: $\frac{TP}{TP+FN}$;
- f1 score (оцінка f1) – враховує як показник позитивного значення, так і показник відкликання та має наступну формулу: $2 * \frac{precision * recall}{precision + recall}$

За результатами логістичної регресії отримали, що точність класифікації є досить високою – 81%. Це означає, що модель здатна правильно класифікувати більшість спостережень у відповідні класи.

Продуктивність класифікації є одним із показників ефективності моделі. Вона визначається як відношення кількості правильно класифікованих спостережень до загальної кількості спостережень. Чим більше точність, тим краще модель.

Однак, наряду з точністю, важливо також оцінювати інші показники ефективності моделі, такі як чутливість, специфічність, точність і F-міра. Ці показники дають більш повну картину про те, як добре модель розпізнає певний клас та як уникнути помилкових класифікацій.

Тож далі ми оцінимо інші показники, що важливо врахувати при дослідженні даної моделі.

	precision	recall	f1-score	support
0	0.82	0.97	0.89	7009
1	0.73	0.24	0.36	1991
accuracy			0.81	9000
macro avg	0.77	0.61	0.62	9000
weighted avg	0.80	0.81	0.77	9000

Рисунок 3.16. – Оцінка якості класифікатора

Джерело: розроблено автором.

Як бачимо показники не дефолтного портфеля досить вагомі, тож я підстави що його оцінено добре та коректно. Щодо дефолтного портфеля, то про його так не можна сказати. Адже показники досить низькі та не досягають навіть й 50%.

Метод найближчих сусідів. Наступним розглянемо метод найближчих сусідів, який є простим, але ефективним алгоритмом класифікації та регресії в аналізі даних.

Після виконання методу аналізу найближчих сусідів ми можемо оцінити ефективність моделі, використовуючи різні метрики, такі як точність, чутливість, специфічність тощо, а також візуалізувати результати для кращого розуміння. Тобто все так само як і в логістичній регресії. Спочатку візуалізуємо матрицю помилок, а далі знаходимо інші показники та оцінюємо роботу цього метода.

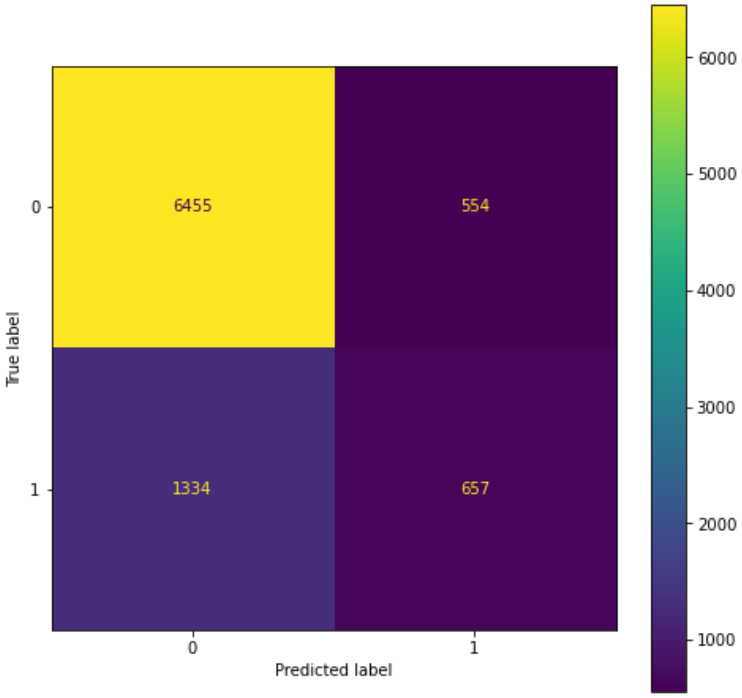

Рисунок 3.17. – Матриця помилок

Джерело: розроблено автором.

