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 004.021:004.8, 336.71

До захисту допущено
Завідувач кафедри ММСА
_____Оксана ТИМОЩУК
«__» _____2025 р.

Магістерська дисертація

на здобуття ступеня магістра
за освітньо-професійною програмою «Системний аналіз і управління»
зі спеціальності 124 «Системний аналіз»
на тему: «Інтелектуальні системи для прогнозування у банківській сфері»

Виконав:
Студент 2 курсу, групи КА-41мп
Міщенко Антон Сергійович _____

Науковий керівник:
ст. викл. каф. ММСА, д. філософії
Гуськова Віра Геннадіївна _____

Консультант з нормокнтролю:
Доктор філософії з сис. аналізу, асистент
Канцедал Георгій Олегович _____

Рецензент:
К.т.н., асистент кафедри ШІ
Осауленко Вячеслав Миколайович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань

Студент: _____

Київ – 2025

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ
Рівень вищої освіти — другий (магістерський)
Спеціальність — 124«Системний аналіз»
Освітньо-професійною програмою «Системний аналіз і управління»

ЗАТВЕРДЖУЮ

Завідувач кафедри ММСА
_____Оксана ТИМОЩУК
«___» _____2025 р.

ЗАВДАННЯ

на магістерську дисертацію студенту

Міщенко Антону Сергійовичу

1. Тема дисертації: Інтелектуальні системи для прогнозування у банківській сфері, науковий керівник дисертації ст. викл. каф. ММСА, д. філософії Гуськова Віра Геннадіївна затверджені наказом по університету від «06» листопада 2025 р. № 4837-с
2. Строк подання студентом дисертації: 16.12.2025
3. Об’єкт дослідження: процес підтримки прийняття рішень у банківських установах під час прогнозування взяття ключових продуктів.
4. Предмет дослідження: методи, моделі та алгоритми машинного навчання й ансамблеві підходи, що забезпечують підвищення точності прогнозування.
5. Перелік завдань, які потрібно розробити:
   1) огляд предметної області;
   2) підготовка бази даних до роботи;
   3) дослідження методів попередньої обробки даних, анонімізації;
   4) вибір та аналіз методів кластеризації, включно з вибором оптимальної кількості кластерів та профілюванням сегментів;
   5) вибір та аналіз алгоритмів для прогнозування поведінкових задач (депозит/кредит);
   6) вибір метрик для оцінювання моделей та інтерпретація результатів;
   7) створення апаратно-програмного прототипу та його тестування;
6. Перелік графічного (ілюстративного) матеріалу:
   1) рисунки;
   2) таблиці;

3) презентація.

7. Орієнтовний перелік публікацій: Міщенко А.С., Гуськова В.Г. Інтелектуальні системи для прогнозування у банківській сфері. Системні науки та інформатика: збірник доповідей IV Всеукраїнської науково-практичної конференції «Системні науки та інформатика», 01–05 грудня 2025 року, Київ. К., НН ІПСА КПІ ім. Ігоря Сікорського, 2025, С. 39-44.

8. Консультанти розділів дисертації

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

9. Дата видачі завдання: 01 вересня 2025 року

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації (МД)	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»	01.09.2025-10.09.2025	виконано
2	Ознайомлення з ДСТУ 3008:2015.	11.09.2025-17.09.2025	виконано
3	Перший розділ. Огляд літературно інформаційних джерел. Аналіз предметної області.	18.09.2025-01.10.2025	виконано
4	Другий розділ. Розробка теоретичного узагальнення методу.	02.10.2025-15.10.2025	виконано
5	Третій розділ. Розробка програмного забезпечення.	16.10.2025-29.10.2025	виконано
6	Третій розділ. Робота над практичним розділом магістерської дисертації.	30.10.2025-12.11.2025	виконано
7	Четвертий розділ. Стартап-проект.	13.11.2025-19.11.2025	виконано
8	Остаточне оформлення роботи	20.11.2025-15.12.2025	виконано

Студент _____ Антон МІЩЕНКО

Науковий керівник дисертації _____ Віра ГУСЬКОВА

РЕФЕРАТ

Магістерська дисертація: 109 с., 9 рис., 20 табл., 2 дод., 19 джерел.

СЕГМЕНТАЦІЯ КЛІЄНТІВ, МАШИННЕ НАВЧАННЯ, ПРОГНОЗУВАННЯ ПОВЕДІНКИ, БАНКІВСЬКИЙ МАРКЕТИНГ, GMM, RANDOM FOREST, DATA LEAKAGE.

