НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 004.852, 51.77

До захисту допущено
Завідувач кафедри ММСА
_____ Оксана ТИМОЩУК
«___» _____ 2024 р.

Магістерська дисертація
на здобуття ступеня магістра
за освітньо-професійною програмою «Системний аналіз фінансового ринку»
зі спеціальності 124 «Системний аналіз»
на тему: «Система оцінки та прогнозування кредитних ризиків у банківському секторі»

Виконав:
Студент 2 курсу, групи КА-22мп
Сумін Олександр Олександрович _____

Науковий керівник:
Доцент кафедри ММСА,
к.ф.-м.н. Шубенкова Ірина Анатоліївна _____

Рецензент:
Професор кафедри ШІ
д.т.н. Данилов Валерій Якович _____

Засвідчую, що у цій магістерській дисертації
немає запозичень з праць інших авторів
без відповідних посилань
Студент (*підпис*): _____

Київ - 2024

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти — другий (магістерський)
Спеціальність — 124 «Системний аналіз»
Освітньо-професійною програмою «Системний аналіз фінансового ринку»

ЗАТВЕРДЖУЮ
Завідувач кафедри ММСА
_____ Оксана ТИМОЩУК
«___» _____ 2024 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Суміну Олександру Олександровичу
(прізвище ім'я по батькові)

1. Тема дисертації: «Система оцінки та прогнозування кредитних ризиків у банківському секторі», науковий керівник дисертації Шубенкова Ірина Анатоліївна к.ф.-м.н., доц., доцент кафедри ММСА,
(прізвище ім'я по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «08» листопада 2023 р. № 5200-с

2. Строк подання студентом дисертації:

3. Об'єкт дослідження: процес кредитування та управління кредитними ризиками для забезпечення фінансової стабільності банків, підприємств та держави.

4. Предмет дослідження: моделі машинного навчання та методи статистичного аналізу виживаності в задачі прогнозування та оцінки кредитних ризиків.

5. Перелік завдань, які потрібно розробити.

1. Огляд предметної області.
2. Аналіз дослідницьких даних.
3. Розробка моделі вирішення задачі для прогнозування та оцінки кредитних ризиків.
4. Тестування отриманих моделей та обрання кращої з них.

6. Перелік графічного (ілюстративного) матеріалу.

1. Рисунки.
2. Таблиці.
3. Презентація.

7. Орієнтовний перелік публікацій.

Сумін О. О., Шубенкова І. А. Системний підхід до аналізу кредитних ризиків в банківському секторі. II Всеукраїнська науково-практична конференція «Системні науки та інформатика», м. Київ, 04-08 грудня 2023 року. С. 213-217.

8. Консультанти розділів дисертації.

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
-	-	-	-

9. Дата видачі завдання: 01 вересня 2023

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Строк виконання етапів магістерської дисертації	Примітка
1	Затвердження теми МД. Ознайомлення зі структурою МД згідно з Положенням про державну атестацію студентів НТУУ «КПІ ім. І. Сікорського»	01.09.2023-10.09.2023	виконано
2	Ознайомлення з ДСТУ 3008-2015 та стандартами ЄСПД	11.09.2023-17.09.2023	виконано
3	Перший розділ. Огляд літературно інформаційних джерел. Аналіз предметної області	18.09.2023-01.10.2023	виконано
4	Другий розділ. Розробка теоретичного узагальнення методу	02.10.2023-15.10.2023	виконано
5	Третій розділ. Розробка програмного забезпечення	16.10.2023-29.10.2023	виконано
6	Третій розділ. Робота над практичним розділом магістерської дисертації	30.10.2023-12.11.2023	виконано
7	Четвертий розділ. Стартап-проект	13.11.2023-19.11.2023	виконано

№ з/п	Назва етапів виконання магістерської дисертації	Строк виконання етапів магістерської дисертації	Примітка
8	Концептуальні висновки. Перспективи розвитку отриманих рішень	20.11.2023-26.11.2023	виконано
9	Оформлення магістерської дисертації	27.11.2023-31.12.2023	виконано

Студент Олександр СУМІН

Науковий керівник дисертації Ірина ШУБЕНКОВА

РЕФЕРАТ

Магістерська дисертація: 102 с., 24 рис., 32 табл., 1 додаток, 17 джерел.

СИСТЕМА ОЦІНКИ ТА ПРОГНОЗУВАННЯ КРЕДИТНИХ РИЗИКІВ У БАНКІВСЬКОМУ СЕКТОРІ

У світлі зростання попиту на кредитування, передбачення кредитних ризиків стає пріоритетним завданням для банків та інших кредитних установ. Прогнозування дефолту та оцінка кредитних збитків визначають основні аспекти кредитної політики, спрямовані на забезпечення фінансової стійкості та відповідального кредитування. В даному контексті використовуються різноманітні методи, серед цих методів важливе місце належить статистичним моделям та методам машинного навчання. Сучасні технології дозволяють аналізувати обширні фінансові дані та враховувати низку факторів, що впливають на кредитоспроможність клієнтів.

Об’єкт дослідження: процес кредитування та управління кредитними ризиками для забезпечення фінансової стабільності банків, підприємств та держави.

Предмет дослідження: моделі машинного навчання та методи статистичного аналізу виживаності в задачі прогнозування та оцінки кредитних ризиків.

Мета роботи: розробка ефективної стратегії управління кредитним ризиком за допомогою прогнозування дефолту та збитків в банківській сфері за допомогою машинного навчання та статистичних методів аналізу виживаності.

На мові Python створено програмний продукт для прогнозування та оцінки кредитних ризиків.

Ключові слова: системний аналіз, кредитні ризики, машинне навчання, моделі виживання.

ABSTRACT

Master's thesis: 102 p., 24 fig., 32 tables, 1 appendix, 17 sources.

SYSTEM FOR ESTIMATING AND FORECASTING CREDIT RISKS IN THE BANKING SECTOR

In face of the growing demand for credit, forecasting credit risks is becoming a priority for banks and other credit institutions. Default forecasting and credit loss assessment determine the main aspects of credit policy aimed at ensuring financial stability and responsible lending. A variety of methods are used in this context, with statistical models and machine learning techniques taking an important place among them. Modern technologies make it possible to analyze extensive financial data and take into account a number of factors that affect the creditworthiness of customers.

The object of research: the process of lending and credit risk management to ensure the financial stability of banks, enterprises and the state.

The subject of research: machine learning models and methods of statistical analysis of survival in the task of forecasting and assessing credit risks.

The purpose of the work is to develop an effective strategy for managing credit risk by predicting default and losses in the banking sector using machine learning and statistical methods of survival analysis.

A software product for forecasting and assessing credit risks was created in Python.

Keywords: system analysis, credit risks, machine learning, survival models.

ЗМІСТ

ВСТУП

В умовах постійного економічної нестабільності та росту попиту на фінансові послуги, питання управління кредитними ризиками набуває особливої важливості для забезпечення стабільності банків та фінансових установ в цілому. З врахуванням динамічного характеру ринку та несприятливих економічних умов, передбачення кредитних ризиків стає критичною умовою успішної фінансової стратегії. В цьому контексті, банки та інші фінансові установи шукають ефективні інструменти та методи для ефективного управління кредитним портфелем.

Нормативне регулювання грає ключову роль у визначенні стандартів та вимог до ефективного управління кредитними ризиками. Базельські нормативи, вперше запроваджені Банком міжнародних розрахунків з метою забезпечення фінансової стабільності, визначають норми для визначення капіталовкладень та ризикованості банківських активів. Ці нормативи визнаються як фундаментальний інструмент для збереження стабільності банківської системи та запобігання фінансовим кризам.

Перший розділ магістерської дисертації спрямований на докладний огляд актуальності проблеми передбачення кредитних ризиків у контексті зростаючого попиту на кредитування та розвитку фінансового ринку. Зокрема, розглядаються виклики, що постають перед сучасними фінансовими установами у зв'язку з необхідністю адаптації до динамічного середовища та забезпеченням відповідності нормативам.

Другий розділ присвячений теоретичному аналізу та детальному огляду застосованих моделей та методів управління кредитними ризиками. Проводиться аналіз традиційних та новаторських підходів до передбачення кредитних ризиків, а також розглядаються переваги та недоліки різних методології.

Третій розділ роботи фокусується на практичних дослідженнях, що включають тестування різних моделей та вибір оптимального підходу для розрахунку кредитних ризиків. Здійснюється оцінка ефективності моделей на реальних даних та визначається їхній внесок у покращення управління кредитним портфелем.

Четвертий розділ, який становить основу для подальшого розвитку досліджень, присвячений побудові власного стартап-проекту. Проаналізовано можливість впровадження інноваційного підходу до управління кредитними ризиками та його потенційний вплив на фінансовий сектор.

Усі ці розділи узагальнюються з метою надання комплексного погляду на проблему управління кредитними ризиками та виявлення оптимальних шляхів вирішення цієї важливої задачі. Магістерська дисертація спрямована на висвітлення не тільки теоретичних аспектів, але й на розробку конкретних рекомендацій та практичних рішень для фінансових установ у сфері управління кредитними ризиками.

РОЗДІЛ 1 СУЧАСНИЙ СТАН ДОСЛІДЖЕНЬ У ГАЛУЗІ ОЦІНЮВАННЯ РИЗИКІВ КРЕДИТУВАННЯ

У даному розділі надається обґрунтування актуальності поставленої задачі, що висвітлює важливість та реальний контекст, в якому вона виникає. Аналізуються сучасні тенденції та виклики, які роблять дану проблему важливою для вивчення та вирішення. Також в розділі розглядаються нормативно-правові акти, які регулюють дану сферу. Аналіз законодавчого середовища допомагає визначити обсяг та межі дії, які необхідно враховувати при вирішенні поставленої задачі. Це включає в себе вивчення законів, положень, та інших правових інструментів, які мають прямий чи опосередкований вплив на розглянуту проблематику.

1.1 Актуальність кредитування та його фінансове забезпечення

Кредитування є важливим інструментом для розвитку економіки, включаючи малий бізнес та сільське господарство. Наукові дослідження показують, що кредитування фізичних осіб має значну роль в одержанні прибутку банків, тому що стабільність кредитних відносин з індивідуальним клієнтом визначає ефективність функціонування банку. Проектне фінансування може бути ефективною формою кредитування інвестиційних проектів, оскільки угоди такі мають більший термін дії і спрямовані на позичальників із вищим рівнем ризику порівняно із звичайними угодами. Кредитування малого бізнесу та його фінансове забезпечення є важливими проблемами, які потребують вирішення, оскільки малі фірми та банки мають проблеми з ефективним співробітництвом через високі кредитні ризики. Аграрні розписки можуть бути альтернативним інструментом кредитування

сільськогосподарських товаровиробників, оскільки вони дозволяють забезпечити фінансове забезпечення сільськогосподарських підприємств.

Отже, кредитування є актуальною темою для дослідження, оскільки воно є важливим інструментом для розвитку економіки та має значний вплив на фінансову стабільність підприємств та держави.

Відповідно до попиту на кредитування, прогноз кредитних ризиків є актуальною задачею для банків та інших кредитних установ, оскільки він дозволяє зменшити ризики та збільшити ефективність кредитування. Це дозволяє банкам та іншим кредитним установам приймати рішення щодо надання кредиту та встановлювати відсоткові ставки в залежності від ризику.

Прогноз дефолту та кредитних збитків представляє собою ключовий складник кредитної політики банків та інших кредитних установ, оскільки він сприяє зниженню ризиків та підвищенню результативності процесу кредитування.

Для прогнозу дефолту та кредитних збитків використовують різноманітні методи, включаючи статистичні моделі, машинне навчання та інші підходи. Наприклад, один із таких методів - це використання моделі логістичної регресії, яка дозволяє оцінити ймовірність дефолту на основі різних факторів (кредитна історія, рівень доходу, тощо).

Загалом, прогноз кредитних ризиків є необхідною складовою стратегії фінансових установ, оскільки він сприяє кращому управлінню ризиками та підвищує ефективність процесу надання кредитів.

1.2 Базельські нормативи, імплементація в Україні та вплив на функціонування систем управління ризиками

Збереження стабільності банківської системи сьогодні є найважливішим завданням для будь-якої світової економіки. З цією метою Міжнародним банком розрахунків (BIS) був створений набір правил, відомий як Базель III. Цей набір правил призначений для покращення оцінки та прогнозування ризиків банківськими установами, а також для зменшення можливих збитків під час кредитування.

Базельські нормативи складаються з трьох підгруп: Базельський комітет з банківського нагляду I (Basel I), Базельський комітет з банківського нагляду II (Basel II) та Базельський комітет з банківського нагляду III (Basel III). Кожна з цих підгруп містить вимоги до капіталу банків та інших фінансових установ, а також рекомендації щодо управління ризиками та забезпечення фінансової стабільності.

Базель III - це третій етап в розвитку міжнародних стандартів, які встановлюють вимоги до капіталу банків та інших фінансових установ. Ці нормативи були з метою забезпечення стабільності фінансової системи та запобігання фінансовим кризам та представляють собою комплекс правил і вимог, які регулюють діяльність банків та інших фінансових установ з метою забезпечення стабільності фінансової системи та запобігання фінансовим кризам.

Основні зміни в Базельських нормативах III включають введення нових вимог до капіталу, зокрема, вимог до капіталу з урахуванням ризиків. Це допомагає банкам більш точно оцінювати та керувати своїми ризиками. Крім того, нормативи містять вимоги до ліквідності та управління ризиками, включаючи вимоги щодо ефективної системи управління ризиками та звітності.

Базельські нормативи III відіграють важливу роль у забезпеченні фінансової стійкості та підтримці функціонування фінансових установ. Вони

сприяють запобіганню фінансовим кризам та регулюванню діяльності банків та інших фінансових інституцій з метою збереження стабільності фінансової системи.

Базельські нормативи впливають значно на функціонування систем управління ризиками в Україні. Відповідно до методичних рекомендацій Базель III, Національний банк України, що виступає національним регулятором банківського сектору в Україні, ввів Постанову №351 «Про затвердження Положення про визначення банками України розміру кредитного ризику за активними банківськими операціями» (зі змінами).

Українські банки зобов'язані дотримуватись вимог цього Положення, яке враховує принципи та рекомендації Базельського комітету з банківського нагляду. Вони повинні розвивати ефективну систему управління ризиками, яка дозволяє виявляти, оцінювати та контролювати ризики, пов'язані з їх діяльністю. Згідно з Положенням, банки повинні створити внутрішні положення відповідно до вимог цього документа та проводити розрахунок розміру кредитного ризику в тестовому режимі.

Розрахунок значення кредитного ризику індивідуального позичальника, який визначається відповідно до Постанови №351, проводиться за визначеною формулою:

$$CR = PD * LGD * EAD$$

PD – ймовірність дефолту

LGD – збитки в разі дефолту

EAD – експозиція під ризиком дефолту

Зазначимо, що формула для розрахунку кредитного ризику, яка визначена в Постанові №351 від Національного банку України, відображає консервативний підхід до оцінки кредитних ризиків. Цей підхід передбачає врахування можливості впливу кризових подій на ризик позичальника, і, таким чином, сприяє більш точному та обережному управлінню кредитними ризиками банків.

Важливо враховувати, що в умовах фінансової нестабільності та економічних криз формули та методи, які базуються на консервативних оцінках ризиків, можуть стати ефективними інструментами для банків та регуляторів. Це допомагає зменшити можливі збитки та зберегти стабільність фінансової системи.

Такий підхід до оцінки кредитних ризиків відображає основний принцип Базельських нормативів, які ставлять перед собою завдання забезпечити надійність та стійкість банківської системи в умовах можливих криз та стрес-тестів.

Отже, вплив Базельських нормативів на системи управління ризиками в Україні є значущим, оскільки вони встановлюють строгі вимоги до капіталу, ліквідності та ризиків, а також визначають стандарти управління ризиками та звітності. Дотримання цих вимог допомагає українським банкам ефективно враховувати та управляти ризиками в умовах відповідності міжнародним стандартам.

Не менш важливим документом, який регулює основні принципи банківської діяльності в галузі кредитування, є Міжнародні стандарти фінансової звітності (МСФЗ 9), відомі також як IFRS (International Financial Reporting Standards). Ці стандарти розробляються та затверджуються Радою з міжнародних стандартів бухгалтерського обліку (IASB, International Accounting Standard Board). МСФЗ 9 є частиною англосаксонської традиції фінансового обліку та мають світовий статус. Вони містять важливі визначення та правила щодо обліку та узгодження кредитних операцій та ризиків. Ці стандарти визначають принципи визнання та оцінки кредитних інструментів, включаючи визнання втрат, пов'язаних із кредитами.

МСФЗ 9 впливають на звітність фінансових установ та компаній, включаючи банки, і вимагають від них встановлення конкретних підходів до оцінки та управління кредитними ризиками відповідно до міжнародних стандартів фінансового звітності. Це важливий інструмент для забезпечення

прозорості фінансової інформації банків та інших компаній на світовому ринку.

Згідно з МСФЗ 9, оцінка кредитного ризику включає в себе поняття очікуваного кредитного збитку, а також рекомендацію формувати резерви для покриття цих очікуваних збитків. У порівнянні з Постановою №351 Національного банку України (ПНБУ №351), практика вказує на більшу чутливість оцінки за МСФЗ 9.

Метод оцінки кредитного збитку, запропонований МСФЗ 9, на перший погляд схожий на метод оцінки кредитного ризику за ПНБУ №351. Однак практика показує значні відмінності в оцінюванні компонентів цих методів. Наприклад, ймовірність дефолту за МСФЗ 9 розраховується з урахуванням поточного впливу макроекономічних факторів, тоді як в ПНБУ №351 вона "усереднюється".

Ці відмінності в методологіях впливають на точність та надійність оцінки кредитного ризику, роблячи оцінку за МСФЗ 9 більш чутливою до змін у макроекономічному середовищі. Врахування цих факторів допомагає фінансовим установам краще оцінювати та управляти ризиками та формувати адекватні резерви для покриття очікуваних кредитних збитків.

1.3 Ризики та збитки при кредитуванні

Вкажемо основні ризики, пов'язані з кредитуванням, згідно з дослідженнями, проведеними в Україні.

1. **Високі процентні ставки**, які можуть зростати, що може призвести до проблем з погашенням кредиту.
2. **Недостатня регуляторна база та низька культура кредитування.**
3. **Обмеженість у доступі** до кредитних ресурсів підприємств малого і середнього бізнесу.

4. **Фінансова нестабільність** у сільськогосподарських підприємств, **відсутність ліквідних забезпечень** у позичальників та невпевненість сільського населення у прийнятті практик кредитування.

5. **Зміна ринкових умов**, що може призвести до збитків для кредитора.

6. **Недостатність фінансових ресурсів**, якщо позичальник не зможе повернути позику, це може призвести до збитків для кредитора.

Згідно з цих досліджень, кредитний ризик є одним з основних ризиків, пов'язаних з кредитуванням. Кредитний ризик може виникнути, якщо позичальник не зможе повернути позику вчасно або не зможе повернути її взагалі. Це може призвести до збитків для кредитора.

Для зменшення кредитного ризику банки використовують різні методи, такі як аналіз кредитної історії позичальника, встановлення лімітів на кредитування, вимоги до застави та інші. Також, застосування методів аналізу великих даних може допомогти виявити потенційні ризики та запобігти ризикам надмірного кредитування.

Дефолт є найбільш яскравим та серйозним кредитним ризиком, оскільки масштаб несприятливих наслідків для обох сторін кредитних відносин є найбільшим у разі дефолту. Однак, кредитний ризик може бути зменшений за допомогою правильного планування та управління ризиками, таких як аналіз кредитної історії позичальника, встановлення лімітів на кредитування, вимоги до застави та інші.