Вже по матриці помилок бачимо, що модель найближчих сусідів спрацювала гірше та має трішки гірші показники від логістичної регресії. А саме точність класифікації становить 79%. Порівнюючи з попередніми 81%, різниця ніби незначна та все ж є. Переглянемо інші показники. Але тут вже маємо трішки кращі показники оцінки дефолтного портфеля та все ж деякі з них також не доходять позначки в 50%.

	precision	recall	f1-score	support
0	0.83	0.92	0.87	7009
1	0.54	0.33	0.41	1991
accuracy			0.79	9000
macro avg	0.69	0.63	0.64	9000
weighted avg	0.77	0.79	0.77	9000

Рисунок 3.18. – Оцінка якості класифікатора

Джерело: розроблено автором.

Тож робимо висновок, що на даний момент кращою є логістична регресія.

Метод опорних векторів. Останній метод що ми розглянемо це метод опорних векторів, який є потужним і широко використовуваним алгоритмом для класифікації та регресії в аналізі даних.

Метод опорних векторів базується на ідеї знаходження оптимального гіперплощинного розділяючого простору між класами даних. Головна мета SVM – знайти гіперплощину, яка максимізує відстань між найближчими точками кожного класу, відомими як опорні вектори.

При оцінці результатів SVM можна використовувати різні метрики, такі як точність, чутливість, специфічність, а також використовувати візуалізацію для кращого розуміння та інтерпретації результатів. Тобто використаємо ті ж самі метрики, що і в попередніх методах.

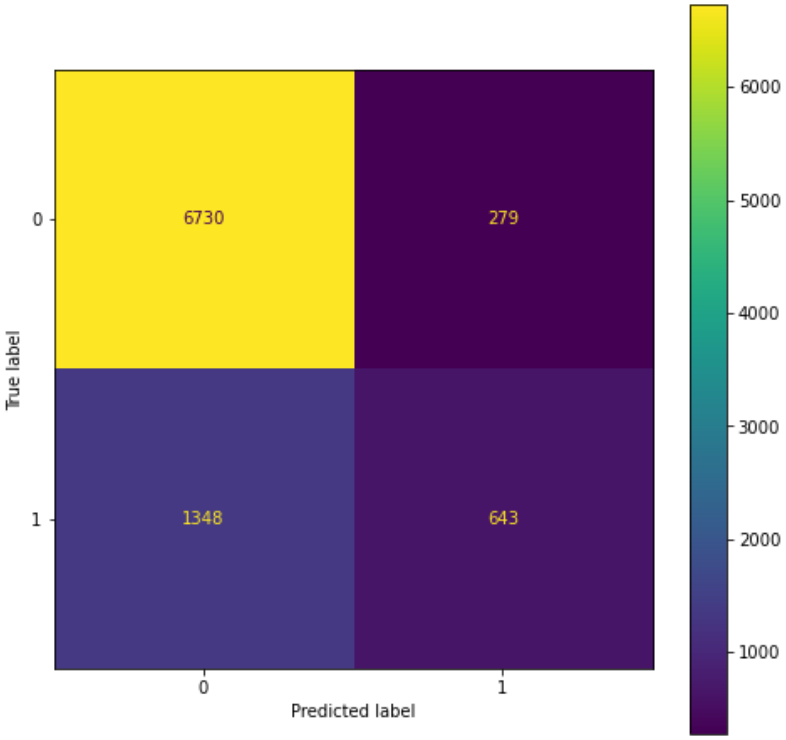

Рисунок 3.19. – Матриця помилок

Джерело: розроблено автором.

При перегляді результатів цього метода, можна звернути на покращення результатів точності. Вони складають 82% й це також дуже добре. Щодо інших показників, то вони також є досить не поганими, а особливо ті що не є в дефолтному портфелі. А щодо дефолтного портфеля то все без змін. Цілком правильно його оцінити не вийде.

	precision	recall	f1-score	support
0	0.83	0.96	0.89	7009
1	0.70	0.32	0.44	1991
accuracy			0.82	9000
macro avg	0.77	0.64	0.67	9000
weighted avg	0.80	0.82	0.79	9000

Рисунок 3.20. – Оцінка якості класифікатора

Джерело: розроблено автором.