Темою роботи є розроблення інтелектуальної системи аналітики портфеля клієнтів банку, яка поєднує сегментацію та прогнозування ключових поведінкових подій.

Об’єктом дослідження є процес підтримки прийняття рішень у банківських установах під час прогнозування взяття ключових продуктів.

Предметом дослідження є методи машинного навчання для сегментації клієнтів, оцінки їхніх ймовірностей взаємодії з банківськими продуктами та побудови персоналізованих рекомендацій.

Метою роботи є створення аналітичного інструментарію, який дозволяє автоматично сегментувати клієнтів, будувати моделі прогнозування їхньої поведінки та формувати дані для таргетованих маркетингових кампаній.

Актуальність роботи зумовлена стрімким переходом банків до персоналізованого маркетингу, необхідністю оптимізації витрат на залучення клієнтів та підвищення точності прогнозів.

У результаті роботи було створено повний машинно-навчальний пайплайн, що включає попередню обробку даних, сегментацію клієнтів, а також побудову моделей прогнозування. Розроблена система формує сегментовані клієнтські бази з оцінкою ймовірності відкриття депозиту чи кредиту, що дозволяє використовувати її як інтелектуальний модуль для підготовки маркетингових кампаній та підвищення ефективності роботи банку.

ABSTRACT

Master's Thesis: 109 pages, 9 fig., 20 tabl., 2 appendices, 19 references.

CLIENT SEGMENTATION, MACHINE LEARNING, BEHAVIORAL FORECASTING, BANKING MARKETING, GMM, RANDOM FOREST, DATA LEAKAGE.

The topic of the thesis is the development of an intelligent analytics system for a bank's client portfolio that integrates segmentation and prediction of key behavioral events.

The object of the study is the decision-support process in banking institutions during forecasting of customer uptake of core financial products.

The subject of the study is the set of machine learning methods for client segmentation, assessment of their probability of interacting with banking products, and the construction of personalized recommendations.

The purpose of the thesis is to create an analytical toolkit that enables automatic segmentation of clients, development of behavioral prediction models, and generation of data for targeted marketing campaigns.

The relevance of the research is driven by the rapid shift of banks toward personalized marketing, the need to optimize customer acquisition costs, and the growing demand for higher accuracy of predictive models.

As a result of the work, a complete machine-learning pipeline was developed, including data preprocessing, client segmentation, and the construction of prediction models. The developed system generates segmented customer datasets with estimated probabilities of opening a deposit or applying for a credit product, which enables its use as an intelligent module for preparing marketing campaigns and increasing the bank's operational efficiency.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ 8
ВСТУП 9
РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ РОБОТИ 11
1.1 Проблематика управління клієнтським портфелем. 11
1.2 Теоретичні основи аналізу клієнтської поведінки. 12
1.3 Методи сегментації клієнтів у банківській сфері 13
1.4 Теоретичні основи прогнозного моделювання у банківській сфері .. 14
1.5 Теоретичні засади підготовки та структурування даних 15
1.6 Інтелектуальні системи підтримки прийняття рішень у банках 16
1.7 Висновки до розділу 1 17
РОЗДІЛ 2 СИСТЕМНИЙ ПІДХІД ДО ПОБУДОВИ ІНСТРУМЕНТАРІЮ ПРОГНОЗУВАННЯ КЛІЄНТСЬКОГО ПОРТФЕЛЮ 19
2.1 Обґрунтування вибору методів прогнозування та сегментації 19
2.2 Архітектура інтелектуальної системи прогнозування клієнтського портфелю 22
2.3 Формування та опис вибірки даних 24
2.4 Модуль попередньої обробки даних 25
2.5 Валідація, інтерпретація та профілювання клієнтських сегментів 27
2.6 Формалізація моделей прогнозування поведінки клієнтів 28
2.7 Оцінювання якості прогнозних моделей 30
2.8 Інтеграція сегментації та прогнозування в систему підтримки прийняття рішень 31
2.9 Висновки до розділу 2 33
РОЗДІЛ 3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНТ ЗА ТЕМОЮ МД 34
3.1 Проблема формування клієнтського портфелю 34
3.2 Опис вхідних даних та їх обробка 35
3.3 Формування набору ознак та моделювання 37
3.4 Сегментація клієнтської бази 39
3.5 Побудова моделей 45
3.6 Результати сегментації: прогноз депозитів і кредитів 48