Кредитний ризик тісно пов'язаний з кредитними збитками, які можна описати як різницю між усіма планованими грошовими потоками, які повинні були б отримати суб'єкт господарювання відповідно до договору, та всіма грошовими потоками, які суб'єкт господарювання очікує одержати (іншими словами, це сума всіх невиплачених грошових обов'язків), знижених за допомогою дисконтування за початковою ефективною процентною ставкою.

Кредитний ризик використовується для оцінювання розміру очікуваних кредитних збитків, що може виникнути через невиконання боржниками своїх фінансових зобов'язань.

Унаслідок унікальних та індивідуальних відносин між сторонами у галузі кредитування, кредитний ризик виявляється як явище з певною ступенем невизначеності та численними особливостями. Фінансові втрати можуть настанути для банку на будь-якому етапі дії контракту, і, як вже було зазначено, це може бути обумовлено різноманітними факторами.

1.4 Регуляторні вимоги розрахунку кредитного ризику

Регуляторні вимоги щодо розрахунку кредитного ризику визначаються законодавством країни та регуляторними органами, такими як Національний банк України. В Україні регуляторні вимоги щодо розрахунку кредитного ризику встановлені №351 "Про затвердження Положення про визначення банками України розміру кредитного ризику за активними банківськими операціями" (зі змінами).

Згідно з цим Положенням, банки повинні розробляти та впроваджувати внутрішні процедури оцінки кредитного ризику, які повинні відповідати вимогам законодавства та регуляторних органів. Внутрішні процедури повинні містити методики та моделі оцінки кредитного ризику, а також критерії класифікації кредитів за рівнем ризику.

Крім того, Положення встановлює вимоги до звітності банків щодо кредитного ризику. Банки повинні подавати до Національного банку України звіти про кредитний ризик, які містять інформацію про портфелі кредитів, класифікацію кредитів за рівнем ризику, величину резервів на покриття можливих збитків від кредитного ризику та іншу інформацію, визначену законодавством та регуляторними органами.

Згідно з цим Положенням, банк, який має понад три роки досвіду банківської діяльності, визначає значення коефіцієнтів LGD (loss given default) на підставі виду застави та рівня покриття боргу заставою, керуючись власними оцінками. Встановлені банком значення коефіцієнтів LGD не

можуть бути меншими, ніж нижчі (кращі) межі діапазонів, передбачених цим Положенням. Банк, який є новим на ринку та функціонує менше трьох років, також визначає значення коефіцієнтів LGD на підставі виду застави та рівня покриття боргу заставою, ґрунтуючись на власних розсудах. Проте встановлені банком значення коефіцієнтів LGD не можуть бути нижчими (кращими), ніж середні значення діапазонів, передбачених цим Положенням.

1.5 Проблематика управління ризиками кредитного портфеля банка

Управління банківськими ризиками є надзвичайно важливою частиною банківської діяльності. Ефективне управління ризиками є ключовим аспектом забезпечення конкурентоспроможності та надійності банку. Принципи управління банківськими ризиками відіграють значну роль в рекомендаціях Базельського комітету з банківського нагляду.

Керівництво банку відіграє ключову роль у визначенні рівня прийнятих ризиків і повинно мати чітке розуміння можливих втрат. Проблеми банку зазвичай виникають тоді, коли менеджмент не розуміє важливості ризик-менеджменту. Тому важливо, щоб вищий рівень керівництва банку належну увагу приділяв сучасній системі управління ризиками та чітко усвідомлював, які ризики можна управляти і які – ні.

Банку слід розробити стратегію управління кожним з виділених ризиків, визначити їх вплив та прийнятний рівень ризику для банку і його вплив на кредитний портфель. Після цього формуються цілі управління ризиками та розробляється методологія оцінки та управління обраним спектром ризиків. Ця методологія включає конкретні показники та кількісні моделі для оцінки ризиків.

Після визначення цілей та методології управління ризиками здійснюється детальна регламентація процесу, що включає в себе дії персоналу банку, межі

відповідальності, структуру та рівень лімітування, а також встановлюється механізм співпраці між різними організаційними одиницями банку. Незалежний контроль за функціонуванням системи управління ризиками дозволяє оцінити відповідність проведених операцій плану та переконатися в їх здатності досягнення передбачених цілей.

Оцінку ризиків повинні проводити різні комітети та структурні підрозділи банку, такі як кредитний комітет та комітет з управління активами-пасивами. Важливу роль в управлінні ризиками відіграє внутрішній аудит.

Ефективне управління ризиками кредитного портфеля банку передбачає детальне розуміння структури та якості цього портфеля. Ризик кредитного портфеля визначається рівнем ризиків його окремих сегментів, кожен з яких має свою специфіку, а також залежить від рівня диверсифікації чи концентрації цих сегментів у структурі портфеля. Основним завданням банку є досягнення максимального прибутку при прийнятному рівні ризиків, що робить дохідність портфеля та рівень ризиків визначальними критеріями його якості.

Ризик кредитного портфеля пов'язаний з якістю та ліквідністю окремих елементів портфеля. Рівень ліквідності кредитного портфеля також важливий, і він залежить від можливості банку повертати позики клієнтів та їх ефективної реалізації. Крім того, ризик окремих елементів кредитного портфеля не завжди відображає їхню якість. Таким чином, управління ризиками кредитного портфеля вимагає комплексного підходу і уваги до всіх аспектів.

Висновки до розділу 1

У даному розділі було обґрунтовано актуальність, практичну значущість та проблематику обраного напрямку дослідження, а також розглянуто поняття кредитування та теоретичні основи оцінки кредитного ризику. Детально висвітлено існуючі вимоги та міжнародні стандарти, що регулюють

банківську діяльність в Україні у сфері кредитування. Завершенням даного розділу є вичерпний огляд предметної області, який в подальших розділах буде використано при розробці методики оцінки та прогнозування фінансового ризику портфеля.

РОЗДІЛ 2 ТЕОРЕТИЧНІ ВІДОМОСТІ

В даному розділі розглянуто теоретичну базу використання моделей машинного навчання та аналізу виживаності як ефективних інструментів для вирішення цих завдань.

Використання моделей машинного навчання виявляється дуже перспективним для прогнозування кредитних ризиків. Алгоритми класифікації дозволяють автоматизувати процес визначення ймовірності дефолту на основі великої кількості факторів. Наприклад, мережі можуть виявити складні неявні зв'язки, що істотно покращує прогностичні можливості.

Аналіз виживаності, або моделі виживання, використовуються для прогнозування часу до настання події, такої як дефолт. Вони дозволяють враховувати часові аспекти та динаміку ризиків. Такі моделі інтегрують фактори, які змінюються з часом, визначаючи ймовірність виживання позичальника на певний час.

2.1 Методи і моделі для вирішення задачі оцінювання ризиків настання дефолту

2.1.1 Логістична регресія

Логістична регресія — це статистичний метод, що використовується для моделювання й прогнозування ймовірностей виникнення подій. Вона є популярним методом у задачах бінарної класифікації, де треба визначити, до якого класу належить об'єкт. Основна ідея логістичної регресії полягає в тому, що вона використовує логістичну функцію для оцінки ймовірності належності об'єкта до певного класу.

Робиться припущення, що ймовірність настання події $y = 1$ дорівнює:

$$\mathbb{P}\{y = 1 \mid x\} = f(z), \text{ де}$$

$$z = \theta^T x = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

$$f(z) = \frac{1}{1 + e^{-z}} - \text{логістична функція (сигмоїда)}$$

Так як y приймає лише значення 0 і 1, то ймовірність прийняти значення 0 дорівнює:

$$\mathbb{P}\{y = 0 \mid x\} = 1 - f(z) = 1 - f(\theta^T x)$$

Для стислості функцію розподілу y при заданому x можна записати в такому вигляді:

$$\mathbb{P}\{y \mid x\} = f(\theta^T x)^y (1 - f(\theta^T x))^{1-y}, \qquad y \epsilon \{0, 1\}$$

Таким чином випадкова величина y має розподіл Бернуллі:

$$y \sim B\big(f(\theta^T x)\big)$$

Оцінка параметрів $\theta_0, \theta_1, \ldots \theta_n$ здійснюється за допомогою методу максимальної правдоподібності. Параметри θ обираються, так щоб максимізувати значення функції правдоподібності на вибірці:

$$\hat{\theta} = \arg\max x_\theta L(\theta) = \arg\max x_\theta \prod \mathbb{P}\{y = y^{(i)} \mid x = x^{(i)}\}$$

, що еквівалентно максимізації логарифма цієї функції:

$$logL(\theta) = \sum log\mathbb{P}\{y = y^{(i)} \mid x = x^{(i)}\}$$

2.1.2 Наївний баєсівський класифікатор

У статистиці наївні класифікатори Байєса — це сімейство лінійних імовірнісних класифікаторів, заснованих на застосуванні теореми Байєса з сильними (наївними) припущеннями незалежності між ознаками. Вони є одними з найпростіших байєсівських мережевих моделей, але в поєднанні з оцінкою щільності ядра вони можуть досягти високого рівня точності.

Наївна Байєсова модель є умовною імовірнісною моделлю: вона призначає ймовірності $p(C_k \mid x_1, x_2 \dots x_n)$ для кожного з можливих результатів або класів заданого екземпляру, який потрібно класифікувати, що представлений вектором кодування деяких ознак.

Основна проблема вище зазначеного формулювання полягає в тому, що за умови великої кількості ознак або ж якщо ознаки приймають велику кількість значень, то застосування моделі на таблицях ймовірності стає неможливим. В цьому випадку необхідно переформулювати модель, щоб зробити її зручнішою. Використовуючи теорему Байєса, розкладемо умовну ймовірність як:

$$p(C_k \mid x) = \frac{p(C_k)p(x \mid C_k)}{p(x)}$$

На практиці цікавить лише чисельник цього дробу, оскільки знаменник не залежить від значення C і ознак x_i, так що знаменник фактично постійний. Чисельник еквівалентний моделі спільної ймовірності:

$$p(C_k, x_1, x_2 \dots x_n)$$

, яке можна переписати таким чином, використовуючи правило ланцюга для повторних застосувань визначення умовної ймовірності:

$$p(C_k, x_1, x_2 \dots x_n) = p(x_1, x_2 \dots x_n, C_k) = p(x_1 \mid x_2 \dots x_n, C_k)p(x_2 \dots x_n, C_k)$$
$$= p(x_1 \mid x_2 \dots x_n, C_k)p(x_2 \mid x_3 \dots x_n, C_k)p(x_3 \dots x_n, C_k) =$$
$$= \cdots = p(x_1 \mid x_2 \dots x_n, C_k)p(x_2 \mid x_3 \dots x_n, C_k) \dots p(x_{n-1} \mid x_n, C_k)p(x_n, C_k)$$

Тепер у гру вступають “наївні” припущення про умовну незалежність: припустимо, що всі функції x взаємно незалежні, обумовлені категорією C_k. Згідно з цим припущенням:

$$p(x_i \mid x_{i+1} \dots x_n, C_k) = p(x_i, C_k)$$
$$p(C_k)p(x \mid C_k) = p(C_k) \prod p(x_i, C_k)$$

Для побудови класифікатора з ймовірнісної моделі наївний класифікатор Байєса поєднується із правилом прийняття рішень. Загальне правило полягає в тому, щоб вибрати гіпотезу, яка є найбільш ймовірною, щоб мінімізувати ймовірність неправильної класифікації; це відоме як правило максимального

апостеріорного або MAP рішення. Відповідний класифікатор, класифікатор Байєса, є функцією, яка призначає мітку класу $\hat{y} = \mathrm{C}_k$ наступним чином:

$$\hat{y} = \arg\max_{k \in \{1,2 \ldots K\}} p(C_k) \prod p(x_i, C_k)$$

Пріор класу може бути обчислений, припускаючи рівно імовірні класи, тобто $p(C_k) = \frac{1}{K}$ або шляхом розрахунку оцінки ймовірності класу з навчального набору.

2.1.3 Дерева прийняття рішень

Дерево прийняття рішень представляє собою ієрархічну модель для підтримки процесу прийняття рішень. Ця модель використовує деревоподібну структуру для представлення рішень та їхніх можливих наслідків, охоплюючи результати випадкових подій, витрати ресурсів і корисність. Вона є одним із методів відображення алгоритму, що базується виключно на умовних операторах керування.

Зазвичай дерева рішень використовуються в області дослідження операцій, зокрема в аналізі рішень, для допомоги у визначенні стратегії, яка максимально ймовірно досягне поставленої мети. Також цей підхід популярний у сфері машинного навчання.

Дерева рішень, які використовуються в інтелектуальному аналізі даних, бувають двох основних типів:

- **дерева класифікації;**
- **дерева регресії.**

Термін дерево класифікації та регресії (Classification And Regression Tree, CART) — це загальний термін, який використовується для позначення будь-якої з наведених вище процедур, вперше введений Брейманом у 1984 році.

Алгоритм CART призначений для створення бінарних дерев рішень, де кожен вузол при розгалуженні має лише два нащадки. У випадку бінарних

дерев, кожен об'єкт може перейти в один з двох нащадків на основі певного критерію. Для алгоритму CART поведінка об'єктів у виділеній групі визначається часткою модального (найчастішого) значення вихідної ознаки. Групи вважаються виділеними, якщо частка цього значення є достатньо великою

Перевагою алгоритму CART є певна гарантія того, що у випадку наявності шуканих детермінацій у досліджуваній сукупності вони будуть виявлені. Крім того, CART дозволяє не обмежуватися лише одним значенням вихідної ознаки, а здійснювати пошук усіх можливих значень, для яких можна визначити відповідний пояснюючий вираз.

Алгоритми побудови дерева рішень зазвичай працюють зверху вниз шляхом вибору змінної на кожному кроці, яка найкраще розбиває безліч елементів. Різні алгоритми використовують різні метрики для виміру кращого рішення. Вони зазвичай вимірюють однорідність цільової змінної на підмножинах. Ці метрики застосовуються до кожного підмножини і значення комбінуються (наприклад, обчислюється середнє) для отримання міри якості розбиття.

Використовуваний в алгоритмі дерев класифікації та регресії критерій Джині є мірою, наскільки часто випадково обраний елемент з набору неправильно позначається, якщо він випадково позначається згідно з розподілом міток у підмножині. Критерій Джині може бути обчислений шляхом підсумовування ймовірності p_i елемента з обраною i міткою, помноженої на ймовірність $\sum_{i \neq i} p_k = 1 - p_i$ помилки категоризації цього елемента. Критерій приймає мінімум, коли всі випадки у вузлі потрапляють до однієї цільової категорії.

Для обчислення критерію Джині для набору елементів з K класами:

$$I_G(p) = \sum p_i \sum_{k \neq i} p_k = \sum p_i(1 - p_i) = 1 - \sum {p_i}^2$$

Переваги:

1. Даний метод є непараметричним, це означає, що для його застосування немає необхідності розраховувати різні параметри імовірнісного розподілу.
2. Для застосування алгоритму CART немає необхідності заздалегідь вибирати змінні, які братимуть участь у аналізі: змінні відбираються безпосередньо під час аналізу на підставі значення індексу Джині.
3. CART легко бореться з викидами: механізм розбиття, закладений в алгоритмі, просто поміщаючи викиди в окремий вузол, що дозволяє очистити наявні дані від шумів.
4. Для застосування цього алгоритму не треба брати до уваги жодних припущень чи припущень перед проведенням аналізу.
5. Великою перевагою є швидкість роботи алгоритму.

Недоліки:

1. Дерева рішень, запропоновані алгоритмом, не є стабільними: результат, отриманий на одній вибірці, не відтворюється на іншій (дерево може збільшуватися, зменшуватися, включати інші предиктори і т.д.)
2. У випадку, коли необхідно побудувати дерево з складнішою структурою, краще використовувати інші алгоритми, оскільки CART може не ідентифікувати правильну структуру даних.

2.1.4 Випадкові ліси

Випадкові ліси, або ліси випадкових рішень, представляють собою метод ансамблевого навчання, який застосовується для класифікації, регресії та інших завдань. Цей метод досягає своєї ефективності шляхом створення множини дерев рішень під час процесу навчання.

Алгоритм навчання випадкових лісів використовує загальну техніку бутстрап-агрегування, або пакування, для дерев, які навчаються. Під час цього

процесу з навчальної множини з відповідями багаторазово вибирається випадкова вибірка замість навчальної множини, і для цих вибірок будуються окремі дерева.

Після завершення навчання можна здійснювати прогнози для зразків, усереднюючи прогнози всіх окремих дерев регресії.

$$\hat{f} = \frac{1}{B}\sum f_b(x)$$

або шляхом прийняття більшості голосів у випадку класифікаційних дерев.

Ця процедура початкового завантаження сприяє покращенню продуктивності моделі, оскільки вона зменшує дисперсію моделі, при цьому не збільшуючи зміщення. Іншими словами, якщо передбачення одного дерева сильно реагують на шум у вибірці, то середнє значення багатьох дерев буде менш чутливим до цього шуму, за умови, що дерева некорельовані. Зазвичай, просте навчання багатьох дерев на одному і тому ж навчальному наборі призводить до високої кореляції між ними (навіть до того рівня, коли одне й те саме дерево може бути отримано багаторазово, якщо алгоритм є детермінованим).

2.1.5 Градієнтний бустинг

Градієнтний бустинг – це техніка машинного навчання, яка використовується, зокрема, у завданнях регресії та класифікації. Він дає модель прогнозування у формі ансамблю слабких моделей прогнозування, тобто моделей, які роблять дуже мало припущень щодо даних, які зазвичай є простими деревами рішень. Якщо дерево рішень є слабким навчальним елементом, отриманий алгоритм називається деревом із посиленням градієнта; зазвичай він перевершує випадковий ліс.

Метод градієнтного бустингу передбачає дійсне значення y. Він шукає: $\hat{F}(x)$ – наближення у вигляді зваженої суми M функцій якогось класу, які називаються базовими (або слабкими) учнями:

$$\hat{F}(x) = \sum \gamma_m h_m(x) + const$$

Відповідно до принципу мінімізації емпіричного ризику, метод намагається знайти наближення $\hat{F}(x)$ що мінімізує середнє значення функції втрат на навчальній множині, тобто мінімізує емпіричний ризик. Це робиться, починаючи з моделі, що складається з постійної функції $F_0(x)$, і поступово розширює його жадібним способом:

$$F_0(x) = \arg\min_\gamma \sum L(y_i, \gamma)$$

$$F_m(x) = F_{m-1}(x) + (\arg\min_{h_m \epsilon H}[\sum L(y_i, F_{m-1}(x_i) + h_i(x_i))])(x)$$

На жаль, вибір найкращої функції h_m на кожному кроці для довільної функції втрат є обчислювально нездійсненною проблемою оптимізації в цілому, тому обмежимося підходом до спрощеної версії задачі. Ідея полягає в тому, щоб застосувати градієнтний спуск до цієї проблеми мінімізації (функціональний градієнтний спуск). Основна ідея полягає в тому, щоб знайти локальний мінімум функції втрат шляхом повторення $F_{m-1}(x)$.

$$F_m(x) = F_{m-1}(x) - \gamma \sum \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i)), \qquad \gamma > 0$$

Крім того, ми можемо оптимізувати γ знайшовши γ_m, для якого функція втрат має мінімум:

$$\gamma_m = argmin_\gamma \sum L\big(y_i, F_m(x_i)\big) =$$

$$= argmin_\gamma \sum L\big(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))\big)$$

Якщо розглядати неперервний випадок, модель було б оновлено відповідно до наступних рівнянь:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))$$

Однак у дискретному випадку, функцію-кандидата h обраємо, як найближчу до градієнта L. В даному випадку алгоритм набуває наступного вигляду:

1. Обираємо модель з постійним значенням

$$F_0(x) = argmin_\gamma \sum L(y_i, \gamma)$$

2. На кожному кроці $m = 1, 2 \dots M$

   a. Обчислюємо псевдозалишки:

$$r_{im} = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}$$

   b. Підібрати базовий навчальний елемент (або слабкий навчальний елемент, наприклад, дерево) $h_m(x)$ наближений до псевдозалишків, тобто навчити його за допомогою навчального набору $\{(x_i, r_{im})\}$

   c. Обчислити γ_m, розв'язавши задачу одновимірної оптимізації:

$$\gamma_m = argmin_\gamma \sum L\big(y_i, F_{m-1}(x_i) - \gamma h_m(x_i)\big)$$

   d. Оновити модель:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

3. Вихід з моделлю $F_M(x)$

2.2 Оцінка Каплана-Майєра

Аналіз виживаності – це галузь статистики, що вивчає очікуваний проміжок часу до настання певної події, зокрема смерті біологічних організмів або відмови механічних систем. У різних наукових галузях цей аналіз може називатися теорією надійності або аналізом надійності (у інженерії), аналізом тривалості чи моделюванням тривалості (у економіці) та аналізом історії подій (у соціології).