На даний момент оцінки кредитного ризику ця модель є найбільш точною.

Результати дослідження.

Щодо висновків, оцінки кредитного ризику вище названими методами. Ми маємо наступні результати:

Таблиця. 3.6. – Показники логістичної регресії

Метрика	Змінна	%
accurancy (точність моделі)		81
precision (точність)	0	82
	1	73
recall (запам’ятовування)	0	97
	1	24
f1-score (оцінка f1)	0	89
	1	36

Джерело: розроблено автором.

Таблиця. 3.7. – Показники методу найближчих сусідів

Метрика	Змінна	%
accurancy (точність моделі)		79
precision (точність)	0	83
	1	54
recall (запам’ятовування)	0	92
	1	33
f1-score (оцінка f1)	0	87
	1	41

Джерело: розроблено автором.

Таблиця. 3.8. – Показники методу опорних векторів

Метрика	Змінна	%
accurancy (точність моделі)		82
precision (точність)	0	83
	1	70
recall (запам’ятовування)	0	96
	1	32
f1-score (оцінка f1)	0	89
	1	44

Джерело: розроблено автором.

На основі цих показників ми можемо зробити висновок, що найкращими методами для даного аналізу є логістична регресія та метод опорних квадратів. Але порівнюючи їхні значення спостерігаємо деякі відмінності. Й роблячи висновок з проведених досліджень, можна стверджувати, що метод опорних векторів є кращою моделлю порівняно з іншими розглянутими методами, такими як логістична регресія та метод найближчих сусідів.

Основними причинами, чому метод опорних векторів може бути визнаний кращою моделлю, включають його здатність працювати з даними високих розмірностей, загальну здатність та ефективність у вирішенні завдань класифікації та регресії. Метод опорних векторів також може ефективно працювати навіть у випадках, коли дані не є повністю роздільними лінійно.

Зважаючи на вищезазначені переваги та успішні результати, отримані за допомогою методу опорних векторів, можна зробити висновок, що краща модель для нашої задачі аналізу даних є саме та що використовує даний метод. Однак, враховуючи контекст застосування та особливості даних, важливо враховувати інші фактори та проводити додаткові експерименти для підтвердження цього висновку.

Таким чином, з урахуванням проведених досліджень і результатів аналізу, ми можемо прийти до висновку, що метод опорних векторів є найбільш ефективною моделлю для розв’язання нашої задачі аналізу даних.

ВИСНОВКИ

У роботі було досліджено сутність кредитних ризиків та проведено детальний аналіз теоретичних аспектів управління ними. Також було розглянуто сучасні методи оцінювання кредитних ризиків. Крім того у роботі проведено комп'ютерне моделювання кредитного ризику з використанням методів машинного навчання.

У першому розділі проаналізовано теоретичні аспекти управління кредитними ризиками. Висвітлено поняття кредитного ризику, його складові та вплив на фінансову стабільність банків та інших фінансових установ. Також розглянуто основні підходи до управління кредитними ризиками, зокрема, диверсифікацію, лімітування та створення резервів для кредитних портфелів. Проведено дослідження моделі оцінки кредитного ризику, такі як: модель Кредитного скорингу FICO, а також модель VantageScore.

У другому розділі досліджено сучасні методи оцінювання кредитних ризиків. Розглянуто основні підходи до кредитного скорингу, включаючи статистичні методи та методи машинного навчання. Також в другому розділі описано методи для фільтрації даних, моделі для прогнозування й особливо важливі критерії адекватності математичних моделей. Зокрема, в пункті 2.3 описано три найголовніші критерія адекватності моделі, а саме Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC).