3.7 Висновки до розділу 3 51
РОЗДІЛ 4 РОЗРОБКА ВЛАСНОГО СТАРТАП ПРОЕКТУ 53
4.1 План розробки стартапу та масштабування його на ринок 54
4.2 Опис ідеї стартап-проекту 55
1.3 Технологічний аудит ідеї проекту 57
1.4 Аналіз ринкових можливостей запуску стартап-проекту 61
1.5 Розроблення ринкової стратегії стартап-проекту 74
1.6 Розроблення маркетингової програми стартап-проекту 79
1.7 Висновки до розділу 4 82
ВИСНОВКИ 84
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ 86
ДОДАТОК А. ЛІСТИНГ ПРОГРАМИ 88
ДОДАТОК Б. ПРЕЗЕНТАЦІЯ 100

ПЕРЕЛІК СКОРОЧЕНЬ

AI — Artificial Intelligence, штучний інтелект.

BI — Business Intelligence, бізнес-аналітика.

BIC — Bayesian Information Criterion, байєсівський інформаційний критерій.

DSS — Decision Support System, система підтримки прийняття рішень.

EDA — Exploratory Data Analysis, розвідувальний аналіз даних.

EMA — Exponential Moving Average, експоненційне ковзне середнє.

FN — False Negative, хибнонегативне рішення.

FP — False Positive, хибнопозитивне рішення.

GMM — Gaussian Mixture Models, модель суміші гауссових розподілів.

KPI — Key Performance Indicator, ключовий показник ефективності.

ML — Machine Learning, машинне навчання.

PCA — Principal Component Analysis, метод головних компонент.

ROC-AUC — Receiver Operating Characteristic – Area Under Curve, площа під ROC-кривою.

ROI — Return on Investment, показник рентабельності інвестицій.

SSE — Sum of Squared Errors, сума квадратів похибок.

SQL — Structured Query Language, мова структурованих запитів.

TN — True Negative, правильно негативне рішення.

TP — True Positive, правильно позитивне рішення.

ВСТУП

У сучасних умовах цифрової трансформації банківський сектор зазнає суттєвих структурних змін, пов’язаних із зростанням конкуренції, динамічністю фінансових ринків та швидким розвитком технологій обробки даних. Ключовим чинником ефективності банківських установ стає здатність точно аналізувати та прогнозувати поведінку клієнтів, адже своєчасне виявлення ризиків, визначення потреб і тенденцій формує основу для конкурентоспроможної кредитної та маркетингової політики. Мільйони транзакцій, багатогранність демографічних та поведінкових характеристик клієнтів, а також вплив зовнішніх економічних факторів створюють надзвичайно складне інформаційне середовище, у якому традиційні аналітичні методи стають недостатніми.

Стрімкий розвиток машинного навчання, інтелектуальних алгоритмів та систем підтримки прийняття рішень відкриває можливості для якісно нового підходу до управління клієнтським портфелем. Моделі класифікації здатні з високою точністю прогнозувати ймовірність відкриття депозиту чи отримання кредиту, а алгоритми кластеризації — виявляти приховані сегменти клієнтів, що раніше залишалися невидимими у традиційній аналітиці. Інтеграція таких інструментів дозволяє банкам формувати персоналізовані фінансові пропозиції, оптимізувати маркетингові бюджети та зменшувати ризики, пов’язані з кредитуванням.

Ця робота спрямована на створення інтелектуальної системи прогнозування поведінки клієнтів банку, яка поєднує алгоритми сегментації, прогнозного моделювання та візуалізації аналітичної інформації. Центральною ідеєю є побудова комплексного рішення, здатного автоматично обробляти великі масиви клієнтських даних, визначати ключові поведінкові патерни, формувати релевантні сегменти та прогнозувати ймовірність відкриття фінансових продуктів. Особливу увагу приділено точності моделей,

відтворюваності процесів аналізу та інтерпретованості результатів - критично важливих факторів для банківської сфери.

Актуальність дослідження зумовлена потребою банків оперативно приймати рішення в умовах високої мінливості економічного середовища, коливання платоспроможності населення та зростаючих вимог регуляторів. Інтелектуальна система, що поєднує машинне навчання, кластерний аналіз та автоматизовану обробку даних, здатна зменшити вплив людського фактора, підвищити якість ризик-менеджменту та забезпечити більш точне таргетування продуктів.