Щоб відповісти на такі питання, необхідно дати визначення терміну життя. Теорія, викладена нижче, передбачає чітко визначені події в певний

час; інші випадки краще розглядати за допомогою моделей, які явно враховують неоднозначні події.

Оцінка Каплана–Майєра, також відома як оцінка обмеження продукту, є непараметричною статистикою, яка використовується для оцінки функції виживання на основі даних за весь період життя. У медичних дослідженнях його часто використовують для вимірювання частки пацієнтів, які живуть протягом певного часу після лікування. В інших галузях оцінка Каплана–Майєра можуть використовуватися для вимірювання тривалості часу, протягом якого люди залишаються безробітними після втрати роботи або, як в нашому випадку, час до настання дефолту по кредитній лінії.

Оцінка функції виживання $S(t)$ (імовірність того, що життя довше ніж t) задано:

$$S(t) = \prod_{i;\, t_i \le t} (1 - \frac{d_i}{n_i})$$

t_i – час, коли сталась принаймні одна подія

d_i – кількість подій (наприклад, настання дефолту)

n_i – кількість осіб, про яких відомо, що вони «вижили» (ще не мали подій або не були піддані цензурі) на момент часу t_i

Графік оцінки Каплана–Майєра являє собою серію спадаючих горизонтальних кроків, які за достатньо великого розміру вибірки наближаються до справжньої функції виживання для цієї сукупності. Значення функції виживання між послідовними різними вибірковими спостереженнями вважається постійним.

Важливою перевагою кривої Каплана–Майєра є те, що метод може враховувати деякі типи цензурованих даних, зокрема праву цензуру, яка виникає, якщо особи виходить (наприклад, дострокове погашення позики) з дослідження або ж втрачається для подальшого спостереження без виникнення події під час останнього спостереження. Якщо не відбувається

скорочення чи цензурування, крива Каплана–Майєра є доповненням до емпіричної функції розподілу .

Висновки до розділу 2

У розділі, присвяченому теоретичному апарату, були розглянуті різноманітні підходи до задачі класифікації та прогнозування. Кожен із розглянутих методів має свої унікальні особливості та застосування, що робить їх ефективними в різних сценаріях.

Логістична регресія виявляється досить ефективною для бінарної класифікації, адже вона добре вирішує завдання прогнозування ймовірностей. Баєсівський класифікатор використовує ймовірнісний підхід, враховуючи апріорні ймовірності класів, і може бути корисним в умовах обмеженого обсягу даних. Дерева прийняття рішень володіють інтерпретованістю та легкістю зрозуміння, але можуть бути схильними до перенавчання. Випадкові ліси вирішують цю проблему шляхом ансамблювання декількох дерев, що забезпечує стабільніші та точніші результати. Градієнтний бустинг використовує послідовне навчання слабких моделей, покращуючи результат на кожному кроці. Його ефективність полягає в здатності пристосовуватися до складних взаємодій у даних.

Оцінка Каплана-Майєра є важливим інструментом для аналізу виживання та ризиків у статистиці, надаючи можливість оцінки часу до подій та порівняння груп.

У цьому розділі було визначено, що вибір конкретного методу чи моделі залежить від конкретної задачі, обсягу даних, характеру змінних та інших факторів. Комбінування різних методів або використання ансамблів може покращити результати моделювання.

РОЗДІЛ 3 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНИХ ЕКСПЕРИМЕНТІВ ОЦІНЮВАННЯ РИЗИКІВ КРЕДИТУВАННЯ

У цьому розділі детально аналізується процес розробки оптимальної моделі для прогнозування та оцінки кредитних ризиків. Висвітлено ключові етапи підготовки даних, вибору та налаштування моделей, які допомагають у визначенні найефективніших стратегій в управлінні ризиками.

Крім того, в даному розділі розглянуто можливі сценарії та стратегія реагування на них, які ґрунтуються на отриманих результатів застосування розробленого аналітичного інструментарію. Ця стратегія спрямована на максимізацію ефективності управління кредитними ризиками, а також може включати рекомендації щодо прийняття рішень у реальному часі на основі аналізу показників.

3.1 Огляд даних для статистичного аналізу та прогнозування

Прогнозування - це формальний аналіз стану досліджуваної системи чи процесу протягом одного, двох чи більше циклів часу відносно теперішнього на основі відповідного статистичного аналізу. Суть прогнозу полягає в наданні науково підкріпленого опису майбутніх подій. Це означає, що прогнози виходять за межі простої передбачуваності і ґрунтуються на наукових доказах, які можна відтворювати незалежно від прогнозиста.

Експертна оцінка, або прогноз професіонала, є одним з проміжних методів формування концепції майбутнього. Важливо враховувати об'єктивність інтерпретації в межах наукового обґрунтування експерта, оскільки ця оцінка поєднує індивідуальне суб'єктивне сприйняття експертом потенційного розвитку прогнозованого процесу та розглядає нездійсненні та несистематизовані фактори.

Статистичні методи аналізу та прогнозу включають в себе глибоку обробку статистичних даних, пов'язаних із вивченим процесом. У даному випадку мають місце такі характеристики дослідження.

1. Більшість вихідних статистичних показників походить із **первинних та вторинних джерел**. **Первинні джерела** включають спеціалізовані соціологічні опитування, переписи та обстеження, які забезпечують дані у форматі, необхідному для управлінських або прогнозних розрахунків. **Вторинні джерела** включають опубліковані дані, що вже були зібрані іншими особами, і хоча вони можуть бути не пов'язані безпосередньо з основним завданням прогнозиста, але все ж важливі для його виконання. Перший етап збору статистичних даних з первинних джерел передбачає чітко сплановану роботу, включаючи виділення спеціальних коштів. Визначення складу показників, розробка плану вибірки та передбачені фіксовані параметри є обов'язковими. В другому випадку вторинні джерела містять вихідні дані, які були вже зібрані іншими особами, не пов'язаними безпосередньо з основним завданням прогнозування, але, тим не менш, можуть мати важливе значення для виконання цього завдання.

2. Умови для джерел статистичних даних. З метою забезпечення високих стандартів якості у сфері статистики, важливо збирати різноманітні статистичні дані як з первинних, так і з вторинних джерел, які відповідають ряду умов.

- **актуальність;**
- **точність і надійність**;
- **репрезентативність.**

Вихідними даними для побудови моделей використано данні однорангової кредитної компанія Lending Club. **Lending Club** - це фінансова компанія, розташована в США, яка працює на принципі однорангового позикодавця. Вони здійснюють посередництво між особами, які мають бажання інвестувати свої кошти, і тими, хто шукає кредит. Інвестори, які вкладають свої гроші через Lending Club, позичають ці кошти іншим особам,

які потребують фінансування. Коли позичальники повертають позики, включаючи відсотки, ці платежі передаються інвесторам. Ця модель приносить вигоду всім учасникам, оскільки зазвичай дозволяє отримати низькі ставки для позичальників та прибуток для інвесторів.

Даний набір даних від Lending Club містить повну інформацію про всі видані позики за період з 2007 по 2015 рік, включаючи інформацію про статус позики та інформацію про останні платежі. Серед змінних в наборі даних можна знайти кредитні рейтинги, кількість фінансових запитів, адреси, включаючи поштові індекси та штати, і дані про збори, спрямовані на відшкодування заборгованості. Загалом, набір даних складається з близько 396030 спостережень і 28 змінних.

3.2 Підготовка вихідних даних

Один із ключових кроків у процесі оцінки ризику - це підготовка даних. Завершена оцінка ризику в значній мірі залежить від обраних вхідних параметрів та їх якості, які використовуються для прогнозування. У цьому дослідженні основою є опція аналізу ймовірності настання негативної події, такої як дефолт, і очікуваного часу її настання. Цей підхід дозволяє оцінити як очікувані, так і неочікувані ризики, пов'язані з видачею позики. Для цього використовуються неперсоніфіковані дані, які містять різноманітні соціальні та фінансові показники, що стосуються особи та умови кредитування, включаючи:

- **Числові:**
  - **Дискретні:**
    - open_acc – кількість відкритих кредитних ліній у кредитній справі позичальника.
    - pub_rec – кількість негативних публічних записів

- total_acc – загальна кількість кредитних ліній у кредитній справі позичальника на даний момент
- mort_acc – кількість іпотечних рахунків.
- pub_rec_bankruptcies – кількість публічних банкрутств
- default_term – період настання дефолту

- **Неперервні:**
  - loan_amnt – зазначена сума позики, на яку просить позичальник. Якщо кредитний відділ зменшить суму кредиту, то це буде відображено в цьому значенні.
  - int_rate – ставка за кредитом
  - installment – щомісячний платіж, який повинен сплатити позичальник у разі отримання кредиту.
  - annual_inc – річний дохід, вказаний позичальником під час реєстрації.
  - dti – співвідношення, розраховане з використанням загальних місячних платежів позичальника на загальну суму боргових зобов'язань, за винятком іпотеки та запитаної кредитної позики, поділених на місячний дохід позичальника, який сам звітує.
  - revol_bal – загальний поновлюваний баланс кредиту
  - revol_util – коефіцієнт використання поновлюваної лінії або сума кредиту, яку використовує позичальник, відносно всього наявного поновлюваного кредиту.

- **Категоріальні:**
  - **Номінальні:**
    - term – кількість платежів за кредитом.
    - emp_title – назва посади, яку вказує позичальник під час подання заявки на кредит
    - home_ownership – статус власності на житло, наданий позичальником під час реєстрації

- verification_status – вказує, чи був дохід підтверджений компанією, джерелом доходу або ж не підтверджено
- loan_status – статус кредиту
- purpose – категорія, яку надає позичальник для запиту на позику
- title – назва позики, надана позичальником
- initial_list_status – статус позики і визначає, чи була позика спочатку зареєстрована на повному чи частковому ринку
- application_type – вказує, чи є кредит індивідуальною заявкою чи спільною заявкою з двома співпозичальниками
- address – адреса реєстрації позичальника

- **Порядкові:**
  - grade – категорія кредитний рейтинг
  - sub_grade – підкатегорія кредитного рейтингу
  - emp_length – тривалість трудового досвіду

В таблиці 3.1 наведено загальну інформацію про дані.

Таблиця 3.1 – Загальна інформація про вибірку

№	Column	Non-Null Count	DType
1	loan_amnt	396030	float64
2	term	396030	object
3	int_rate	396030	float64
4	installment	396030	float64
5	grade	396030	object
6	sub_grade	396030	object
7	emp_title	373103	object
8	emp_length	377729	object
9	home_ownership	396030	object
10	annual_inc	396030	float64
11	verification_status	396030	object

Продовження таблиці 3.1

12	issue_d	396030	object
13	loan_status	396030	object
14	default_term	77673	float64
15	purpose	396030	object
16	title	394274	object
17	dti	396030	float64
18	earliest_cr_line	396030	object
19	open_acc	396030	float64
20	pub_rec	396030	float64
21	revol_bal	396030	float64
22	revol_util	395754	float64
23	total_acc	396030	float64
24	initial_list_status	396030	object
25	application_type	396030	object
26	mort_acc	358235	float64
27	pub_rec_bankruptcies	395495	float64
28	address	396030	object

Перший і надзвичайно важливий етап в аналізі даних та в будь-якому дослідженні, пов'язаному з інформацією, - це очистка даних. Цей процес визначає якість та надійність отриманих результатів і впливає на кінцевий успіх проекту. В цьому етапі відбувається перевірка даних на наявність помилок, неточностей та видалення зайвої інформації.

Очищення даних включає в себе кілька ключових аспектів.

1. **Обробка відсутніх даних:** потрібно вирішити, що робити з відсутніми значеннями. Це може включати в себе їх заповнення середніми значеннями, медіаною чи видаленням відповідних записів.

2. **Усунення дублікатів:** дублікати даних можуть призвести до перенавчання або незбалансованого набору даних. Вони повинні бути ідентифіковані і вилучені.

3. **Перевірка формату даних:** впевніться, що дані відповідають встановленому формату. Наприклад, дати мають бути в однаковому стандарті, а текстові дані - відформатовані правильно.

4. **Первинний аналіз даних:** перевірка даних на відповідність дійсним обмеженням та правилам. Наприклад, переконайтеся, що числові дані в межах реалістичних значень.

Перший етап підготовки даних успішно завершено, забезпечено їх коректний формат та гарантовано, що вони не містять небажаних дублікатів. Тепер перед нами стоїть завдання обробити відсутні дані та провести їх валідацію. Ці наступні кроки є критичними для забезпечення якості та достовірності нашого аналізу.

Обробка відсутніх даних є важливою (на рисунку 3.1 продемонстровано кількість відсутніх даних в вибірці), оскільки вони можуть вплинути на точність наших висновків. Існує кілька підходів до цього завдання. Наприклад, відновити відсутні даних, заповнивши їх середніми значеннями, медіаною чи іншими статистичними показниками. Цей процес допоможе уникнути втрати важливої інформації та зберегти репрезентативність наших даних.

```
'emp_title':
        number of missing values '22927' ==> '5.789%'
'emp_length':
        number of missing values '18301' ==> '4.621%'
'default_term':
        number of missing values '318357' ==> '80.387%'
'title':
        number of missing values '1756' ==> '0.443%'
'revol_util':
        number of missing values '276' ==> '0.070%'
'mort_acc':
        number of missing values '37795' ==> '9.543%'
'pub_rec_bankruptcies':
        number of missing values '535' ==> '0.135%'
```

Рисунок 3.1 – Абсолютна та відносна кількість пропущених даних

У нашому наборі даних спостерігається наступна особливість: кількість унікальних записів у полі **emp_title** становить 173105. Зрозуміло, що це занадто велика кількість унікальних назв посад та спроба перетворити всі ці унікальні назви на фіктивну змінну ознаки була б надзвичайно ресурсомістким завданням і, імовірно, не давала б великої користі для аналізу даних. Більше того, це може призвести до надмірного розширення набору даних та вплинути на ефективність аналітичного процесу.

Тому прийнято рішення видалити поле **emp_title** із нашого набору даних. Це дозволить зосередитися на інших змінних ознак та спростить аналіз. Важливо завжди враховувати практичність та ефективність обробки даних у процесі вивчення інформації, і прийняття рішень щодо видалення занадто розширених змінних.

Перед тим як визначити метод відновлення відсутніх даних в ознаці **emp_length**, необхідно провести її аналіз. Цей аналіз може бути вирішальним для вибору найкращого підходу до відновлення пропущених даних. Після ретельного дослідження виявилося, що не існує статистично значущих різниць між різними категоріями в ознаці **emp_length** (таблиця 3.2). Це означає, що тривалість роботи не впливає на інші характеристики або результати в даному контексті.

Таблиця 3.2 – Статус кредиту відповідно стажу роботи

	Fully Paid	**Charged Off**
< 1 року	0.79	0.21
2 роки	0.81	0.19
3 роки	0.80	0.20
4 роки	0.81	0.19
5 років	0.81	0.19
6 років	0.81	0.19
7 років	0.81	0.19

Продовження таблиці 3.2

8 років	0.80	0.20
9 років	0.80	0.20
> 10 років	0.82	0.18

З урахуванням цього аналізу та відсутності значущої інформації, прийнято рішення видалити ознаку **emp_length** із нашого набору даних. Це дозволить спростити аналіз, зменшити розмірність даних та покращити продуктивність моделі без втрати інформації або значущого впливу на результати нашого дослідження.

У випадках, коли певні ознаки не приносять значущого внеску до аналізу або не впливають на результати, їх видалення може бути розумним кроком для оптимізації обробки даних та моделювання.

Також було прийнято рішення замість відновлення ознаки **title** видалити її з набору даних. Це рішення було обґрунтовано рядом факторів, які вказують на те, що інші ознаки можуть краще відображати інформацію про ціль кредиту.

Одним із головних аргументів є наявність ознаки **purpose**, яка була згенерована на основі ознаки **title**. Ознака **purpose** у чіткіший та більш зрозумілий спосіб характеризує мету кредиту та визначає її за ключовими категоріями. Видалення ознаки спрощує набір даних та зменшує розмірність без втрати важливої інформації. Додатково, це рішення також сприяє уникненню зайвої кореляції між ознаками, що могло б призвести до перенавчання моделі. Видалення ознаки **title** допомагає у зниженні ризику збурення моделі і поліпшенні її узагальнюючої здатності.

Враховуючи ці аргументи, видалення ознаки **title** було прийнято як логічний крок для покращення обробки даних та моделювання, забезпечуючи більш точний та ефективний аналіз цілей кредиту у відповідності до їх ключових категорій.

Ознаки **revol_util** та **pub_rec_bankruptcies** мають надзвичайно малу кількість відсутніх даних. У цьому випадку прийнято рішення видалити записи, в яких відсутні значення в цих ознаках.

Вибір видалення записів був обґрунтований декількома факторами. По-перше, оскільки об'єм відсутніх даних був надзвичайно малим, ця дія не вплине на загальний обсяг даних великою мірою. По-друге, видалення записів з відсутніми значеннями в цих ознаках дозволяє нам попрацювати з більш чистим та повним набором даних, що може покращити якість та надійність аналітичних результатів.

Серед невідомих даних з пропусками виявилася лише одна змінна, а саме **mort_acc**. Важливо пояснити, що пропуски в ознаці **default_term** обґрунтовані відсутністю настання події, а саме дефолту, та не можуть бути відновлені.

Під час розв'язання цього завдання, розглянуто кілька можливих підходів до заповнення пропущених даних, враховуючи, що немає жодного універсального методу, що підходить до всіх ситуацій.

Один із можливих підходів - це побудова простої моделі для заповнення пропущених даних, наприклад, лінійної регресії. Інший підхід - це заповнення пропущених значень на основі середнього, медіани або іншої статистики; розбиття стовпців на категорії та встановлення NaN як окремої категорії. Варіантів багато, і важливо розглядати їх залежно від конкретного контексту і даних.

У випадку **mort_acc**, здавалося, що існує кореляція з ознакою **total_acc** (продемонстровано на рисунку 3.2). Тому вирішено згрупувати дані за ознакою **total_acc** і обчислити середнє значення **mort_acc** в межах кожної категорії. Цей підхід дозволяє використовувати інформацію з інших записів, щоб заповнити відсутні значення в даному записі. Цей метод може бути більш точним і відповідати логіці даних більше, ніж деякі інші підходи.