У третьому розділі проведено комп'ютерне моделювання кредитного ризику з використанням методів машинного навчання обраного вхідного масиву даних. Проведено дослідження щодо кращого програмного забезпечення для даного дослідження та обрано програму Python, за низкою вагомих аргументів. В роботі було використано різні алгоритми машинного навчання, такі як: логістична регресія, метод найближчих сусідів та метод опорних векторів, для

побудови моделей оцінки кредитного ризику. Проведено експерименти та оцінено ефективність розроблених моделей.

За результатами цього дослідження, було визначено більш ефективний метод для аналізу кредитного ризику – метод опорних квадратів. Хоча всі методи були приблизно на однаковому рівні, він показав найкращі показники серед інших. Наприклад, показник accurancy (показник точності) методу опорних квадратів становить 82%, коли в логістичній регресії цей показник становить 81%, а в методі найближчих сусідів взагалі – 79%. Також показник f1-score (показник ефективності) показує досить великі результати по робочому кредитному портфелю, а саме – 89%, й 44% по дефолтному портфелю. Хоча логістична регресія має такі ж показники за робочим портфелем, але вона має гірші показники за дефолтним – 36%. Й метод найближчих сусідів має наступні показники за робочим портфелем – 87% та за дефолтним – 41%.

Отже, на основі проведених досліджень можна зробити висновок, що використання методів машинного навчання у оцінюванні кредитних ризиків є перспективним та ефективним підходом. Ці методи дозволяють покращити точність та передбачуваність оцінки кредитного ризику, знизити ризики для кредитних установ та покращити прийняття рішень щодо надання кредитів. Однак, слід пам’ятати, що використання машинного навчання також потребує належної підготовки та обробки даних, а також врахування етичних та регуляторних аспектів.

Майбутні дослідження в галузі оцінювання кредитних ризиків можуть включати розробку нових моделей, вдосконалення існуючих методів машинного навчання, а також дослідження впливу інших факторів на кредитний ризик, таких як соціальні мережі та поведінкові дані.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Особливості банківського кредиту. Освіта в Україні – Освіта.ua. URL: https://osvita.ua/vnz/reports/econom_pidpr/19050/ (дата звернення: 11.05.2023).

2. Кравченко В. Сутність та економічна природа інвестиційного ризику. LivingFo. URL: https://livingfo.com/sutnist-ta-ekonomichna-pryroda-investytsijnoho-ryzyku/ (дата звернення: 11.05.2023).

3. Verma E. What is financial risks and its types?. Simplilearn. URL: https://www.simplilearn.com/financial-risk-and-types-rar131-article (дата звернення: 13.05.2023).

4. Поняття, структура та фактори, що формують кредитний ризик. Finalearn - Фінансова аналілика. URL: http://www.finalearn.com/lifers-1662-1.html (дата звернення: 15.05.2023).

5. Поляруш І. М. Скоринг, як вдосконалений механізм оцінки потенційного позичальника банком – демонстрація процесу обробки даних. Ефективна економіка. 2015. № 11. URL: http://www.economy.nayka.com.ua/?op=1&z=4510 (дата звернення: 15.05.2023).

6. Brock T. What is credit scoring? Purpose, factors, and role in lending. The Investopedia. URL: https://www.investopedia.com/terms/c/credit_scoring.asp (дата звернення: 15.05.2023).

7. Кредитний скоринг: як фінансові установи визначають, чи можна нам дати у борг. PaySpaceMagazine. URL: https://psm7.com/uk/analytics/chernovik-6.html (дата звернення: 15.05.2023).

8. O'Shea B. What is a vantagescore?. NerdWallet. URL: https://www.nerdwallet.com/article/finance/vantagescore-fico-score-the-difference (дата звернення: 15.05.2023).

9. Analysis of financial statements. Corporate Finance Institute. URL: https://corporatefinanceinstitute.com/resources/accounting/analysis-of-financial-statements/ (дата звернення: 15.05.2023).

10. Мосьондз О. Б. Аналіз фінансового стану підприємства: сутність і необхідність. Ефективна економіка. 2012. № 3. URL: http://www.economy.nayka.com.ua/?op=1&z=1016 (дата звернення: 15.05.2023).