Практична цінність роботи полягає у можливості застосування розробленого підходу для банківських установ, фінансово-аналітичних підрозділів та маркетингових департаментів. Побудована система може бути інтегрована у внутрішні інформаційні процеси банку та використовуватися для формування персоналізованих пропозицій, прогнозування попиту на продукти та підвищення загальної прибутковості клієнтського портфелю.

РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ РОБОТИ

1.1 Проблематика управління клієнтським портфелем

Управління клієнтським портфелем є одним з ключових напрямів діяльності банківських установ, оскільки саме структура та якість клієнтської бази визначають фінансові результати банку, рівень його стійкості та здатність адаптуватися до ринкових змін. Клієнтський портфель можна визначити як сукупність фізичних та юридичних осіб, які взаємодіють із банком, користуючись його продуктами та послугами. Він відображає не лише поточний стан активності клієнтів, а й потенційні можливості для зростання, формування доходів, оптимізації ризиків та стратегічного розвитку. У цьому контексті клієнтський портфель виступає основою для прийняття управлінських рішень у сферах маркетингу, кредитної політики, ризик-менеджменту та операційної діяльності.

Разом з тим сучасні умови функціонування банківської системи суттєво ускладнюють процес аналізу та управління клієнтською базою. Однією з ключових проблем є різноманітність і великий обсяг даних, які генеруються кожним клієнтом у процесі взаємодії з банком. Традиційні інструменти аналізу часто не здатні оперативно опрацьовувати такі потоки інформації, що призводить до втрати важливих закономірностей і тенденцій. Додаткові виклики виникають у зв'язку зі зростанням кількості каналів комунікації, цифровізацією фінансових послуг та підвищенням очікувань клієнтів щодо персоналізації сервісу.

Суттєвим фактором є також висока конкуренція на банківському ринку. Клієнти мають доступ до широкого спектра фінансових інструментів та можуть легко змінювати банк у разі незадоволення умовами обслуговування. Це зумовлює необхідність глибокого розуміння поведінки клієнтів, їх потреб, мотивації та потенційної реакції на маркетингові або ризикові рішення банку. Динамічність ринку, зокрема зміни економічної ситуації, рівня доходів

населення або регуляторних вимог, додатково ускладнює аналіз та потребує адаптивних інструментів прогнозування.

У таких умовах особливої актуальності набуває прогнозування та сегментація клієнтів. Сегментація дозволяє об’єднати клієнтів у групи зі схожими характеристиками, що спрощує розробку персоналізованих продуктів і визначення стратегічно важливих сегментів. Прогнозне моделювання, у свою черзі, допомагає оцінити ймовірність певних дій клієнтів, таких як відкриття депозиту, отримання кредиту або відтік з банку. Це підвищує ефективність управлінських рішень, оптимізує маркетингові бюджети та мінімізує кредитні ризики.

Обмеження традиційних методів аналізу є однією з ключових причин впровадження інтелектуальних систем. Класичні статистичні інструменти, такі як регресійні моделі або експертні підходи, не завжди здатні враховувати нелінійні залежності, складні взаємозв'язки між ознаками та великі обсяги даних. Вони вимагають значної кількості припущень і часто не можуть забезпечити високу точність прогнозів у динамічному середовищі. Це призводить до втрати інформації, зниження ефективності маркетингових кампаній та підвищених ризиків кредитування.

Таким чином, проблематика управління клієнтським портфелем банку зумовлена комплексом факторів — високою конкуренцією, швидкою зміною ринку, різноманітністю джерел даних і необхідністю точного прогнозування поведінки клієнтів. Для її вирішення потрібні сучасні методи аналізу, які поєднують машинне навчання, сегментацію та інтелектуальні системи підтримки прийняття рішень, що дозволяють підвищити ефективність управління та адаптуватися до швидкоплинних умов ринку.

1.2 Теоретичні основи аналізу клієнтської поведінки

Аналіз клієнтської поведінки є важливим елементом банківського менеджменту, оскільки дозволяє виявляти потреби клієнтів, оцінювати їхню

активність та формувати стратегії взаємодії. Під клієнтською поведінкою розуміють сукупність дій та рішень клієнта, що проявляються у використанні ним фінансових продуктів та каналів обслуговування.