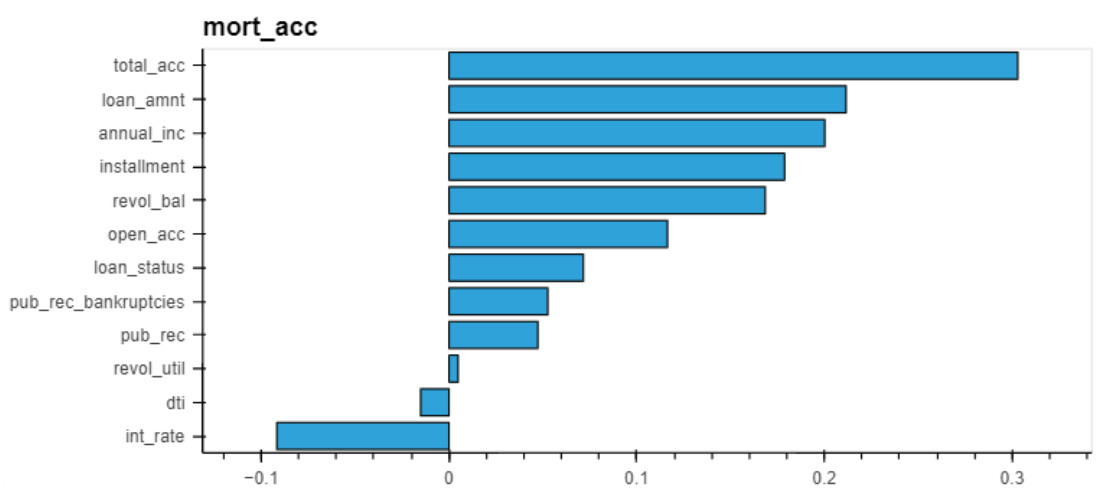

Рисунок 3.2 – Кореляція ознаки mort_acc

Проте варто пам’ятати, що не існує одного "правильного" методу розв'язання пропущених даних, і вибір методу залежить від конкретної ситуації та мети аналізу. У даному випадку, використання кореляції між **mort_acc** і **total_acc** видається логічним і обґрунтованим підходом для заповнення пропущених значень.

Після обробки відсутніх даних ми переходимо до валідації. Цей крок передбачає перевірку даних на відповідність дійсним обмеженням та правилам. Ми впевнимося, що числові дані знаходяться в межах реалістичних значень та що текстові дані відповідають встановленому формату. Цей процес гарантує нам, що ми працюємо з надійними та достовірними даними, на яких можна будувати нашу подальшу аналітичну роботу. Завершення обробки відсутніх даних та їх валідація позначають важливий етап у підготовці даних для подальшого використання. Це дозволяє нам рухатися вперед і проводити аналіз, спираючись на здорові та надійні дані. Попередньо розглянемо кореляцію між даними (теплова карта кореляції зображено на рисунку 3.3).

Рисунок 3.3 – Теплова карта кореляції даних

В даних присутня майже ідеальна кореляція між змінними **loan_amnt** та **installment**, дослідимо ці характеристики детальніше (на рисунку 3.4, 3.5 продемонстровано гістограму даних та діаграму розмаху цих ознак відповідно):

- **loan_amnt** – зазначена сума позики, на яку просить позичальник;
- **installment** – щомісячний платіж, який повинен сплатити позичальник у разі отримання кредиту.

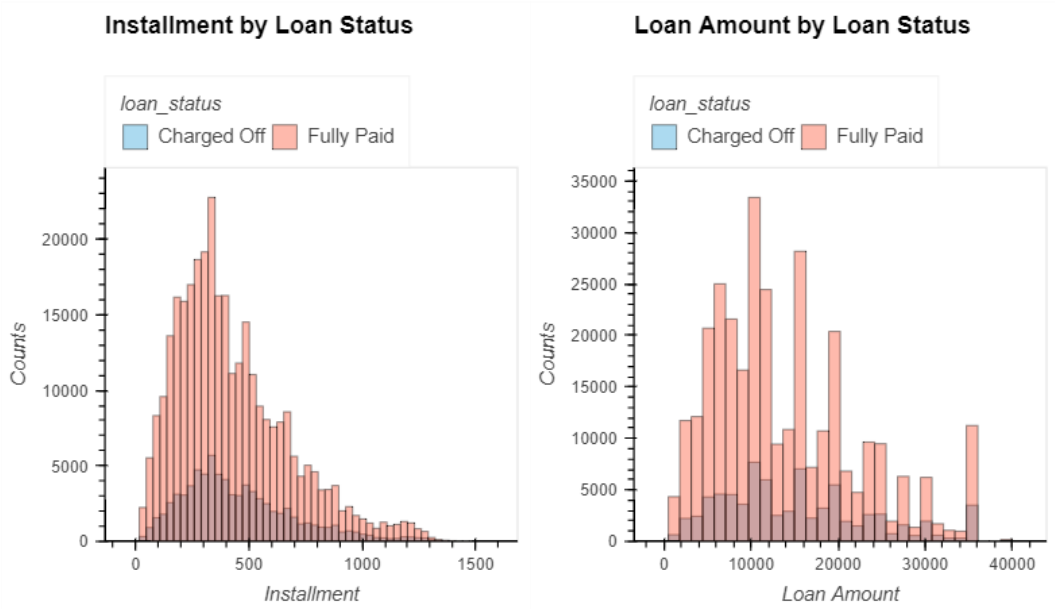

Рисунок 3.4 – Гістограма даних **loan_amnt** і **installment**

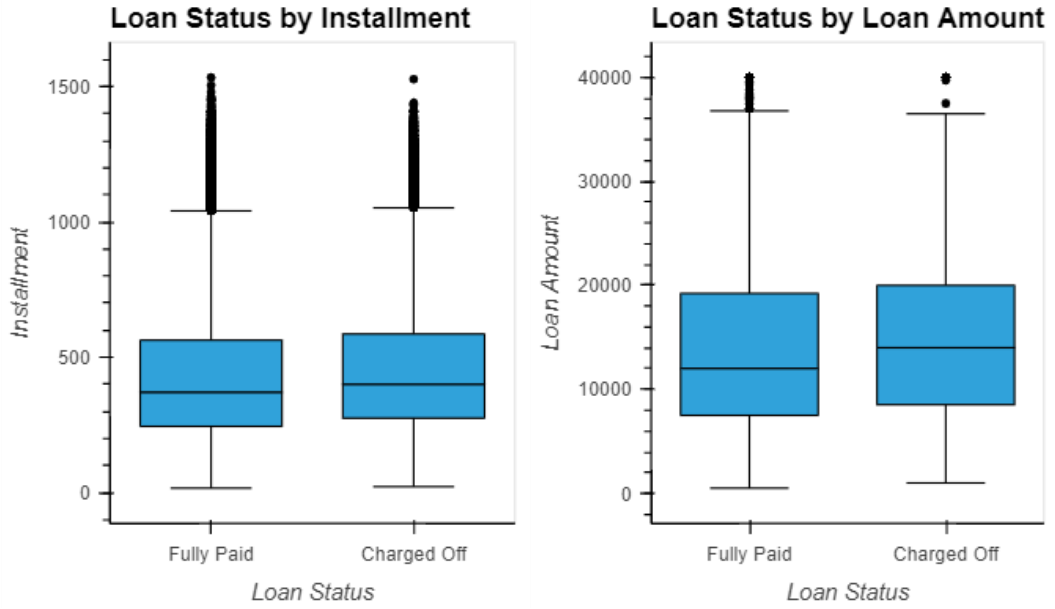

Рисунок 3.5 – Діаграма розмаху даних **loan_amnt** і **installment**

Вже з аналізу опису змінних стає зрозумілою причина високої кореляції між ними.

Наступним етапом нашого дослідження буде аналіз кореляції між статусом виплати кредиту та кредитним рейтингом позичальника. Кредитний рейтинг позичальника є однією з ключових характеристик, яка має суттєвий вплив на ймовірність повернення позики та рішення кредитора щодо надання кредиту. Цей рейтинг відображає кредитну історію та надійність позичальника і відіграє важливу роль у фінансових операціях.

Основна ціль нашого аналізу полягатиме в встановленні, чи існує статистично значуща кореляція між кредитним рейтингом та статусом виплати кредиту. Це дозволить нам з'ясувати, чи існують зв'язки між рівнем надійності позичальників і їхньою здатністю вчасно повертати кредити. Відомо, що кредитний рейтинг може бути індикатором фінансової відповідальності, і відповідно, його важливість у сфері кредитування є високою. На рисунку 3.6 наведено розподіл виплат кредиту відповідно кредитного рейтингу (класу та підкласу).

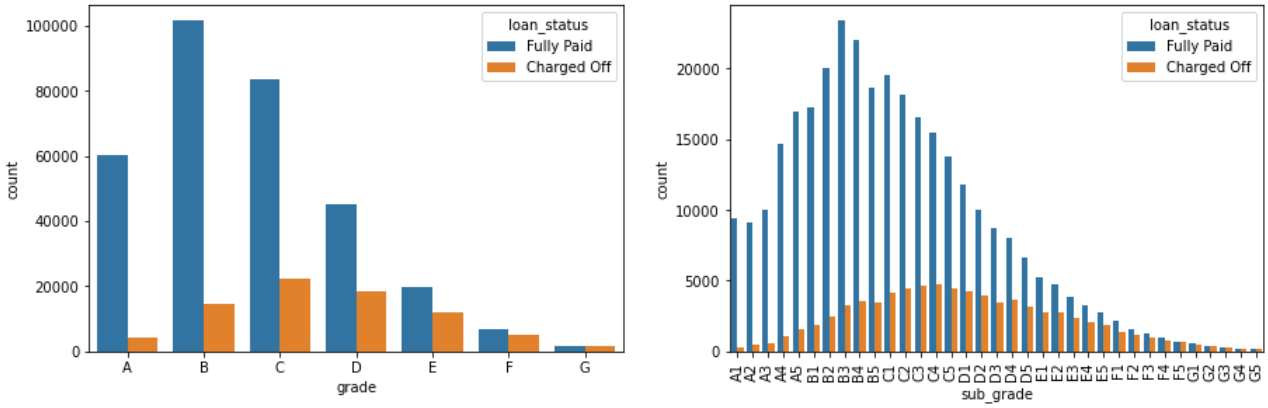

Рисунок 3.6 – Розподіл виплати кредиту відповідно кредитного класу

Результати дослідження також показали, що підкласифікація кредитного рейтингу в середньому мало відрізняється всередині групи. Це означає, що для подальших досліджень можливо обмежитися лише основним класом позичальника, не розглядаючи додаткові підкатегорії. Такий підхід дозволить нам зосередитися на основних тенденціях і виявити кореляційні зв'язки між статусом виплати кредиту та кредитним рейтингом, що мають ключове значення для нашого аналізу та прийняття рішень у сфері кредитування.

У нашому наступному етапі дослідження ми спрямуємо увагу на аналіз кореляції між типом власності житла, такими як власність або оренда, і статусом виплати кредиту. Це дослідження важливо для розуміння взаємозв'язку між умовами проживання позичальників та їхньою спроможністю вчасно відшкодовувати кредити.

Статус власності житла може бути ключовим фактором, який впливає на фінансову стабільність та здатність позичальників виконувати свої зобов'язання перед кредиторами. Наприклад, особи, які мають власність, можуть бути більш фінансово стабільними, оскільки вони мають значний актив, який може бути використаний в разі фінансових труднощів. З іншого боку, орендарі можуть стикатися з іншими фінансовими викликами, і це може вплинути на їхню спроможність вчасно повертати кредити.

Дослідження кореляції між типом власності житла та статусом виплати кредиту (дивитися рисунок 3.7) дозволить нам з'ясувати, чи існують певні тенденції або залежності між цими двома змінними. Це може бути корисною інформацією для кредиторів та фінансових установ, які мають на меті оцінити ризики та приймати рішення щодо надання кредитів.

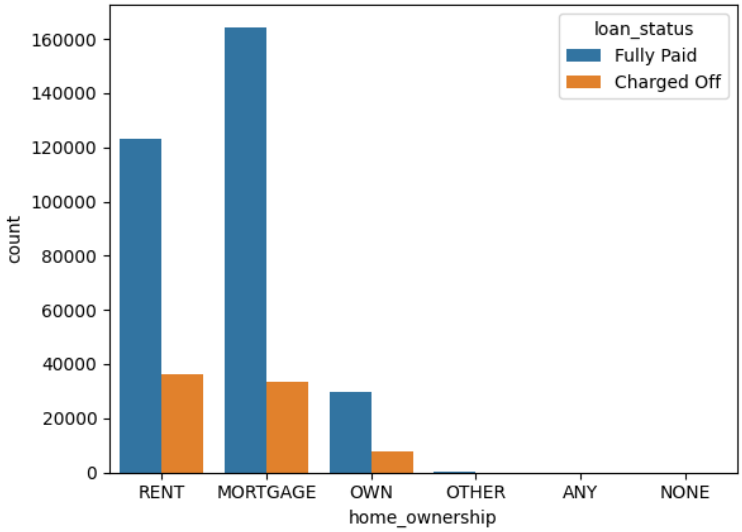

Рисунок 3.7 – Розподіл виплати кредиту відповідно типу власності житла

Проведений аналіз даних надав цінну інформацію щодо кореляції між класом власності житла та статусом виплати кредиту. Результати аналізу підтвердили присутність взаємозв'язку між цими двома факторами. У процесі аналізу також було враховано наявність деяких малих категорій, які можуть впливати на результати дослідження. З метою оптимізації та спрощення аналізу було прийнято рішення об'єднати ці категорії в одну загальну категорію. Цей підхід дозволив більш чітко розглянути взаємозв'язок між

типом власності житла та статусом виплати кредиту, враховуючи основні тенденції та зменшивши кількість категорій для подальшого аналізу.

На наступному етапі нашого аналізу звернемо увагу на деякі економічні показники, які можуть мати велике значення для оцінки кредитної ситуації позичальників. Ці показники включають в себе наступне.

1. **Debt-to-Income Ratio (DTI)** - це важливий показник, який визначає, скільки відсотків доходу позичальника витрачається на виплату зобов'язань, включаючи кредити та позики. Високий DTI може бути ознакою фінансового стресу та вплинути на здатність позичальника вчасно виплачувати кредити.

2. **Кількість відкритих кредитних ліній (Open Accounts)** та **загальна кількість кредитних ліній (Total Accounts)** вказують на кількість активних кредитів та кредитних зобов'язань позичальника. Ці показники можуть свідчити про рівень фінансової активності та здатність позичальника керувати багатьма кредитами одночасно.

3. **Revolving Balance** та **Revolving Utilization** вказують на баланс та використання кредитних ліній зі зворотнім оборотом, таких як кредитні картки. Великий баланс або висока використаність можуть сигналізувати про високий рівень кредитного навантаження.

Аналіз цих економічних показників допоможе краще зрозуміти фінансову стабільність та кредитну здатність позичальників, а також виявити можливі кореляції та взаємозв'язки між ними. Відомо, що ці показники можуть мати важливе значення для оцінки ризику неповернення кредитів та забезпечення фінансової стабільності.

Аналіз DTI (див. рисунок 3.8, 3.9) дозволить визначити, чи позичальники з високим DTI мають тенденцію до прострочки платежів та як це впливає на їхню кредитну історію. Дослідження кількості відкритих та загальних кредитних ліній (див. рисунок 3.10) допоможе розірвати їхню взаємозв'язок та визначити, як кількість активних кредитів впливає на фінансовий стан позичальників.

Звернення уваги на Revolving Balance (див. рисунок 3.11) та Revolving Utilization (див. рисунок 3.12) допоможе з'ясувати, як велика кількість заборгованості на кредитних картах та їхня активність використання можуть впливати на спроможність вчасно повертати кредити та забезпечувати стабільність фінансів.

Цей аналіз допоможе зрозуміти важливі залежності та визначити фактори, які можуть бути важливими для прийняття рішень щодо надання кредитів та оцінки кредитного ризику.