11. Methods in complex analysis. Mathematics of waves and materials. URL: https://www.mwmresearchgroup.org/methods-in-complex-analysis.html (дата звернення: 17.05.2023).

12. Ніколаєнко Ю. В. Процес управління кредитним ризиком як складова банківського менеджменту. Економічна наука. 2015. С. 99–102. URL: http://www.investplan.com.ua/pdf/23_2015/20.pdf (дата звернення: 17.05.2023).

13. Diversification credit. International risk management institute. URL: https://www.irmi.com/term/insurance-definitions/diversification-credit (дата звернення: 17.05.2023).

14. Kagan J. What is a credit limit? How it's determined and how to increase it. The Investopedia. URL: https://www.investopedia.com/terms/c/credit_limit.asp (дата звернення: 17.05.2023).

15. Risk limit. Open risk manual. URL: https://www.openriskmanual.org/wiki/Risk_Limit (дата звернення: 18.05.2023).

16. Reserve accounting. AccountingTools. URL: https://www.accountingtools.com/articles/what-is-reserve-accounting.html (дата звернення: 17.05.2023).

17. Alpert G. What is a loan loss provision? Definition and use in accounting. The Investopedia. URL: https://www.investopedia.com/terms/l/loanlossprovision.asp (дата звернення: 17.05.2023).

18. Logistic regression and machine learning. International Business Machines. URL: https://www.ibm.com/topics/logistic-

regression#Logistic+regression+and+machine+learning (дата звернення: 18.05.2023).

19. Edgar T. W. Logistic regression. ScienceDirect. URL: https://www.sciencedirect.com/topics/computer-science/logistic-regression (дата звернення: 18.05.2023).

20. Decision tree. GeeksForGeeks. URL: https://www.geeksforgeeks.org/decision-tree/ (date of access: 18.05.2023).

21. Chen J. What is a neural network?. The Investopedia. URL: https://www.investopedia.com/terms/n/neuralnetwork.asp (дата звернення: 18.05.2023).

22. What is the k-nearest neighbors algorithm?. International Business Machines. URL: https://www.ibm.com/topics/knn (дата звернення: 18.05.2023).

23. An introduction to exponential smoothing for time series forecasting in Python. Simplilearn. URL: https://www.simplilearn.com/exponential-smoothing-for-time-series-forecasting-in-python-article (дата звернення: 25.05.2023).

24. Hyndman R. J., Athanasopoulos G. Autoregressive models. Forecasting: principles and practice. Australia. URL: https://otexts.com/fpp2/AR.html (дата звернення: 25.05.2023).

25. Hyndman R. J., Athanasopoulos G. Moving average models. Forecasting: principles and practice. Australia. URL: https://otexts.com/fpp2/MA.html (дата звернення: 25.05.2023).

26. Hyndman R. J., Athanasopoulos G. Seasonal ARIMA models. Forecasting: principles and practice. Australia. URL: https://otexts.com/fpp2/seasonal-arima.html (дата звернення: 25.05.2023).

27. Mehandzhiyski V. What Is an ARMA Model?. 365DataScience. URL: https://365datascience.com/tutorials/time-series-analysis-tutorials/arma-model/ (дата звернення: 25.05.2023).

28. Zajic A. What Is Akaike Information Criterion (AIC)?. Builtin. URL: https://builtin.com/data-science/what-is-aic (дата звернення: 25.05.2023).

29. DelSole T. Correcting the corrected AIC. ScienceDirect. URL: https://www.sciencedirect.com/science/article/pii/S0167715221000262 (дата звернення: 25.05.2023).

30. Bauldry S. Bayesian Information Criterion. ScienceDirect. URL: https://www.sciencedirect.com/topics/social-sciences/bayesian-information-criterion (дата звернення: 25.05.2023).

31. Mean squared error: definition and example. Statistics How To. URL: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/ (дата звернення: 25.05.2023).

32. Absolute error & mean absolute error (MAE). Statistics How To. URL: https://www.statisticshowto.com/absolute-error/ (дата звернення: 25.05.2023).