У теоретичній літературі поведінка клієнтів розглядається як результат впливу соціально-демографічних, економічних та поведінкових факторів, що формують індивідуальну модель взаємодії з банком. Одним із базових підходів є концепція життєвого циклу клієнта, яка дає змогу описати етапи його розвитку від залучення до можливого відтоку та обґрунтувати відповідні інструменти управління [1].

Суттєве значення має також сегментація клієнтів, що дозволяє поділити їх на групи зі схожими характеристиками та підвищити точність прогнозування поведінкових реакцій. У сучасних дослідженнях все більшої ваги набувають поведінкові та мотиваційні ознаки, які забезпечують глибше розуміння моделей прийняття клієнтами фінансових рішень.

Додатковим теоретичним аспектом є ризик-орієнтований підхід, який враховує ймовірність виникнення небажаних подій, таких як прострочення чи відтік клієнтів. Використання таких підходів дозволяє підвищити якість кредитної політики та оптимізувати управління клієнтською базою.

1.3 Методи сегментації клієнтів у банківській сфері

Сегментація клієнтів є важливим інструментом банківського менеджменту, оскільки дозволяє поділити клієнтів на відносно однорідні групи, що спрощує розробку продуктів, підвищує ефективність маркетингових кампаній та покращує управління ризиками. Сутність сегментації полягає у виявленні спільних характеристик клієнтів та формуванні сегментів, для яких можна застосовувати диференційовані підходи.

У класичному розумінні сегментація ґрунтується на соціально-демографічних, економічних та поведінкових ознаках. Однак зі зростанням обсягів даних та ускладненням поведінкових моделей клієнтів дедалі більшого

значення набувають алгоритмічні методи сегментації. Найпоширенішими є методи кластерного аналізу, які дозволяють автоматично виокремлювати групи клієнтів на основі схожості їхніх характеристик [2].

Сегментація може базуватися на ієрархічних або щільнісних алгоритмах, проте у банківській сфері частіше застосовуються саме центроїдні методи, оскільки вони забезпечують достатню інтерпретованість та придатність для практичних управлінських рішень.

Таким чином, методи сегментації є важливою складовою аналізу клієнтської бази, а використання алгоритмічних підходів забезпечує підвищення точності, масштабованості та об’єктивності результатів.

1.4 Теоретичні основи прогнозного моделювання у банківській сфері

Прогнозне моделювання є одним з ключових інструментів аналітики у банківській галузі, оскільки дозволяє оцінити ймовірність майбутніх дій клієнтів та формувати обґрунтовані управлінські рішення. У контексті управління клієнтським портфелем прогнозування використовується для оцінки ймовірності відкриття депозитів, залучення кредитів, зміни активності або відтоку клієнтів. Застосування таких моделей сприяє зменшенню фінансових ризиків та підвищенню ефективності маркетингових стратегій.

Теоретичною основою прогнозного моделювання є класифікаційні та регресійні методи, що дозволяють встановлювати залежності між вхідними ознаками клієнтів та їхньою майбутньою поведінкою. У банківській практиці одними з найпоширеніших є лінійні моделі, зокрема логістична регресія, яка використовується для оцінювання ймовірності подій, що мають бінарний характер. Перевагою таких моделей є їх висока інтерпретованість та здатність надавати оцінку важливості кожної ознаки.

Разом з тим розвиток машинного навчання привів до широкого застосування нелінійних та ансамблевих методів, таких як Random Forest та Gradient Boosting. Ці моделі здатні враховувати складні взаємозв'язки між ознаками, працювати з великою кількістю параметрів та забезпечувати високу

точність прогнозів у порівнянні з класичними підходами. Їхня ефективність зумовлена механізмом агрегування результатів множини дерев рішень, що дозволяє зменшувати ризик переобучення та підвищувати узагальнювальну здатність моделі.

Важливим аспектом прогнозного моделювання є оцінювання якості моделей, яке ґрунтується на використанні метрик, що характеризують точність, повноту, ступінь помилок та здатність моделі розрізняти класи. Серед таких метрик найчастіше застосовуються Accuracy, Precision, Recall, F1-score та ROC-AUC. Кожна з них дає можливість оцінити модель з різних позицій та визначити її придатність для конкретних управлінських задач.