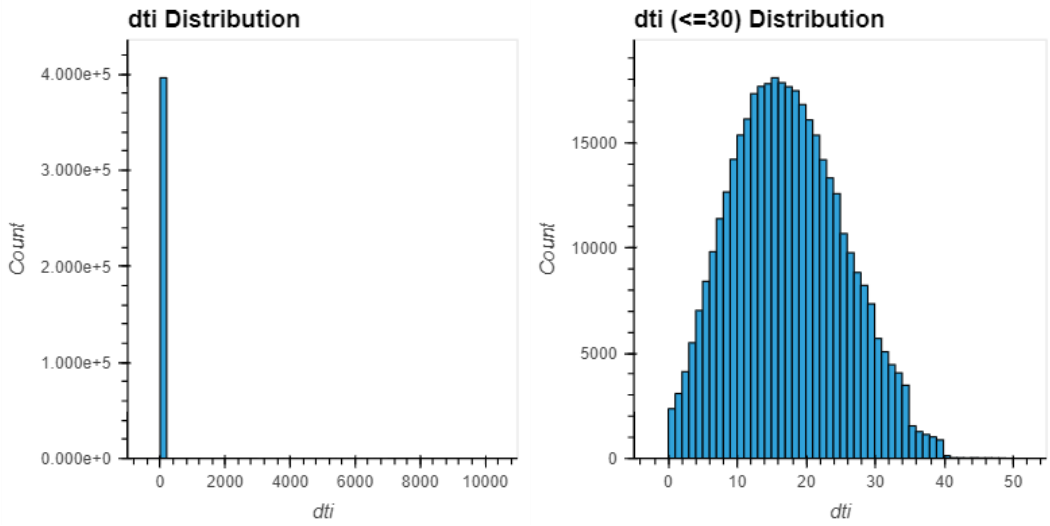

Рисунок. 3.8 – Гістограма розподілу **dti**

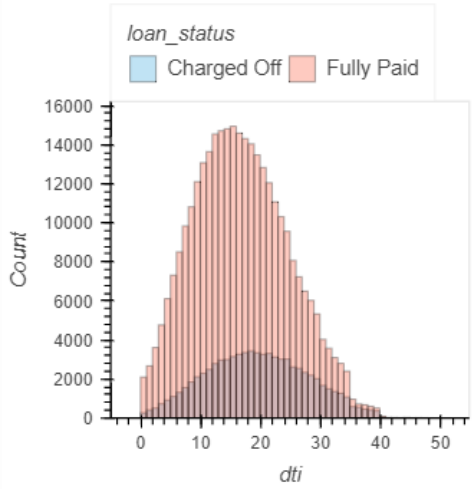

Рисунок. 3.9 – Гістограма розподілу **dti** відповідно кредитного статусу

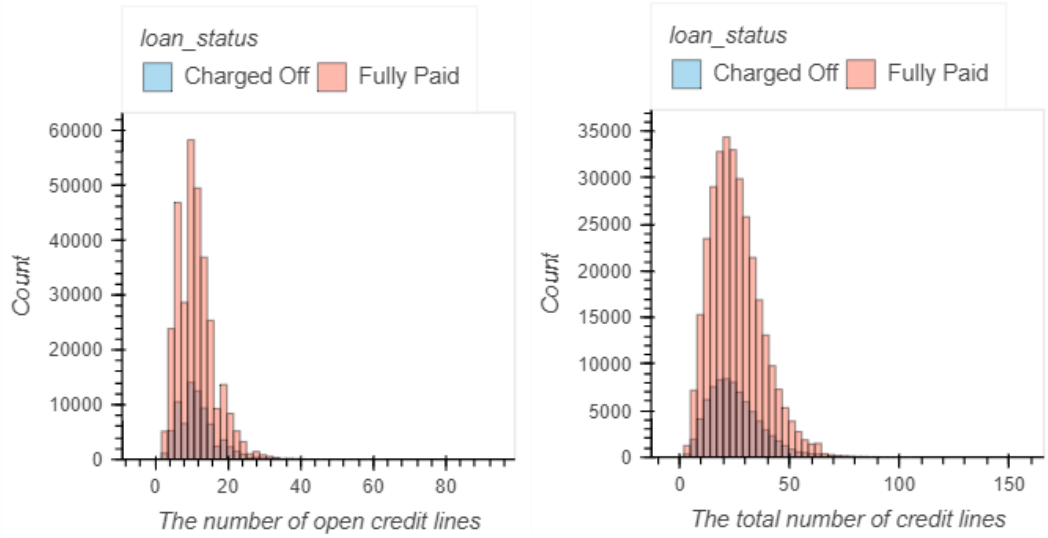

Рисунок. 3.10 – Гістограми розподілу **open_acc** та **total_acc**

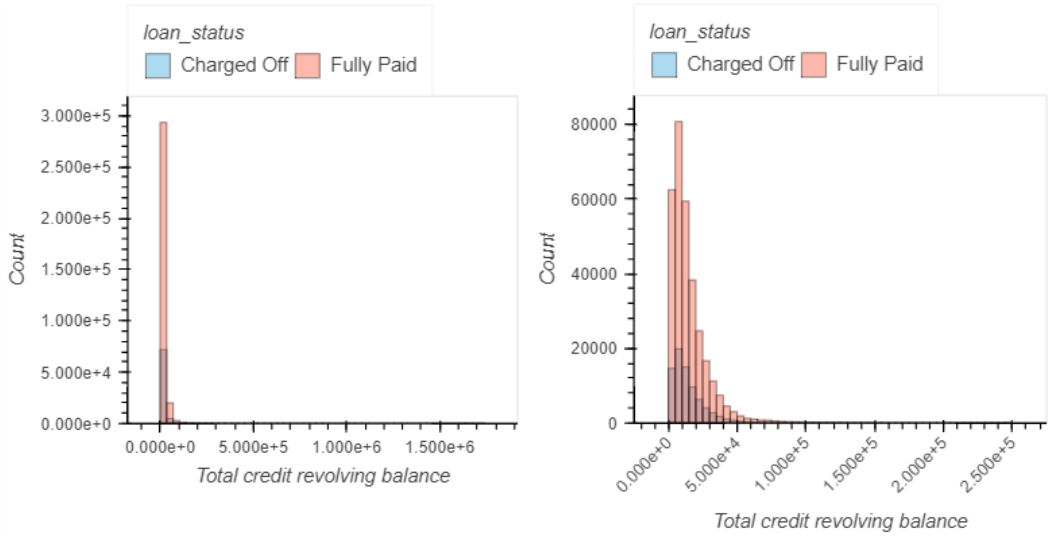

Рисунок. 3.11 – Гістограми розподілу **revol_bal**

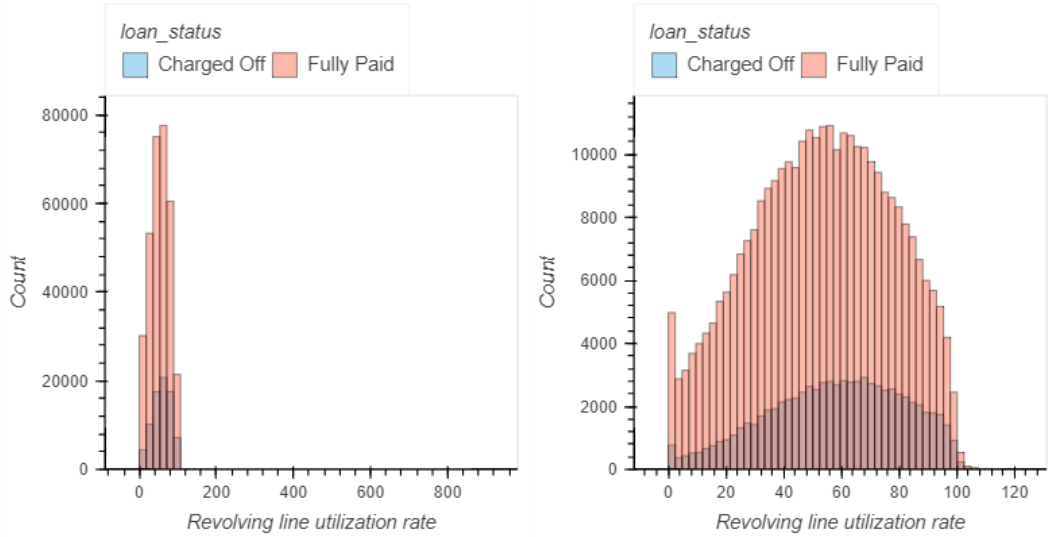

Рисунок. 3.12 – Гістограми розподілу **revol_util**

На основі проведених досліджень були виявлені деякі значущі викиди в даних, що можуть мати важливий вплив на аналіз кредитної ситуації позичальників. Серед найбільш помітних викидів можна виділити наступні.

1. Лише 296 позичальників мають показник Debt-to-Income Ratio (DTI) більше 40. Високий DTI свідчить про велику частку доходу, яка витрачається на погашення зобов'язань, і це може бути ознакою фінансових труднощів.

2. Лише 217 позичальників мають більше 40 відкритих кредитних ліній. Велика кількість відкритих кредитних ліній може вказувати на високий рівень фінансової активності та ризик надмірного заборгованості.

3. Лише 266 позичальників мають більше 80 кредитних ліній у своїй кредитній справі. Це свідчить про значну різноманітність та кількість кредитних зобов'язань у деяких кредитних справах.

4. Лише 397 позичальників мають поновлюваний баланс більше 250000. Великий поновлюваний баланс може бути ознакою великої кількості кредитних зобов'язань або високої кредитної активності.

5. Лише 27 позичальників мають коефіцієнт використання поновлюваної лінії більше 120. Високий коефіцієнт використання поновлюваної лінії може бути ознакою надмірного використання кредитних ресурсів та ризику неповернення кредиту.

Ці викиди в даних важливо враховувати при аналізі кредитного ризику та прийнятті рішень щодо надання кредитів. Вони можуть свідчити про потенційні фінансові труднощі та ризики, які можуть виникнути у відношенні до деяких позичальників.

На основі аналізу даних було прийнято рішення щодо перетворення деяких змінних в бінарну форму, що може спростити аналіз та моделювання. Однією з таких змінних є **pub_rec_bankruptcies**, яка вказує на кількість записів про банкрутство в кредитній історії позичальника.

Аналіз показав, що кількість записів з значенням **pub_rec_bankruptcies** більше 1 не є інформативною для наших досліджень через низку кількість записів. Тому було прийнято рішення перетворити цю змінну на бінарну

форму, де значення 1 вказуватиме на наявність записів про банкрутство в кредитній історії позичальника, а значення 0 - на відсутність таких записів.

Аналогічний підхід застосований до інших змінних, таких як **pub_rec** (кількість публічних записів про дефолт позичальника) та **mort_acc** (кількість іпотечних кредитів). Перетворення цих змінних в бінарну форму дозволить зберегти важливу інформацію про наявність або відсутність певних подій (наприклад, банкрутства) та спростить аналіз та моделювання даних.

Це рішення допоможе покращити якість та точність наших аналітичних моделей та приймати більш обґрунтовані рішення у сфері кредитування.

Під час обробки та підготовки даних для аналізу та моделювання було прийнято рішення видалити деякі ознаки, які загалом не мають великого впливу на ймовірність повернення кредиту. Це дозволяє спростити модель та зменшити кількість зайвих змінних. Вкажемо деякі з ознак, які були видалені.

1. Ознака **address** – ця ознака не виявила значущого впливу на ймовірність повернення кредиту позичальником. Інформація про адресу може бути менш важливою для оцінки кредитної спроможності, тому було прийнято рішення видалити її.

2. Ознака **earliest_cr_line** – дата відкриття першого кредитного рахунку також не виявила значущого впливу на нашу модель. Ця інформація, хоч і важлива для кредиторів, не допомогла покращити передбачення повернення кредиту в нашому конкретному аналізі.

3. Ознака **issue_d** – ця ознака вказує на дату видачі кредиту. Проте вона вважається витоком даних, оскільки при роботі з реальними даними нам не завжди буде відомо, коли буде видана позика чи взагалі видана за використанням нашої моделі. Таким чином, у нас не буде дати видачі для реальних даних, і ця ознака не має практичного значення для нашого аналізу.

4. Ознака **initial_list_status** – також була видалена, оскільки дослідження показало, що джерело кредитування не впливає на поведінку позичальника та не сприяє покращенню передбачення повернення кредиту.

Видалення цих ознак дозволило зробити модель більш простою та ефективною, концентруючись на ключових факторах, які впливають на рішення про надання кредиту та оцінку кредитного ризику.

3.3 Побудова моделей прогнозування ризиків кредитного портфелю

Оцінка якості моделей буде виконуватися за допомогою валідації на 20% вибірки. Для більш якісної класифікації введемо також балансування класів цільової змінної методом SMOTE. Основною метрикою для оцінки якості моделі буде F1 для позитивного класу, проте також для кожної моделі буде побудована ROC-крива, та обраховані основні метрики якості класифікації.

3.3.1 Побудова моделей прогнозування ймовірності дефолту

Першим кроком аналізу була спроба використання неглибокої моделі – логістичної регресії. Логістична регресія – це алгоритм машинного навчання, який використовується для бінарної класифікації та оцінки ймовірності настання певної події. Вона є добрим початковим пунктом для аналізу даних і може надати важливу інформацію щодо впливу окремих факторів на нашу цільову змінну.

Результати роботи моделі логістичної регресії представлені в таблиці 3.3, а також зображені на графіках на рисунках 3.13 і 3.14. Ці результати допоможуть нам оцінити, наскільки ефективно логістична регресія працює та чи варто розглядати більш складні моделі для подальшого аналізу.

Таблиця 3.3 – Результати логістичної регресії

Target	Precision	Recall	F1	Accuracy
Non-Default	0.66	0.64	0.65	
Default	0.65	0.68	0.66	
avg				0.66

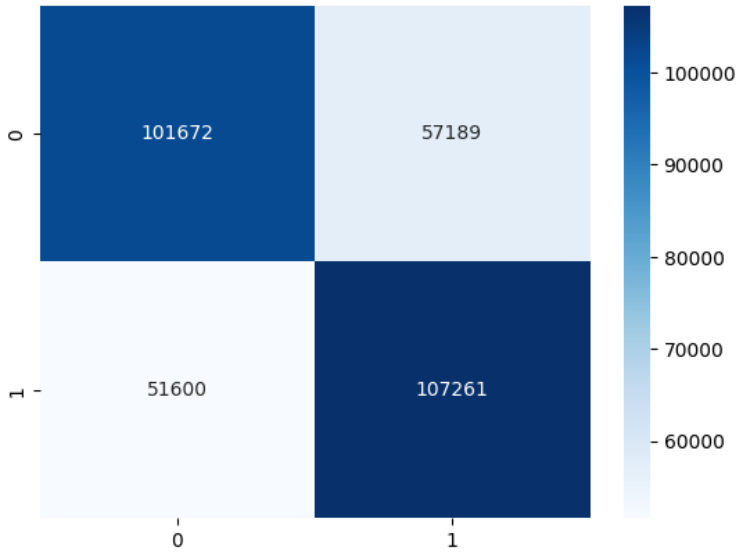

Рисунок 3.13 – Матриця спряженості. Логістична регресія

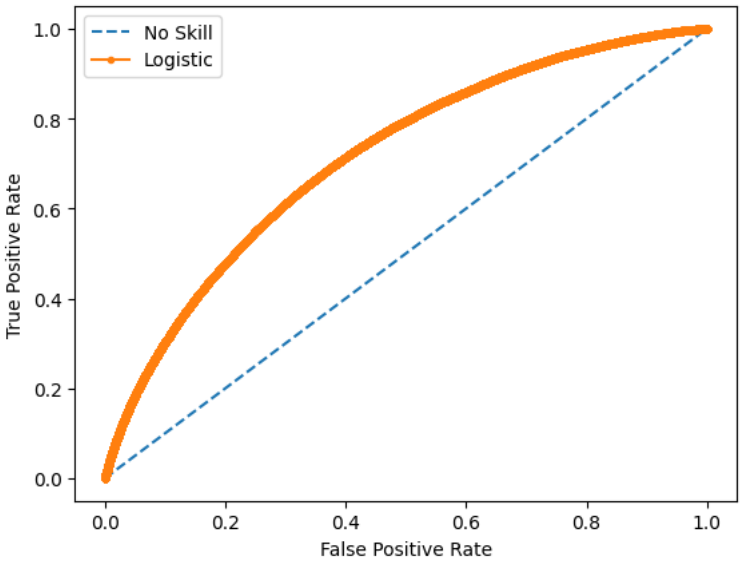

Рисунок 3.14 – ROC-крива. Логістична регресія

Для розширення спектра моделей та збільшення точності нашого аналізу вирішено застосувати ймовірнісний класифікатор, який використовує теорему

Баєса для визначення ймовірності приналежності спостереження до одного з класів при припущенні (наївному) незалежності змінних. Цей класифікатор, відомий як наївний баєсівський класифікатор, часто використовується для задач класифікації.

Результати роботи наївного баєсівського класифікатора будуть представлені в таблиці 3.4, а також проілюстровані на графіках на рисунках 3.15 і 3.16. Ці результати нададуть можливість оцінити, наскільки цей метод покращує точність та ефективність нашої моделі порівняно з іншими підходами.

Таблиця 3.4. Результати наївного баєсівського класифікатора

Target	**Precision**	**Recall**	**F1**	**Accuracy**
Non-Default	0.62	0.71	0.66	
Default	0.66	0.57	0.61	
avg				0.64

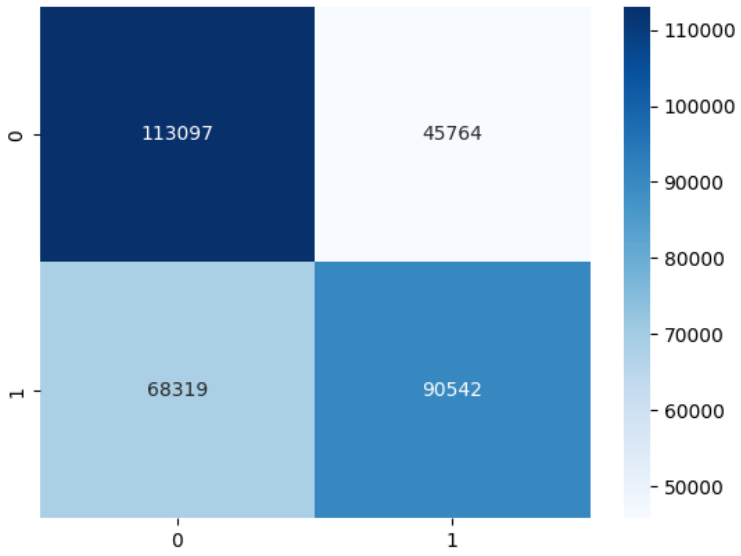

Рисунок 3.15 – Матриця спряженості. Наївний баєсівський класифікатора

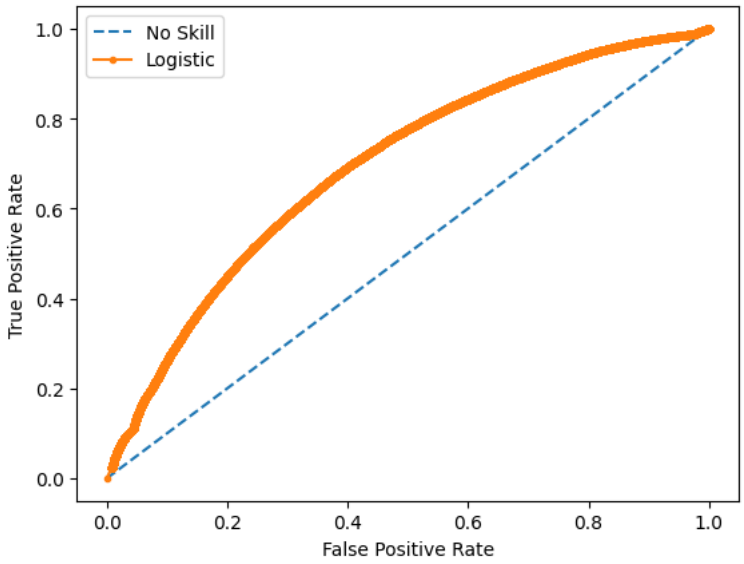

Рисунок 3.16 – ROC-крива. Наївний баєсівський класифікатора

Після застосування логістичної регресії та наївного баєсівського класифікатора переходимо до іншого потужного інструменту – дерева прийняття рішень. Дерева прийняття рішень широко використовуються в інтелектуальному аналізі даних і мають за мету побудувати модель, яка може прогнозувати значення цільової змінної на основі великої кількості вхідних змінних. Цей метод дозволяє розглянути складні взаємозв'язки між змінними та покращити точність прогнозів.

Результати роботи моделі дерев прийняття рішень будуть представлені в таблиці 3.5, а також відображені на візуалізаціях на рисунках 3.17 і 3.18. Ці результати дають можливість оцінити, наскільки дерева прийняття рішень ефективно працюють з нашими даними та чи є вони конкурентоспроможними в порівнянні з іншими моделями для подальшого аналізу.

Таблиця 3.5 – Результати дерева прийняття рішень

Target	Precision	Recall	F1	Accuracy
Non-Default	0.99	0.98	0.98	
Default	0.98	0.99	0.98	
avg				0.98

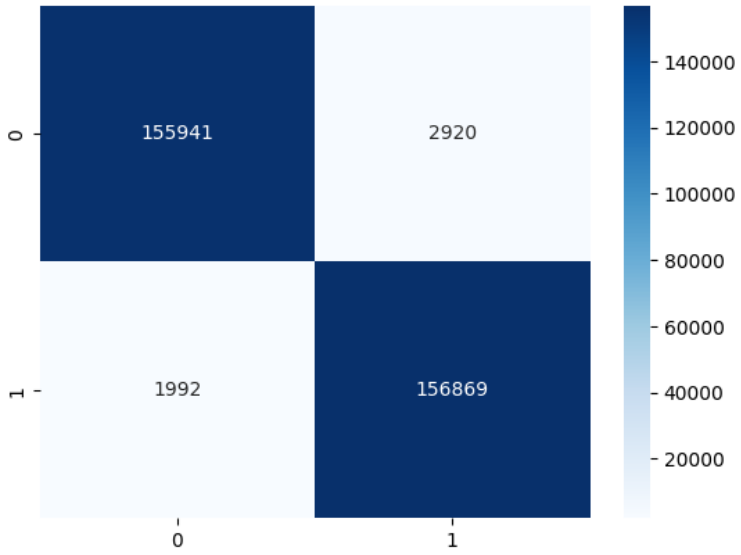

Рисунок 3.17 – Матриця спряженості. Дерево прийняття рішень

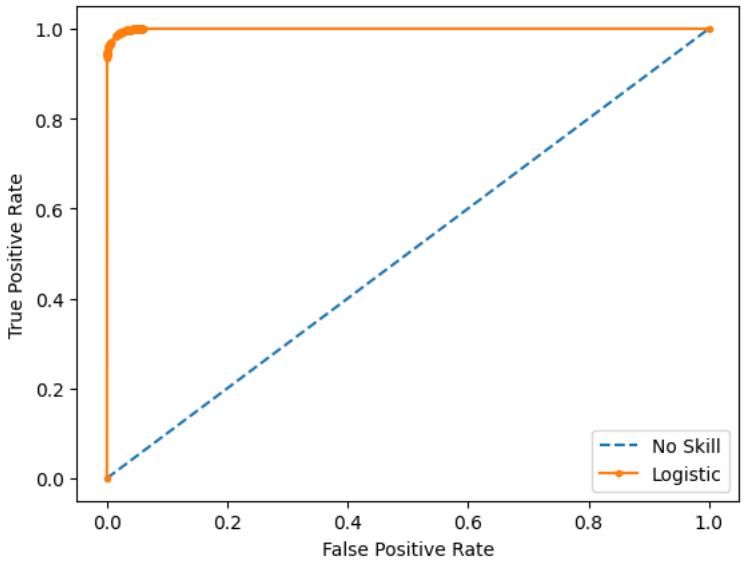

Рисунок 3.18 – ROC-крива. Дерево прийняття рішень

Для подальшого аналізу та покращення ефективності моделей застосуємо модель беггінгу над випадковими деревами – Random Forest. Цей алгоритм є

потужним і досить універсальним інструментом для класифікації та регресії. Мінімізація функції втрат в Random Forest відбувається за допомогою критерію Джині, який допомагає покращити якість розділення вузлів дерева та побудувати більш точну модель.