33. Coefficient of determination (R squared): definition, calculation. Statistics How To. URL: https://www.statisticshowto.com/probability-and-statistics/coefficient-of-determination-r-squared/ (дата звернення: 25.05.2023).

34. BasuMallick C. What is MATLAB? Working, functions, and applications. SpiceWorks. URL: https://www.spiceworks.com/tech/devops/articles/what-is-matlab/ (дата звернення: 25.05.2023).

35. What is R coding language and why is it so important?. Emeritus. URL: https://emeritus.org/blog/coding-r-coding-language/ (дата звернення: 25.05.2023).

36. Pedamkar P. What is SAS?. Educba. URL: https://www.educba.com/what-is-sas/ (дата звернення: 25.05.2023).

37. Python introduction. W3schools. URL: https://www.w3schools.com/python/python_intro.asp (дата звернення: 25.05.2023).

38. Мірошниченко І. ARIMA. Githack. URL: https://raw.githack.com/Aranaur/aranaur-apero/main/content/talk/2022-forecasting-if/lecture/08.html#/title-slide (дата звернення: 26.05.2023).

39. Mudgalvivek. Machine learning : confusion matrix (Error Matrix). Medium. URL: https://medium.com/@mudgalvivek2911/machine-learning-confusion-matrix-error-matrix-c518bca18de7 (дата звернення: 25.05.2023).

40. Default of credit card clients dataset. Kaggle. URL: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset?resource=download (дата звернення: 05.05.2023).

ДОДАТОК А

ЛІСТИНГ ПРОГРАМИ

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams['figure.figsize'] = (8, 8)
pd.options.display.float_format = '{:.3f}'.format
credit_risk = pd.read_csv('UCI_Credit_Card.csv')
credit_risk.head()
#формати та кільсть значень. перевірка на пропуски
credit_risk.info()
credit_risk.drop(['ID'], axis=1, inplace=True)
#редагування значень змінної "Освіта"
credit_risk['EDUCATION'].replace({0:1,1:1,2:2,3:3,4:4,5:1,6:1},
inplace=True)
credit_risk.EDUCATION.value_counts()
credit_risk['MARRIAGE'].replace({0:3,1:1,2:2,3:3}, inplace=True)
credit_risk.MARRIAGE.value_counts()
credit_risk.SEX.value_counts()
credit_risk['default.payment.next.month'].value_counts()
#перегляд основної описової статистики
credit_risk.describe()
#перевірка на викиди змінної "Вік"
credit_risk[['AGE']].boxplot(figsize = (10, 10));
#розвідувальний аналіз + візуалізація (дефолт)
credit_risk[['default.payment.next.month']].value_counts().plot.bar();
#розвідувальний аналіз + візуалізація (навчання + дефолт)
sns.countplot(data = credit_risk, x = 'EDUCATION', hue =
'default.payment.next.month');
```

```
plt.xticks(rotation=45);
#розвідувальний аналіз + візуалізація (вік)
sns.distplot(credit_risk['AGE']);
#розвідувальний аналіз + візуалізація (вік)
sns.distplot(credit_risk['AGE']);
#візуальний аналіз
pd.plotting.scatter_matrix(credit_risk[['BILL_AMT1', 'PAY_AMT1']],
figsize=(8, 8), diagonal='kde');
#візуальний аналіз
pd.plotting.scatter_matrix(credit_risk[['BILL_AMT2', 'PAY_AMT2']],
figsize=(8, 8), diagonal='kde');
#визначимо матрицю незалежних змінних та вектор залежної змінної
X = credit_risk.drop(credit_risk[['default.payment.next.month']], axis=1)
print(X.head())
y = credit_risk['default.payment.next.month']
print(y.head())
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
sc.fit(X)
X_std = sc.fit_transform(X)
#навчальну вибірку використовуємо для навчання моделі
#тестову вибірку використовуємо для перевірки якості побудованої
моделі
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
   X_std, y, stratify = y, test_size=0.3, random_state=0)
#створення і навчання класифікатора  на  навчальному наборі даних
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
#навчаємо модель на навчальній вибірці - X_train, y_train
```