Таким чином, прогнозне моделювання у банківській сфері базується на поєднанні теоретичних методів статистики та сучасних алгоритмів машинного навчання. Використання різних типів моделей дозволяє підвищити точність прогнозів, забезпечити адаптивність рішень та створити основу для побудови інтелектуальних систем управління клієнтським портфелем.

1.5 Теоретичні засади підготовки та структурування даних

Підготовка даних є одним з найважливіших етапів побудови аналітичних та прогнозних моделей у банківській сфері. Якість даних безпосередньо визначає точність, стабільність та інтерпретованість результатів, тому правильна організація цього процесу має ключове значення для будь-якої системи прогнозування. Банківські дані зазвичай включають демографічні характеристики клієнтів, транзакційну історію, показники кредитної активності, поведінкові та соціально-економічні фактори, що потребують узгодження та інтеграції.

Однією з теоретичних проблем підготовки даних є наявність пропусків, шумів та аномальних значень. Пропуски можуть виникати через технічні збої, різні формати даних або неповну взаємодію клієнта з банком. У таких випадках застосовуються методи імпутації, які дозволяють замінити відсутні значення середніми величинами, модами або прогнозними оцінками.

Аномальні значення, своєю чергою, можуть спотворювати результати моделювання, тому їх виявлення та усунення є необхідною складовою процесу.

Наступним важливим аспектом є нормалізація та масштабування ознак, що забезпечує коректну роботу моделей, особливо тих, які чутливі до різниці в масштабах даних. Категоріальні ознаки потребують кодування, яке перетворює текстові або номінальні значення у числовий формат, придатний для алгоритмів машинного навчання. Широко використовуються методи one-hot кодування та порядкового перетворення.

Особливу увагу необхідно приділяти проблемі витоку інформації, яка виникає тоді, коли у моделювання потрапляють ознаки, що відображають події, невідомі на момент прийняття рішення. Така ситуація призводить до штучного завищення точності моделей та їх непридатності для реального застосування. Тому теоретичні засади підготовки даних передбачають чітке розмежування ознак, допустимих для навчання, та тих, що слід вилучити.

У підсумку підготовка даних є етапом, що поєднує очищення, трансформацію, структурування та контроль якості входових параметрів. Без належної організації цього процесу неможливо побудувати ефективну систему прогнозування клієнтської поведінки.

1.6 Інтелектуальні системи підтримки прийняття рішень у банках

Інтелектуальні системи підтримки прийняття рішень є сучасним інструментом управління, що дозволяє банкам аналізувати великі обсяги даних, прогнозувати ризики та формувати обґрунтовані стратегії. Такі системи інтегрують методи машинного навчання, статистичні моделі та аналітичні механізми для автоматизації процесів оцінювання клієнтів і прийняття управлінських рішень.

У класичному вигляді система підтримки рішень складається з кількох компонентів — бази даних, аналітичного модуля, механізму моделювання та інтерфейсу користувача. У банківській сфері ці компоненти доповнюються

модулями прогнозування, сегментації та моніторингу, що дозволяє системі не лише аналізувати історичні дані, а й формувати прогнози майбутніх подій, таких як ймовірність оформлення кредиту або відкриття депозиту.

Важливою особливістю інтелектуальних систем є їх здатність адаптуватися до змін ринку та поведінки клієнтів. За рахунок використання алгоритмів машинного навчання такі системи можуть оновлювати знання та коригувати параметри моделей у відповідь на нові дані. Це підвищує стійкість до змін зовнішнього середовища та забезпечує актуальність прогнозів.

Значну роль відіграє також інтерпретованість системи підтримки рішень, оскільки банківські менеджери повинні розуміти логіку прийняття рішень. Інтелектуальні системи забезпечують візуалізацію результатів, пояснення прогнозів та інструкції щодо дій, що сприяє використанню моделі у практичних умовах.

Таким чином, інтелектуальні системи підтримки рішень є невід'ємною складовою сучасного банківського управління. Вони забезпечують підвищення ефективності аналізу даних, оптимізацію маркетингових і кредитних стратегій та створюють основу для впровадження інновацій у сфері клієнтської аналітики.

1.7 Висновки до розділу 1

У першому розділі було розглянуто теоретичні засади управління клієнтським портфелем банку та аналітичні інструменти, що формують основу для побудови інтелектуальних систем прогнозування. Визначено, що проблематика управління клієнтською базою зумовлена високою конкуренцією, динамічністю ринку та зміною поведінкових моделей клієнтів, що потребує застосування сучасних методів аналізу та прогнозування.