Результати роботи моделі Random Forest будуть наведені в таблиці 3.6, а також проілюстровані на графіках на рисунках 3.19 і 3.20. Ці результати допоможуть оцінити, наскільки Random Forest покращує точність та ефективність нашої моделі порівняно зі деревами прийняття рішень.

Таблиця 3.6 – Результати Random Forest

Target	**Precision**	**Recall**	**F1**	**Accuracy**
Non-Default	0.99	0.95	0.97	
Default	0.95	0.99	0.97	
avg				0.97

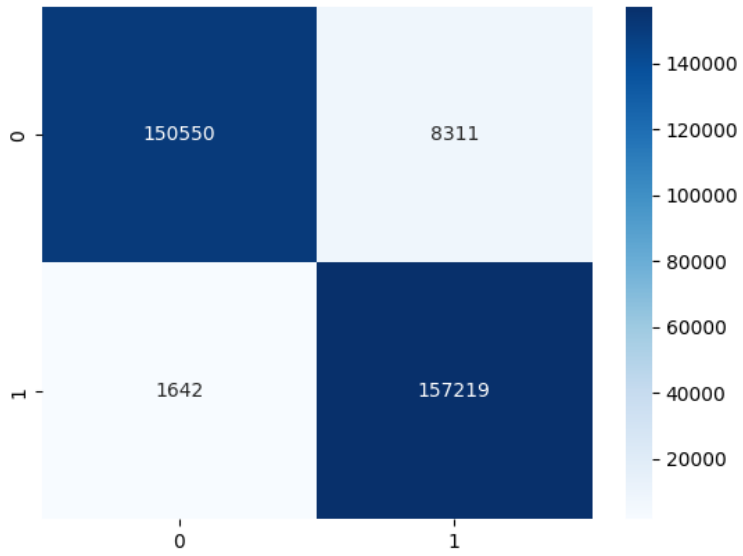

Рисунок 3.19 – Матриця спряженості. Random Forest

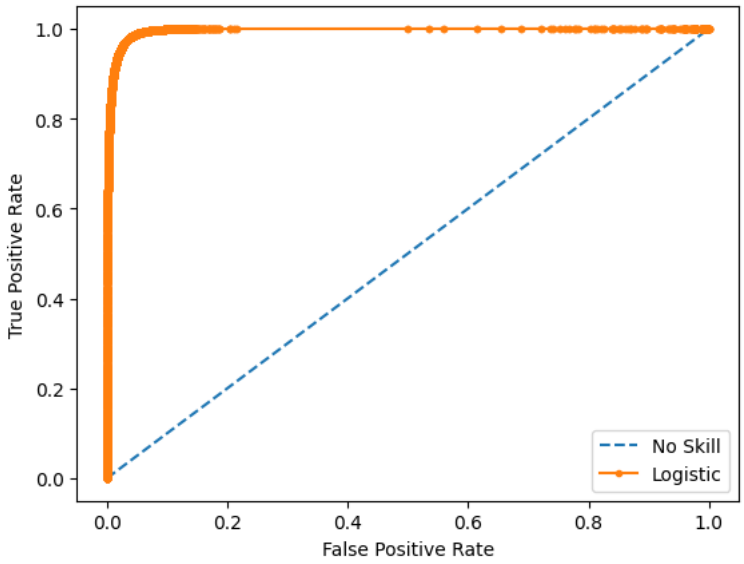

Рисунок 3.20 – ROC-крива. Random Forest

Для оцінки підходу бустингу вирішено використати модель бустингу - XGBoost. XGBoost є однією з найпопулярніших та найефективніших реалізацій алгоритму градієнтного бустингу на деревах. Цей алгоритм відзначається своєю високою точністю та здатністю підвищувати якість прогнозів шляхом поетапного покращення моделі.

Основна ідея XGBoost полягає в використанні простого правила: найкращий алгоритм – той, що максимально зменшує помилку, отриману на попередніх ітераціях. Це досягається шляхом мінімізації вектора антиградієнту функції втрат, в нашому випадку – середньоквадратичної помилки. XGBoost дозволяє покращити точність моделі та досягти більш надійних результатів завдяки своїй ефективності та здатності до роботи з великими обсягами даних.

Результати роботи моделі XGBoost наведено в таблиці 3.7, яка містить важливі метрики для оцінки ефективності моделі. Також, результати ілюстровані на графіках на рисунках 3.21 і 3.22, що дозволяє візуально порівняти ефективність XGBoost з іншими моделями та визначити, наскільки цей метод може покращити точність та якість прогнозів у конкретному завданні.

Таблиця 3.7 – Результати XGBoost

Target	Precision	Recall	F1	Accuracy
Non-Default	0.82	0.97	0.89	
Default	0.96	0.79	0.87	
avg				0.88

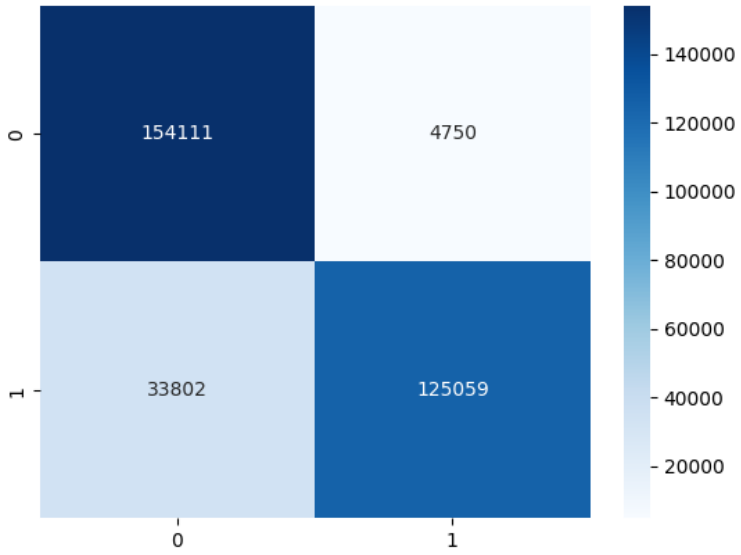

Рисунок 3.21 – Матриця спряженості. XGBoost

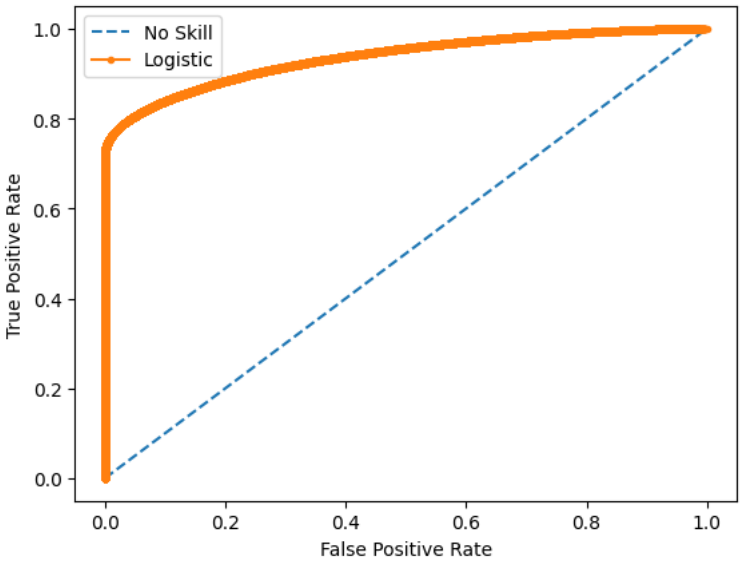

Рисунок 3.22 – ROC-крива. XGBoost

Перейдемо до порівняння результатів якості моделей на основі тренувальних та валідаційних вибірок. Це етап важливий для оцінки того,

наскільки моделі загалом здатні узагальнювати інформацію та робити правильні прогнози на нових даних, які не використовувалися під час тренування. Порівняння результатів на тренувальній та валідаційній вибірках допоможе нам визначити, чи виникає проблема перенавчання та оцінити якість прогнозів моделей на нових даних.

На цьому етапі оцінимо метрики якості для кожної моделі на тренувальних та валідаційних даних (результати наведено в таблиці 3.8). Це дозволить визначити, яка з моделей найкраще впоралася з завданням класифікації та чи є необхідність у подальшій оптимізації та покращенні моделі.

Таблиця 3.8 – Порівняння результатів роботи моделей

Модель	**Training Accuracy**	**Validation Accuracy**
Logistic Regression	0.66	0.65
Gaussian Naive Bayes	0.64	0.68
Decision Tree	0.98	0.69
Random Forest	0.97	0.75
XGBoost	0.88	0.80

Найкращим варіантом виявилася модель бустингу, оскільки високі результати Decision Tree та Random Forest на тренувальній вибірці було отримано через перенавчання цих моделей, а зменшення кількості тренувальних записів не призвело до покращення результатів на валідаційній вибірці. Для подальшого покращення результатів роботи моделі рекомендується розглянути декілька варіантів.

1. **Експертна зміна критичної межі ймовірності дефолту.** Забезпечити можливість експертної корекції порогу ймовірності, який визначає віднесення кредитів до класу дефолту. Аналіз та оптимізація цього параметра можуть допомогти досягти більш точних прогнозів.

2. **Донавчання моделі на свіжих даних.** Періодично перенавчати модель на оновлених даних, що краще відображають поточну економічну ситуацію. Це може сприяти підтримці актуальності та точності моделі в мінливому середовищі.

3. **Застосування ансамблю різних моделей.** Використання ансамблю декількох моделей може покращити загальну ефективність системи. Поєднання бустингу з іншими алгоритмами класифікації дозволяє отримати більш робастну та стійку до різноманітних умов моделі.

Ці кроки можуть допомогти оптимізувати та покращити результати моделі бустингу, забезпечуючи більш точні та стійкі прогнози в контексті кредитного скорингу.

3.3.2 Побудова моделей розрахунку очікуваного періоду настання дефолту

Для проведення оцінки очікуваного періоду настання дефолту вирішено скористатися оцінкою Каплана-Майєра, яка є важливим інструментом в аналізі виживання та оцінці ризику. Спершу проведено аналіз та оцінку часу виживаності в кредиту середньостатистичного позичальника (див. рисунок 3.23). Це дозволить розуміти, як довго в середньому позичальники виконують умови перед настанням дефолту. Для більшої точності та розгалуженості наших оцінок, також проведемо дослідження в межах груп обумовлених оцінок кредитного рейтингу. Це допоможе з'ясувати, чи існують різниці у часі виживаності кредиту між різними групами позичальників з різними кредитними рейтингами.

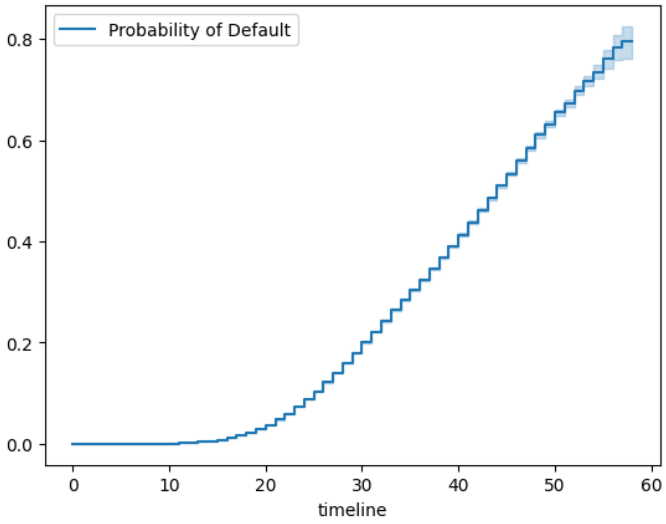

Рисунок 3.23 – Ймовірність дефолту в період

Загальна статистика надає важливі уявлення про те, коли можуть виникати проблеми з виплатами кредитів. Зокрема, виявляється, що проблеми виникають вже невдовзі після отримання кредиту. Це може свідчити про можливе втручання шахраїв, оскільки їхня кількість не є незначною та виглядає дещо аномальною. Однак основний період, коли починаються проблеми з виплатою, припадає на 12-24 місяців після отримання кредиту. Можливо, ця динаміка пов'язана зі змінами в економічній ситуації в країні.

Після цього періоду ризик дефолту зростає, і це може бути викликано різними факторами, такими як зміни в економіці, які не можуть бути передбачені позичальником на довготривалу перспективу. Важливо враховувати ці динаміки під час управління ризиками та прийняття рішень щодо надання кредитів, оскільки вони можуть впливати на результативність портфеля кредитів.

Перейдемо до аналізу виживаності в межах групи за кредитним рейтингом, результати наведено на рисунку 3.24.

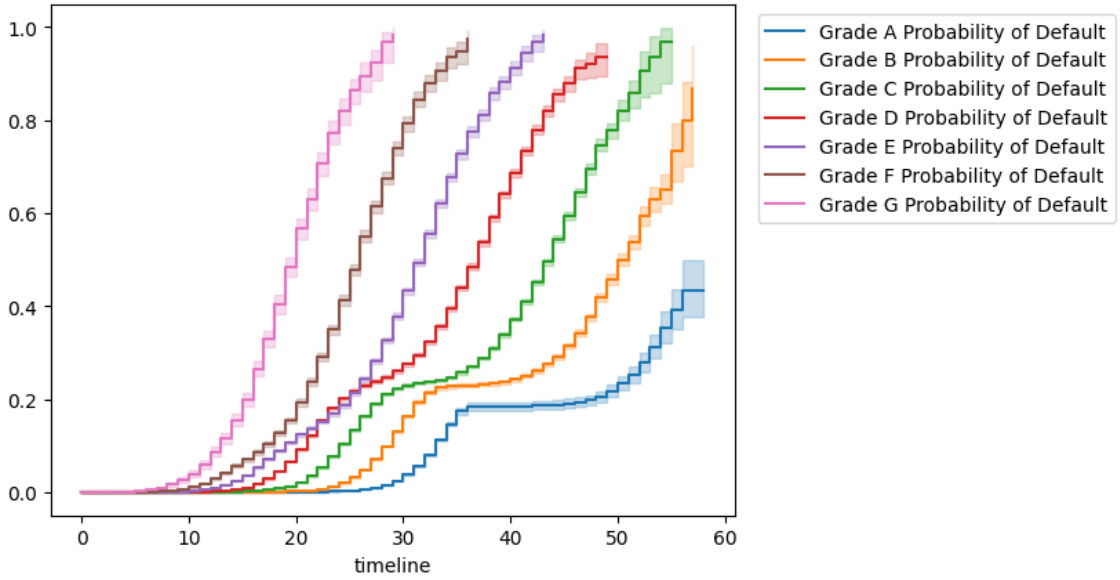

Рисунок 3.24 – Ймовірність дефолту в період відповідно рейтингу

Проводячи аналіз виживаності виплат кредитів в залежності від групи кредитного рейтингу, можемо отримати цінні висновки. Це підтверджує, що позичальники із вищим кредитним рейтингом представляють менший ризик для кредиторів. Наприклад, клієнти класу A зазвичай зіштовхуються з проблемами погашення кредиту ближче до 30 або 50 місяця, що означає, що це відбувається під кінець строку виплат. Низька ймовірність цього може вказувати на те, що ці зміни пов'язані зі структурними змінами в економіці, а не з небажанням позичальника виконувати свої зобов'язання.

З іншого боку, позичальники з рейтингом нижче класу E зазвичай припиняють погашати кредит вже протягом 1-2 років. Таким клієнтам, ймовірно, слід надавати більш жорсткі умови кредитування, наприклад, вищу процентну ставку, обмежену суму кредиту і так далі, для страхування своїх ризиків.

3.3.3 Зведена модель розрахунку кредитних ризиків

Для оцінки економічного ризику кредитного портфелю перейдемо до зведеної моделі розрахунку. У загальному випадку, функція втрат може бути визначена за такою формулою:

$$\mathcal{L} = S * \left(1 + InterestRate\frac{T}{12}\right) * \left(1 - \frac{E(\tau)}{T}\right),$$

$$E\mathcal{L} = \sum \mathcal{L}P\{default\},$$

S – розмір виданої позики

$InterestRate$ – відсоткова ставка

$P\{default\}$ – ймовірність настання дефолту

$E(\tau)$ – очікуваний час настання дефолту (в місяцях)

T – період кредитування (в місяцях)

Серед невідомих змінних в формулі залишаються ймовірність настання дефолту та очікуваний час до настання дефолту (в місяцях). Для отримання цих даних використано моделі, натреновані на наших даних. Зокрема, модель XGBoost допоможе визначити ймовірність дефолту для кожного кредиту. Після цього, застосовуючи оцінку Каплана-Майєра, отримано оціночний період виплат відповідно до кредитного рейтингу та терміну видачі кредиту. Ці дані будуть ключовими для моделі розрахунку економічного ризику кредитного портфелю.

На основі наведених розрахунків, ми можемо розглядати три можливі сценарії:

- **Критичний сценарій:**

$$E\mathcal{L}_c > S$$

В цьому сценарії витрати на позики переважають дохід, тобто за таких обставин, витрати переважатимуть тіло початкового кредиту, що може призвести до фінансових труднощів і необхідності у пошуку додаткових

ресурсів для вирішення цієї ситуації. Для подолання цього сценарію слід розглядати такі кроки, як відмова в виданні нових позик або перегляд умов видачі позик (збільшення відсоткової ставки, зменшення розміру позики тощо).

- **Песимістичний сценарій:**

$$E\mathcal{L}_p > S * InterestRate\frac{T}{12}$$

У цьому сценарії передбачається, що витрати переважатимуть над чистим доходом, тобто тіло кредиту буде погашено, однак через неповний чистий дохід від процентів, не вдасться повністю компенсувати витрати. Це може призвести до невідповідності між витратами і доходом, що потребує уважного фінансового планування та можливо коригування стратегії для забезпечення стійкості фінансового стану. Для подолання цього сценарію також може бути необхідним розглядати перегляд умов видачі позик (збільшення відсоткової ставки, зменшення розміру позики тощо).

- **Оптимістичний сценарій:**

$$E\mathcal{L}_o = 0$$

В цьому сценарії загальні витрати на позики дорівнюють нулю, що означає, що дохід переважає всі можливі витрати. В такому випадку не потрібно вживати додаткових стратегій, і можна продовжувати роботу в звичайному режимі.

Залежно від сценарію, кредитори можуть приймати різні рішення та розглядати різні стратегії для зменшення економічного ризику в управлінні кредитним портфелем.

Отже, очікуванні витрати:

$$E\{expected\ loss\} = E\mathcal{L} = \sum \mathcal{L}P\{default\}$$

$$E\{unexpected\ loss\} = 3\sqrt{E(E\mathcal{L}^2) + E^2(E\mathcal{L})}$$

На прикладі, деякої вибірки розрахуємо очікувані та неочікувані витрати, та порівняємо з реальними даними, результати наведено в таблицях 3.9-3.12.