```
lr.fit(X_train, y_train)
#робимо прогноз цільової змінної на тестовій вибірці - X_test
y_pred_lr = lr.predict(X_test)
#побудуємо матрицю помилок
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
#матриця помилок
cm = confusion_matrix(y_test, y_pred_lr)
#візуалізація матриці помилок
cmd = ConfusionMatrixDisplay(cm, display_labels=['0','1'])
#display_label - задаються назви класів (0 - 1-й клас, 1 - 2-й клас)
cmd.plot()
#оцінимо якість класифікатора
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_lr))
#створення і навчання класифікатора  на  навчальному наборі даних
from sklearn.neighbors import KNeighborsClassifier
KNN = KNeighborsClassifier()
KNN.fit(X_train, y_train)
#робимо прогноз на тестовій вибірці - X_test
y_pred_KNN = KNN.predict(X_test)
#побудуємо матрицю помилок
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
#матриця помилок
cm = confusion_matrix(y_test, y_pred_KNN)
#візуалізація матриці помилок
cmd = ConfusionMatrixDisplay(cm, display_labels=['0','1'])
#display_label - задаються назви класів
cmd.plot()
```

```
#оцінимо якість класифікатора
print(classification_report(y_test, y_pred_KNN))
#створення і навчання класифікатора  на  навчальному наборі даних
from sklearn.svm import SVC
svm = SVC()
svm.fit(X_train, y_train)
y_pred_svm = svm.predict(X_test)
#побудуємо матрицю помилок
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
#матриця помилок
cm = confusion_matrix(y_test, y_pred_svm)
#візуалізація матриці помилок
cmd = ConfusionMatrixDisplay(cm, display_labels=['0','1'])
# display_label - задаються назви класів
cmd.plot()
#оцінимо якість класифікатора
print(classification_report(y_test, y_pred_svm))
```

Для читабельної візуалізації було створено новий файл

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams['figure.figsize'] = (8, 8)
pd.options.display.float_format = '{:.3f}'.format
credit_risk = pd.read_csv('UCI_Credit_Card.csv')
credit_risk.head()
#формати та кільсть значень. перевірка на пропуски
credit_risk.info()
```

```
credit_risk.drop(['ID'], axis=1, inplace=True)
#редагування значень змінної "Освіта"
credit_risk['EDUCATION'].replace({0:1,1:1,2:2,3:3,4:4,5:1,6:1},
inplace=True)
credit_risk['EDUCATION'].replace({1:"Коледж",2:"Університет",3:"Школа",
4:"Інше"}, inplace=True)
credit_risk.EDUCATION.value_counts()
credit_risk['MARRIAGE'].replace({0:3,1:1,2:2,3:3}, inplace=True)
credit_risk['MARRIAGE'].replace({1:"Одружений",2:"Неодружений",3:"Інш
е"}, inplace=True)
credit_risk.MARRIAGE.value_counts()
credit_risk['SEX'].replace({1:"Чоловік",2:"Жінка"}, inplace=True)
credit_risk.SEX.value_counts()
credit_risk['default.payment.next.month'].replace({0:"Not
default",1:"Default"}, inplace=True)
credit_risk['default.payment.next.month'].value_counts()
#розвідувальний аналіз + візуалізація (дефолт)
credit_risk[['default.payment.next.month']].value_counts().plot.bar();
#розвідувальний аналіз + візуалізація (навчання + дефолт)
sns.countplot(data = credit_risk, x = 'EDUCATION', hue =
'default.payment.next.month');
plt.xticks(rotation=45);
#розвідувальний аналіз + візуалізація (стать + дефолт)
sns.countplot(data = credit_risk, x = 'SEX', hue =
'default.payment.next.month');
plt.xticks(rotation=45);
```