Таблиця 3.9 – Тестування на повній вибірці

Portfolio size	$5,581,306,925.00
Expected loss	$930,238,195.00
Unexpected loss	$1,860,476,390.01
Real loss	$823,418,695.29

Таблиця 3.10 – Тестування на вибірці розміром N=25000

Portfolio size	$352,665,400.00
Expected loss	$58,667,169.11
Unexpected loss	$117,334,338.22
Real loss	$51,916,256.83

Таблиця 3.11 – Тестування на вибірці розміром N=50000

Portfolio size	$704,084,850.00
Expected loss	$117,426,534.61
Unexpected loss	$234,853,069.21
Real loss	$103,634,970.60

Таблиця 3.12 – Тестування на вибірці розміром N=250000

Portfolio size	$3,532,607,925.00
Expected loss	$589,538,313.79
Unexpected loss	$1,179,076,627.58
Real loss	$521,710,847.61

Висновки до розділу 3

За результатами проведеного аналізу можна зробити висновок, що розроблена модель досить точно оцінює фінансові ризики кредитного портфелю. Похибка цієї моделі в середньому становить близько +10%. Такий рівень похибки характеризує розроблену модель як досить строгу, оскільки вона здійснює оцінку ризиків з певним запасом. Іншими словами, модель схильна до відсіву деяких позитивних кредитних запитів, що може призвести до відмови в виданні позики, особливо в разі ризикових запитів.

Ця консервативність моделі означає, що ми надаємо перевагу запобіганню надмірного ризику перед прийняттям ризикових рішень. Це важливо для забезпечення стабільності кредитного портфелю та уникнення значних втрат. Незважаючи на деяку втрату можливостей щодо видання позик, такий підхід сприяє зниженню ризику та збереженню фінансової стійкості компанії.

У разі, якщо маємо справу з високо ризиковими запитами, цей підхід є більш прагматичним, оскільки він сприяє зменшенню можливості невиплат та збереженню фінансової надійності кредитора.

РОЗДІЛ 4 РОЗРОБКА СТАРТАП ПРОЕКТУ

Цей розділ спрямований на проведення маркетингового аналізу стартап-проекту з метою визначення його потенціалу для введення на ринок і можливих стратегій впровадження. Фінансові установи без попереднього досвіду в кредитній сфері можуть використовувати передбачену модель для швидкого входу на ринок та для подальшого вдосконалення на основі власних даних. Це означає, що проблема, яку вирішує стартап, є актуальною і може здобути успіх на ринку, ставши основним вибором для цільової аудиторії.

4.1 План розробки стартапу та масштабування його на ринок

Розглянемо план розробки стартапу та його введення на ринок. Першим етапом є проведення маркетингового аналізу, який включає наступне.

1. **Конкурентний аналіз:** дослідження методів вирішення схожих проблем, які вже використовуються на ринку.
2. **Формування ідеї проекту та визначення цільової аудиторії:** розробка унікальної ідеї проекту та визначення, для кого він призначений.
3. **Розробка стратегії введення продукту на ринок:** основана на аналізі ринкового середовища.

Після цього наступним етапом є організація самого стартапу, яка включає в себе наступні пункти.

1. **План розробки та запуску продукту:** складення докладного плану та створення графіка розробки та запуску продукту.
2. **Оцінка обсягу виробництва та ресурсів:** визначення необхідних ресурсів для виконання плану.
3. **Розрахунок витрат:** визначення витрат на реалізацію проекту та запуск.

Далі, потрібно провести фінансово-економічний аналіз та оцінку ризиків стартап-проекту, включаючи наступні показники.

1. **Визначення обсягу інвестиційних втрат:** розрахунок втрат, пов'язаних з інвестиціями.

2. **Розрахунок фінансово-економічних показників:** обчислення таких показників, як собівартість, ціна продукту/послуги, податковий збір та чистий прибуток, а також оцінка інвестиційної привабливості проекту, таких як рентабельність продажів та період окупності проекту.

3. **Визначення основних ризиків проекту та їх управління:** визначення потенційних ризиків та розробка стратегій для їх зменшення.

На заключному етапі, розробляються заходи з комерціалізації продукту, що є ключовим для росту та розширення бізнесу. Це включає наступні кроки.

1. **Дослідження інтересів інвесторів та бізнесів:** вивчення потреб та інтересів потенційних інвесторів та бізнес-партнерів.

2. **Складання інвестиційної пропозиції:** розробка пропозиції, яка включає опис продукту, його розмір та потенційні шляхи розвитку.

3. **Вибір каналів комунікації:** визначення способів спілкування з потенційними інвесторами та зацікавленими сторонами.

Далі в розділі буде представлено результати виконання кожного з наведених етапів.

4.2 Опис ідеї стартап-проекту

Стартап-проект спрямований на вирішення проблеми, пов'язаної із передбаченням кредитних ризиків у банківському секторі. Основною ідеєю цього стартапу є розробка системного підходу до визначення можливих втрат, що виникають внаслідок дефолту по кредитній лінії. В таблиці 4.1 представлена інформаційна характеристика стартап-проекту.

Таблиця 4.1 – Інформаційна карта проекту

Назва проекту	Credit Risk Guard
Коротка анотація	Наш продукт пропонує інноваційний підхід до вирішення проблеми прогнозування кредитних ризиків у банківському секторі, спрямований на розробку системи, яка дозволить банкам систематично визначати можливі втрати внаслідок дефолту по кредитним лініям. Наш підхід заснований на використанні аналітичних методів та машинного навчання, що допоможе банкам удосконалити свої стратегії управління кредитними ризиками та знизити втрати.
Термін реалізації проекту	12 місяців
Необхідні ресурси	Інфраструктура та офісні приміщення Технології та програмне забезпечення Фінансування Доступ до даних, партнерські відносини з банками Захист інтелектуальної власності
Опис проблеми, яку вирішує проект	Головною проблемою, яку вирішує цей стартап, є необхідність банків та фінансових установ визначати ймовірність дефолту позичальників, а також визначати можливі втрати, пов'язані з дефолтом по кредитним лініям.
Головні цілі та завдання проекту	Надати бізнесу ефективний інструмент для зниження кредитних витрат та підвищення фінансової стійкості

Продовження таблиці 4.1

Очікувані результати	Забезпечення можливості розширення бізнесу та залучення нових клієнтів і інвесторів завдяки інноваційному підходу до розрахунку кредитних ризиків. Покращення інфраструктури та обслуговування клієнтів після введення продукту на ринок.

4.3 Технологічний аудит ідеї проекту

В даному розділі розглянуто концепцію стартапу та проведено аналіз конкурентного середовища. У таблиці 4.2 подано опис ідеї стартапу.

Таблиця 4.2 – Опис ідеї проекту

Зміст ідеї	**Напрямки застосування**	**Вигоди для користувача**
Надати банкам та фінансовим установам ефективного інструмента для зменшення кредитних втрат і підвищення їхньої фінансової стійкості.	Система може бути впроваджена в банки для покращення процесів кредитування, а також управління кредитними ризиками та оптимізації кредитних портфелів.	Підвищення фінансової стійкості банку Швидше та надійне кредитування Зменшення витрат на кредитний аналіз

Продовження таблиці 4.2

Зміст ідеї	**Напрямки застосування**	**Вигоди для користувача**
Проект також передбачає інтеграцію цієї системи з існуючими банківськими процесами та системами управління ризиками.	Небанківські фінансові установи, такі як кредитні спілки та фінансові компанії, можуть використовувати цю систему для зниження кредитних ризиків та покращення управління кредитами.	Зменшення ризику невдач Зниження вартості кредитів Покращена якісь обслуговування клієнтів
	Страхові компанії можуть використовувати інноваційний підхід до прогнозування ризиків для визначення премій та страхових тарифів.	Збільшення доступності страхування Можливість отримання індивідуальних страхових рішень Більш точне оцінювання ризику

Наступним кроком проведено порівняльний аналіз конкурентів проекту, результати аналізу представлено у таблиці 4.3. На сьогоднішній день не існує загального рішення, яке підходило б для всіх банків. Зазвичай банки розробляють систему управління ризиками в рамках власних проектів, але іноді звертаються за допомогою до зовнішніх консультантів.

Таблиця 4.3 – Порівняльний аналіз конкурентів проекту

Техніко-економічні характеристики ідеї	**(потенційні) товари/концепції конкурентів**		**W**	**N**	**S**
	Стартап проект «Credit Risk Guard»	**Внутрішня розробка банку**			
Ціна проекту	Низька	Низька		+	
Експертність	Може бути різною	Висока		+	
Доступ та повнота даних	Низька	Висока	+		
Швидкість інтеграції	Висока	Низька			+
Знання продукту	Високе	Середнє			+

Аналіз конкурентоздатності вказує на те, що конкуренти мають одну перевагу: доступність та повноту даних. Банки володіють актуальним набором даних для вдосконалення своїх систем, проте ця проблема вирішується під час укладення договору щодо співпраці з метою досягнення найкращих результатів відповідно до стратегії банку. В той же час, спеціалісти з високим рівнем досвіду мають можливість розробити систему, яка може бути швидко інтегрована в загальну банківську систему управління. При цьому вони здатні зберегти гнучкість системи, дозволяючи легко вносити зміни в майбутньому.

Наступним кроком проведено аналіз реальних можливостей з технічного втілення ідеї проекту. Результати аналізу продемонстровано в таблиці 4.4.

Таблиця 4.4 – Технологічна здійсненність продукту

Ідея проекту	**Технології реалізації**	**Наявність технологій**	**Доступність технологій**
Надати банкам та фінансовим установам ефективного інструмента для зменшення кредитних втрат і підвищення їхньої фінансової стійкості.	Мова Програмування Python	Існуючі бібліотеки, такі як Pandas, Numpy, Sklearn та Matplotlib	Є доступними та безкоштовними для використання
	Мова Програмування R	Різні фреймворки та бібліотеки	Є доступними та безкоштовними для використання
Обрана технологія реалізації проекту: мова програмування Python			

Серед двох обраних мов програмування було прийнято рішення використовувати Python. Ця мова має більш широке застосування в різних галузях програмування, і, отже, її використання спрощує та прискорює процес розробки. Докладна документація мови та її бібліотек також позитивно впливає на продуктивність роботи. Тим самим Python є найкращим вибором для даного проекту.

4.4 Аналіз ринкових можливостей запуску стартап-проекту

У даному розділі проведено аналіз ринкових можливостей запуску стартап-проекту. У таблиці 4.5 подано характеристику потенційного ринку.

Таблиця 4.5 – Характеристика потенційного ринку

Показники стану ринку (найменування)	**Характеристика**
Кількість головних гравців, од	25
Загальний обсяг продаж	727 млрд. грн.
Динаміка ринку	Зростає
Наявність обмежень для входу	Нормативні документи державного рівня на відповідність ПЗ до законів України
Специфічні вимоги до стандартизації та сертифікації	Постанова НБУ №351, Наказ НБУ №141, тощо
Середня норма рентабельності в галузі, %	25%

У таблиці 4.6 подано характеристику потенційних клієнтів.

Таблиця 4.6 – Характеристика потенційних клієнтів

Назва пункту	**Опис**
Потреби, що формує ринок	Управління ризиками
Цільова аудиторія (цільові сегменти ринку)	Комерційні банки та інші фінансові установи, що ведуть кредитну діяльність, страхові компанії
Відмінності у поведінці різних потенційних цільових груп клієнтів	У банків різний підхід до аналізу, висновків та підходу до постановки задач щодо управління ризиками
Вимоги споживачів до товару	Висока швидкість обробки великої кількості інформації; Інструкція щодо використання продукту; Технічна підтримка та супровід продукту

Наступним етапом проведено аналіз факторів загроз і можливостей. Це допоможе оцінити потенційні перешкоди при введенні продукту на ринок, ретельно проаналізувавши загрози. З іншого боку, розрахунок факторів можливостей дозволить нам ідентифікувати всі сприятливі умови та, за можливості, використати їх на користь проекту. Результати аналізу факторів загроз наведено в таблиці 4.7.

Таблиця 4.7 – Фактори загроз

Фактор	**Зміст загрози**	**Можлива реакція**
Високий поріг потреб ринку у сфері стандартизації та ліцензування	Значні витрати пов'язані з процедурою ліцензування, тестуванням та впровадженням систем, а також правовими питаннями	Розробити систему розробки програмного забезпечення, яка дотримується всіх відомих технічних та юридичних вимог та умов.
Зміна потреб користувачів	Клієнтам буде потрібна система з додатковими вимогами та можливостями	Для цього слід передбачити можливість розширення системи та підвищення ступеня модульності.
Зміна економічної ситуації	Зміни в економічній ситуації можуть призвести до невідповідності моделі оцінки новим умовам та реаліям ринку	Для запобігання цій загрозі, слід розглянути можливість динамічного тренування моделі під час роботи.

Результати аналізу факторів можливостей наведено в таблиці 4.8.

Таблиця 4.8 – Фактори можливостей

Фактор	**Зміст загрози**	**Можлива реакція**
Конкуренція	Відсутність прямих аналогів	Пристосування проекту до конкретних вимог та особливостей українського ринку
Зміна регуляційних нормативів	Здатність до гнучкого налаштування відповідно до чинних нормативно-правових актів	Перехід до нової системи управління ризиками та збитками

Також проведено аналіз пропозиції, в ході якого були визначено загальні характеристики конкурентної ситуації. Результати наведено в таблиці 4.9.

Таблиця 4.9 – Ступеневий аналіз конкуренції

Особливості середовища	**В чому виражена характеристика**	**Вплив на діяльність**
Вид конкуренції: **Монополія**	На ринку присутні декілька постачальників конкурентів, але їх товар дещо відрізняється від нашого проекту	Забезпечення якості системи, сталий розвиток, вдосконалення, оновлення та обслуговування.
За рівнем конкурентної боротьби: **міжнародна**	Ринок збуту та конкуренти представлені на міжнародному рівні	Забезпечення відповідності системи виконанню міжнародних стандартів управління ризиками

Продовження таблиці 4.9

За галузевою ознакою: **внутрішньогалузева**	Ця система може застосовуватися для оцінки різноманітних видів ризиків, але обмежена однією конкретною галуззю	Забезпечення якості та гнучкості відповідно до вимог управління ризиками
Конкуренція за видом товару: **товарно-видова**	Конкуренція між різними типами та системами оцінки ризиків	Розробити систему, враховуючи як недоліки, так і переваги конкретних методів
За характером конкурентних переваг: **нецінова**	Методи пропонують різний рівень якості та точності прогнозів	Підвищувати якість прогнозів у розробленій системі

Після проведення аналізу конкуренції важливо виконати докладний аналіз відносних умов конкуренції в галузі, використовуючи модель «П'ять сил» Майкла Портера. Результати аналізу наведено в таблиці 4.10.

Таблиця 4.10 – Ступеневий аналіз конкуренції за моделлю «П’яти сил»

Складові аналізу				
Прямі конкуренти в галузі	**Потенційні конкуренти**	**Постачальники**	**Клієнти**	**Товаро замінник**
SAS	Існуючі методи прогнозування	-	Якість системи та підтримка оновлень	Відомий розробник з наявними обчислювальними потужностями
Висновки	Знижена інтенсивність конкуренції, а також значний попит на подібні рішення.	Можливості для входу на ринок буде легшою через відсутність прямих конкурентів	Вимоги клієнта до якості та простоти застосування продукту	Випустити програмний продукт, який буде відповідати вимог клієнтів та законним нормам

На основі аналізу конкуренції (таблиця 4.9), характеристик ідеї проекту (таблиця 4.2), вимог споживачів до товару (Таблиця 4.5) та факторів маркетингового середовища (таблиці 4.6, 4.7), можна скласти перелік факторів конкурентоспроможності – таблиця 4.11.

Таблиця 4.11 – Фактори конкурентоспроможності

Фактор конкурентоспроможності	**Обґрунтування факторів**
Час впровадження	Швидший час запуску дозволить залучити більше клієнтів
Функціональність	Гнучкість та динамічність системи забезпечить перевагу на ринку
Клієнтська підтримка	Супровід та підтримка готового продукту забезпечить вищу якість та збільшить довіру клієнтів

На основі отриманих даних проведено аналіз сильних та слабких сторін стартап-проекту – таблиця 4.12.

Таблиця 4.12 – Порівняльний аналіз сторін проекту

Фактор конкуренто-спроможності	**Бали (1-20)**	**Рейтинг товарів-конкурентів**						
		-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
Час впровадження	20	+						
Функціональність	15			+				
Клієнтська підтримка	16				+			

Останнім кроком маркетингового дослідження можливостей для реалізації системи підтримки прийняття рішень як стартап-проекту є створення SWOT-аналізу (матриці сильних (Strength) та слабких (Weak) сторін, загроз (Troubles) та можливостей (Opportunities)) на основі конкурентних та маркетингових загроз та можливостей, а також сильних і слабких сторін. Матриця SWOT-аналізу продемонстрована у таблиці 4.13.

Таблиця 4.13 – SWOT-аналіз проекту

Сильні сторони	**Слабкі сторони**
- Швидке впровадження системи - Підтримка та адаптація системи в реальному часі - Функціональність	- Складність впровадження системи через нормативні обмеження - Складність виходу на міжнародний ринок ускладнюється через відмінність в нормативних обмеженнях між різними країнами
Можливості	**Загрози**
- Наявність вакантного ринкового сегменту без прямих конкурентів	- Зміна нормативних актів, що регулюють діяльність - Зміна потреб клієнтів

На основі SWOT-аналізу можемо ідентифікувати альтернативні стратегії ринкової поведінки для впровадження стартап-проекту на ринок та приблизний оптимальний час їхньої реалізації, з урахуванням потенційних проектів конкурентів, які можуть бути запущені на ринок. Альтернативи продемонстровано в таблиці 4.14.

Таблиця 4.14 – Альтернативи ринкового впровадження

Альтернатива ринкової поведінки	**Ймовірність отримання ресурсів**	**Строки реалізації**
Дистрибуція обмеженої версії програмного продукту на безоплатній основі	70%	12 місяців
Вихід на ринок з нижчою якістю прогнозування	40%	8 місяців

Продовження таблиці 4.14

Вихід на ринок через пряме партнерство з установою	80%	10 місяців

4.5 Розроблення ринкової стратегії стартап-проекту

В даному розділі наведено аналіз ринкової стратегії стартап проекту, з врахуванням цільової аудиторії продукту. В таблиці 4.15 наведено аналіз цільової аудиторії запропонованого товару.

Таблиця 4.15 – Цільова аудиторія проекту

Опис профілю цільової групи потенційних клієнтів	**Готовність споживачів сприйняти продукт**	**Орієнтовний попит у межах цільової групи (сегменту)**	**Інтенсивність конкуренції в сегменті**	**Простота входу у сегмент**
Комерційні банки	Висока	Високий	Висока	Середня
Державні банки	Висока	Середній	Низька	Важка
Страхові компанії	Середня	Низький	Висока	Середня

В таблиці 4.16 наведено базову стратегію розвитку.

Таблиця 4.16 – Базова стратегія розвитку

Обрана альтернатива	**Стратегія охоплення ринку**	**Ключові конкурентоспроможні позиції**	**Базова стратегія розвитку**
Дистрибуція обмеженої версії програмного продукту на безоплатній основі	Стратегія сегментації ринку	Цінова політика Функціональність Якість	Стратегія інноваційного розвитку

В таблиці 4.17 наведено базову стратегію конкурентної поведінки.

Таблиця 4.17 – Базова стратегія конкурентної поведінки

Чи є проект «першопрохідцем» на ринку?	Ні
Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Пошук нових та залучення існуючих
Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Ні
Стратегія конкурентної поведінки	Диференціація продукції

В таблиці 4.18 наведено стратегію позиціювання.

Таблиця 4.18 – Стратегія позиціювання

Вимоги до товару цільової аудиторії	Вимоги клієнта до якості та простоти застосування продукту
Базова стратегія розвитку	Стратегія інноваційного розвитку
Ключові конкурентоспроможні позиції власного стартап-проекту	Час впровадження Функціональність Клієнтська підтримка
Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)	Надійність Інновації Експертність

4.6 Розроблення маркетингової програми стартап-проекту

Після проведеного всебічного аналізу, можливо повністю висвітлити основні переваги потенційного продукту та розробити стратегію маркетингових комунікацій. В таблиці 4.19 продемонстровано ключові переваги потенційного товару.

Таблиця 4.19 – Ключові переваги концепції потенційного товару

Потреба	**Вигода, яку пропонує товар**	**Ключові переваги перед конкурентами**
Швидкість впровадження	Споживачі мають можливість оперативно та ефективно інтегрувати та використовувати продукт	Швидкість інтеграції системи є однією з ключових переваг продукту, що гарантує інтерес клієнтів

Продовження таблиці 4.19

Функціональність	Продукт має розширений спектр функцій і можливостей, які перевершують базові очікування клієнта	Користувачі можуть отримувати більше користі від продукту, що сприяє їхній задоволеності та лояльності до бренду
Клієнтська підтримка	Можливості отримати високоякісну та надійну допомогу, консультації та обслуговування з боку розробника після придбання продукту	Клієнтська підтримка є важливою частиною стратегії обслуговування клієнтів та може визначити успіх на ринку

В таблиці 4.20 наведено концепцію маркетингових комунікацій.

Таблиця 4.20 – Концепція маркетингових комунікацій

Специфіка поведінки цільових клієнтів	Банки та інші фінансові установи, які ведуть кредитну діяльність, що виконують контрольно-керуючі функцію у банківській сфері
Канали комунікацій	- Участь у конференціях - Публікації у виданнях та професійних видавництвах - Новини у галузі інформаційних технологій
Ключові позиції	Позиція, заснована на відображенні високої якості та надійності оцінки та прогнозування кредитних ризиків
Завдання рекламного повідомлення	- Розповісти про новий продукт і його переваги - Продемонструвати функціональні переваги власного продукту над конкурентами.

Висновки до розділу 4

Під час аналізу розглядалася розробка стратегій виходу на ринок та маркетингових стратегій. Варто зауважити, що ринок, на якому проект функціонує, є сприятливим, оскільки відсутні прямі конкуренти для даного продукту, а відповідні завдання виконуються внутрішніми підрозділами компанії. Запропонований продукт дозволяє значно зекономити ресурси, які б інакше були витрачені на власний розробки, і швидко досягти результатів. Це відкриває можливості для швидкого становлення ключовим гравцем на ринку.

Крім того, були вивчені сильні та слабкі сторони проекту, проведено SWOT-аналіз, а також проаналізовано конкурентів та цільову аудиторію. На основі цих досліджень було розроблено концепцію маркетингової стратегії для визначених цільових аудиторій.

ВИСНОВКИ

У рамках даної магістерської дисертації було виконано комплексний аналіз та дослідження моделей машинного навчання з метою їх використання у сфері прогнозування та оцінювання кредитного ризику, зокрема, у випадку можливого настання дефолту. В дисертації наявне ретельне вивчення теоретичних аспектів застосування моделей машинного навчання та статистичних методів аналізу виживаності, а також їхніх практичних застосувань у контексті кредитної сфери.

Основна увага була приділена аналізу даних, зібраних з відкритих джерел однорангової кредитної компанії Lending Club. Це надало можливість отримати конкретні результати, підкріплені реальними прикладами та висновками. У ході практичних досліджень було виявлено, що розроблена модель ефективно визначає ризик дефолту та забезпечує надійну оцінку кредитних збитків у разі його виникнення.

Згідно з отриманими результатами, розроблений метод може служити важливим інструментом для фахівців у галузі кредитного аналізу та ризик-менеджменту. Модель може бути використана для прийняття обґрунтованих рішень щодо видачі кредитів, оптимізації процесів ризик-менеджменту та забезпечення більш точного прогнозу фінансових наслідків.

Результати даного дослідження не лише розширюють теоретичні знання про моделі машинного навчання у фінансовій сфері, але й надають практичний внесок у розвиток методів кредитного аналізу. Зокрема, вони сприяють підвищенню ефективності процесів видачі кредитів, а також зменшенню можливих фінансових ризиків для кредитних установ.

Завершуючи дисертацію, можна зазначити, що розроблений метод не тільки поглиблює наше розуміння моделей машинного навчання та статистичних методів аналізу виживаності в фінансовому аналізі, але і

відкриває нові перспективи для подальших досліджень та розвитку методів управління кредитним ризиком в умовах сучасного фінансового ринку.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Про затвердження Положення про визначення банками України розміру кредитного ризику за активними банківськими операціями: Постанова правління Національно банку України №351 від 30.06.2016р. (зі змінами) URL: https://bank.gov.ua/ua/legislation/Resolution_30062016_351

2. Про затвердження Положення про організацію системи управління ризиками в банках України та банківських групах: Постанова правління Національно банку України №64 від 11.06.2018р.: станом на 30.03.2023. URL: https://zakon.rada.gov.ua/laws/show/v0064500-18#Text

3. Впровадження нових вимог до капіталу відповідно до норм Базеля та законодавства ЄС. URL: https://bank.gov.ua/admin_uploads/article/%D0%A7%D0%B0%D1%83%D1%81_pr_2017-07-19.pdf?v=4

4. Naumenkova S. Basel I, II, III: The development of approaches for strengthening of prudential framework. Bulletin of Taras Shevchenko National University of Kyiv, Economics. 2015. No. 177. URL: https://doi.org/10.17721/1728-2667.2015/177-12/5

5. Міжнародні стандарти фінансової звітності (МСФЗ, МСФЗ для МСП, включаючи МСБО та тлумачення КТМФЗ, ПКТ). URL: https://zakon.rada.gov.ua/laws/show/929_010#Text

6. Річний звіт 2022. URL: https://bank.gov.ua/admin_uploads/article/annual_report_2022.pdf?v=4

7. Кузнєцова Н. В., Бідюк П. І. Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування. Наукові вісті НТУУ «КПІ», 2010. № 1(69). С. 42–53.

8. Бідюк П. І., Кузнєцова Н. В., Терентьєв О. М. Система підтримки прийняття рішень для аналізу фінансових даних. Наукові вісті НТУУ «КПІ», 2011. №1. С. 48-61.

9. Кузнєцова Н. В., Бідюк П. І. Системний підхід до аналізу кредитних ризиків з використанням мереж Байєса. Наукові вісті НТУУ «КПІ», 2008. № 3(59). С. 11-24.

10. Бідюк П. І., Гуськова В. Г. Аналіз кредитоспроможності за допомогою методів інтелектуального аналізу даних. Електронне моделювання, 2019. Т. 41, № 2. С. 111-120.

11. Гуськова В. Г. Методи і моделі інтелектуального аналізу даних для оцінювання фінансових ризиків: дис. … д-ра філософії: 122 – Комп'ютерні науки / КПІ ім. Ігоря Сікорського, Київ, 2020. 197 с.

12. Shen F., Wang R., Shen Y. A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. Technological and Economic Development of Economy, 2019. 26(2). P. 405-429. doi:10.3846/tede.2019.11337.

13. Cao R., Vilar J., Devia R., Andres E. Modelling consumer credit risk via survival analysis. Sort: Statistics and Operations Research Transactions, ISSN 1696-2281, Vol. 33, No. 1, 2009, P. 3-30.

14. Lane R., Looney S., Wansley J. An application of the cox proportional hazards model to bank failure, Journal of Banking & Finance, Vol. 10, Issue 4, 1986, P. 511-531.

15. Dirick L., Claeskens G., Baesens B. Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society, 2017, 68(6), P. 652–665. doi:10.1057/s41274-016-0128-9

16. Thomas L., Reyes E. M. Tutorial: survival estimation for Cox regression models with time-varying coefficients using SAS and R, J. Stat. Softw., 2014, Vol. 61, P. 1-23.

17. Сумін О. О., Шубенкова І. А. Системний підхід до аналізу кредитних ризиків в банківському секторі. II Всеукраїнська науково-практична конференція «Системні науки та інформатика», м. Київ, 04-08 грудня 2023 року. С. 213-217.

ДОДАТОК А ЛІСТИНГ ПРОГРАМИ

```
# Installation of required modules
!pip install -q hvplot
!pip install -q lifelines
!pip install -q scikit-survival


# Importing required modules
import pandas as pd
import numpy as np

from scipy import stats

import matplotlib.pyplot as plt
import seaborn as sns
import hvplot.pandas

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

from imblearn.over_sampling import SMOTE

from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

import xgboost as xgb

from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_auc_score
```

```
from sklearn.metrics import roc_curve

from lifelines import KaplanMeierFitter


# Configuring environment settings
pd.set_option('display.float', '{:.2f}'.format)
pd.set_option('display.max_columns', 50)
pd.set_option('display.max_rows', 50)


# Downloading data
import os
for dirname, _, filenames in os.walk('/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
data = pd.read_csv("/input/data.csv")


# Search for missing data
for column in data.columns:
    if data[column].isna().sum() != 0:
        missing = data[column].isna().sum()
        portion = (missing / data.shape[0]) * 100
        print(f"'{column}':\n\tnumber of missing values '{missing}' ==>
'{portion:.3f}%'")


# Removing unnecessary data
data.drop('emp_title', axis=1, inplace=True)
data.drop('emp_length', axis=1, inplace=True)
data.drop('title', axis=1, inplace=True)

# Recovering missed data
```

```
data.dropna(subset=['revol_util', 'pub_rec_bankruptcies'], inplace=True)
total_acc_avg = data[['total_acc', 'mort_acc']].groupby(by='total_acc').mean().mort_acc

def fill_mort_acc(total_acc, mort_acc):
    if np.isnan(mort_acc):
        return total_acc_avg[total_acc].round()
    else:
        return mort_acc

data['mort_acc'] = data.apply(lambda x: fill_mort_acc(x['total_acc'], x['mort_acc']), axis=1)

data.loc[(data.home_ownership == 'ANY') | (data.home_ownership == 'NONE'), 'home_ownership'] = 'OTHER'


# Removing unnecessary data
data.drop('sub_grade', axis=1, inplace=True)
data.drop('issue_d', axis=1, inplace=True)
data.drop('earliest_cr_line', axis=1, inplace=True)
data.drop('address', axis=1, inplace=True)
data.drop('initial_list_status', axis=1, inplace=True)


# Preparing data for application
target_default = data['loan_status'] == 'Charged Off'
X = data.drop(['loan_status', 'default_term'], axis=1)
X = pd.get_dummies(X, prefix=['term', 'grade', 'home_ownership', 'verification_status', 'application_type', 'purpose'])


# Scaling and splitting data into samples
```

```
scaler = StandardScaler()
X = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, target_default,
test_size=.2, random_state=111122)



# Data synthesis for unbalanced classes
sm = SMOTE(sampling_strategy='minority', random_state=111122)

X_train_oversampled,  y_train_oversampled  =  sm.fit_resample(X_train,
y_train)



# Logistic Regression
clf_lr               =               LogisticRegression(max_iter=10000,
C=0.1).fit(X_train_oversampled, y_train_oversampled)
clf = clf_lr

y_predict_proba = clf.predict_proba(X_train_oversampled)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_train_oversampled,             y_predict,
target_names=['Non-Default', 'Default']))

ns_probs = [0 for _ in range(len(y_train_oversampled))]

lr_probs = clf.predict_proba(X_train_oversampled)
lr_probs = lr_probs[:, 1]

ns_auc = roc_auc_score(y_train_oversampled, ns_probs)
lr_auc = roc_auc_score(y_train_oversampled, lr_probs)

ns_fpr, ns_tpr, _ = roc_curve(y_train_oversampled, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_train_oversampled, lr_probs)
```

```
plt.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic')

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()

cf_matrix = confusion_matrix(y_train_oversampled, y_predict)
sns.heatmap(cf_matrix, annot=True, fmt='.0f', cmap='Blues')

y_predict_proba = clf.predict_proba(X_test)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_test,   y_predict,   target_names=['Non-
Default', 'Default']))

# Gaussian Naïve Classificator
clf_gnb = GaussianNB()
clf_gnb.fit(X_train_oversampled, y_train_oversampled)
clf = clf_gnb

y_predict_proba = clf.predict_proba(X_train_oversampled)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_train_oversampled,            y_predict,
target_names=['Non-Default', 'Default']))

ns_probs = [0 for _ in range(len(y_train_oversampled))]

lr_probs = clf.predict_proba(X_train_oversampled)
lr_probs = lr_probs[:, 1]

ns_auc = roc_auc_score(y_train_oversampled, ns_probs)
```

```
lr_auc = roc_auc_score(y_train_oversampled, lr_probs)

ns_fpr, ns_tpr, _ = roc_curve(y_train_oversampled, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_train_oversampled, lr_probs)

plt.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic')

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()

cf_matrix = confusion_matrix(y_train_oversampled, y_predict)
sns.heatmap(cf_matrix, annot=True, fmt='.0f', cmap='Blues')

y_predict_proba = clf.predict_proba(X_test)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_test,   y_predict,   target_names=['Non-
Default', 'Default']))



# Decision Tree
clf_dt = DecisionTreeClassifier(max_depth=32)
clf_dt.fit(X_train_oversampled, y_train_oversampled)
clf = clf_dt

y_predict_proba = clf.predict_proba(X_train_oversampled)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_train_oversampled,            y_predict,
target_names=['Non-Default', 'Default']))

ns_probs = [0 for _ in range(len(y_train_oversampled))]
```

```
lr_probs = clf.predict_proba(X_train_oversampled)
lr_probs = lr_probs[:, 1]

ns_auc = roc_auc_score(y_train_oversampled, ns_probs)
lr_auc = roc_auc_score(y_train_oversampled, lr_probs)

ns_fpr, ns_tpr, _ = roc_curve(y_train_oversampled, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_train_oversampled, lr_probs)

plt.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic')

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()

cf_matrix = confusion_matrix(y_train_oversampled, y_predict)
sns.heatmap(cf_matrix, annot=True, fmt='.0f', cmap='Blues')

y_predict_proba = clf.predict_proba(X_test)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_test,   y_predict,   target_names=['Non-
Default', 'Default']))



# Random Forest
clf_rf = RandomForestClassifier(max_depth=24)
clf.rf.fit(X_train_oversampled, y_train_oversampled)
clf = clf_rf

y_predict_proba = clf.predict_proba(X_train_oversampled)
```

```
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_train_oversampled, y_predict,
target_names=['Non-Default', 'Default']))

ns_probs = [0 for _ in range(len(y_train_oversampled))]

lr_probs = clf.predict_proba(X_train_oversampled)
lr_probs = lr_probs[:, 1]

ns_auc = roc_auc_score(y_train_oversampled, ns_probs)
lr_auc = roc_auc_score(y_train_oversampled, lr_probs)

ns_fpr, ns_tpr, _ = roc_curve(y_train_oversampled, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_train_oversampled, lr_probs)

plt.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic')

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()

cf_matrix = confusion_matrix(y_train_oversampled, y_predict)
sns.heatmap(cf_matrix, annot=True, fmt='.0f', cmap='Blues')

y_predict_proba = clf.predict_proba(X_test)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_test,   y_predict,   target_names=['Non-
Default', 'Default']))



# XGBoost
```

```
clf_gbt = xgb.XGBClassifier()
clf_gbt.fit(X_train_oversampled, y_train_oversampled)
clf = clf_gbt

y_predict_proba = clf.predict_proba(X_train_oversampled)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
print(classification_report(y_train_oversampled,             y_predict,
target_names=['Non-Default', 'Default']))

ns_probs = [0 for _ in range(len(y_train_oversampled))]

lr_probs = clf.predict_proba(X_train_oversampled)
lr_probs = lr_probs[:, 1]

ns_auc = roc_auc_score(y_train_oversampled, ns_probs)
lr_auc = roc_auc_score(y_train_oversampled, lr_probs)

ns_fpr, ns_tpr, _ = roc_curve(y_train_oversampled, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_train_oversampled, lr_probs)

plt.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic')

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()

cf_matrix = confusion_matrix(y_train_oversampled, y_predict)
sns.heatmap(cf_matrix, annot=True, fmt='.0f', cmap='Blues')

y_predict_proba = clf.predict_proba(X_test)
y_predict = [1 if x>0.5 else 0 for x in y_predict_proba[:, 1]]
```

```
print(classification_report(y_test, y_predict, target_names=['Non-
Default', 'Default']))


# Kaplan-Meier Estimator
kmf = KaplanMeierFitter()

tenure = data['default_term'].fillna(value=0.0)
event = data['loan_status'] == 'Charged Off'

kmf.fit(tenure, event, label='Probability of Default')
kmf.plot_cumulative_density()

grade = data['grade']


# Kaplan-Meier score (according to grade)
kmf = KaplanMeierFitter()

tenure = data['default_term'].fillna(value=0.0)
event = data['loan_status'] == 'Charged Off'

kmf.fit(tenure[grade=='A'], event[grade=='A'], label='Grade A
Probability of Default')
kmf.plot_cumulative_density()

kmf.fit(tenure[grade=='B'], event[grade=='B'], label='Grade B
Probability of Default')
kmf.plot_cumulative_density()

kmf.fit(tenure[grade=='C'], event[grade=='C'], label='Grade C
Probability of Default')
kmf.plot_cumulative_density()
```

```
kmf.fit(tenure[grade=='D'],    event[grade=='D'],    label='Grade    D
Probability of Default')
kmf.plot_cumulative_density()

kmf.fit(tenure[grade=='E'],    event[grade=='E'],    label='Grade    E
Probability of Default')
kmf.plot_cumulative_density()

kmf.fit(tenure[grade=='F'],    event[grade=='F'],    label='Grade    F
Probability of Default')
kmf.plot_cumulative_density()

kmf.fit(tenure[grade=='G'],    event[grade=='G'],    label='Grade    G
Probability of Default')
kmf.plot_cumulative_density()

plt.legend(bbox_to_anchor=(1.02, 1), loc='upper left')
plt.show()


# Testing the developed model
clf = clf_gbt

portfolio = data.sample(250000, random_state=17102023)

portfolio['default_term'] = portfolio['default_term'].fillna(value=0.0)

exp_default_term                =                data.groupby(['grade',
'term'])['default_term'].mean()

curr_loan = portfolio.drop(['loan_status', 'default_term'], axis=1)
curr_loan   =   pd.get_dummies(curr_loan,   prefix=['term',   'grade',
'home_ownership',      'verification_status',      'application_type',
'purpose'])
```

```
curr_loan = scaler.transform(curr_loan)

portfolio['prob_default_gbt'] = clf.predict_proba(curr_loan)[:, 1]

portfolio['exp_loss'] =  portfolio.apply(
    lambda x: x.prob_default_gbt * x.loan_amnt * (1 + x.int_rate *
int(x.term[1:3]) / 1200.0) * (1 - exp_default_term[x.grade, x.term] /
int(x.term[1:3])),
    axis=1
)

print('Expected                loss:                $'                +
"{:,.2f}".format(np.sum(portfolio['exp_loss'])))

portfolio['unexp_loss'] =  portfolio.apply(
    lambda   x:   3*np.sqrt((x.loan_amnt   *   (1   +   x.int_rate   *
int(x.term[1:3]) / 1200.0) * (1 - exp_default_term[x.grade, x.term] /
int(x.term[1:3])))**2) * x.prob_default_gbt - x.exp_loss,
    axis=1
)

print('Unxpected                loss:                $'                +
"{:,.2f}".format(np.sum(portfolio['unexp_loss'])))

portfolio['real_loss'] =  portfolio.apply(
    lambda x: (x.loan_status=='Charged Off') * x.loan_amnt * (1 +
x.int_rate * int(x.term[1:3]) / 1200.0) * (1 - x.default_term /
int(x.term[1:3])),
    axis=1
)

print('Real                  loss:                $'                +
"{:,.2f}".format(np.sum(portfolio['real_loss'])))
```