Проаналізовано теоретичні підходи до вивчення клієнтської поведінки, зокрема модель життєвого циклу клієнта, поведінкові та ризик-орієнтовані концепції. Показано роль сегментації як ключового інструменту структуризації клієнтської бази, а також охарактеризовано основні методи

кластеризації, включно з алгоритмічними підходами, що дозволяють формувати більш точні та об’єктивні сегменти.

Окрему увагу приділено теорії прогнозного моделювання, яке у банківській сфері використовується для оцінки ймовірності майбутніх дій клієнтів. Розглянуто лінійні та ансамблеві методи, їх переваги та обмеження, а також основні метрики оцінки якості моделей.

Також окреслено теоретичні аспекти підготовки даних, які включають очищення, нормалізацію, кодування ознак та запобігання витоку інформації. Наголошено, що якість даних є критичним чинником успішності моделей машинного навчання.

Завершальним елементом теоретичного аналізу стало вивчення інтелектуальних систем підтримки прийняття рішень у банківській сфері, які інтегрують аналітичні модулі, прогностичні алгоритми та механізми інтерпретації результатів. Подібні системи забезпечують автоматизацію управлінських процесів та підвищення ефективності роботи банку.

Узагальнюючи, перший розділ сформував теоретичне підґрунтя для подальшої розробки інструментарію прогнозування клієнтського портфелю, що буде детально розглянуто у наступному розділі.

РОЗДІЛ 2 СИСТЕМНИЙ ПІДХІД ДО ПОБУДОВИ ІНСТРУМЕНТАРІЮ ПРОГНОЗУВАННЯ КЛІЄНТСЬКОГО ПОРТФЕЛЮ

2.1 Обґрунтування вибору методів прогнозування та сегментації

Вибір методів сегментації є ключовим етапом розроблення інтелектуальної системи прогнозування, оскільки саме цей елемент визначає структуризацію клієнтської бази, формування однорідних груп та подальшу точність моделей машинного навчання. У рамках дослідження первинно розглядався алгоритм KMeans як базовий центричний метод кластеризації. Попри його поширеність, аналіз структури банківських даних показав наявність низки обмежень, що унеможливлюють отримання високоточних сегментів. Тому у роботі обґрунтовано перехід до більш гнучкого підходу — Gaussian Mixture Models (GMM), який краще описує реальний розподіл клієнтів та забезпечує високу якість кластеризації [3].

Алгоритм KMeans належить до жорстких кластеризаційних методів і передбачає поділ набору спостережень на K кластерів шляхом мінімізації внутрі кластерної дисперсії. Формальна цільова функція має вигляд:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

де C_i– множина спостережень, що належать і-му кластеру;

μ_i – центр відповідного кластера.

Алгоритм складається з декількох кроків.

1. Ініціалізація випадкових центрів $\mu_1, \ldots, \mu_k$.
2. Призначення об’єктів до найближчого центру за метрикою Евкліда:

$$C_i = \{x: \|x - \mu_i\| \quad = \|x - \mu_j\|\}.$$

3. Оновлення центрів:

$$\mu_i = \frac{4}{|C_i|} \sum_{x \in C_i} x.$$

4. Повторення кроків до збіжності.

Хоча KMeans є швидким та обчислювально ефективним, він має фундаментальні обмеження.

1. Метод коректно працює лише тоді, коли кластери мають форму приблизно круглих хмар даних. У банківських вибірках сегменти часто розтягнуті, еліптичні або перекриті.
2. KMeans фактично припускає, що всі кластери мають однакову варіабельність, що суперечить реальній структурі клієнтів.
3. Кожен об’єкт належить лише одному кластеру. Однак у банківській сфері типово, що клієнт може належати до кількох поведінкових груп з різною ймовірністю (наприклад, молоді активні користувачі, які водночас мають профіль інвестора).
4. Невдалий початковий вибір центрів призводить до локальних мінімумів і поганої якості сегментації.

З огляду на ці недоліки KMeans не забезпечує достатньої гнучкості та точності при сегментації клієнтів банку.

Gaussian Mixture Models (GMM) є статистичним методом, який моделює дані як суміш кількох гауссових розподілів. На відміну від KMeans, GMM допускає різну форму кластерів, допускає різні дисперсії, виконує м’яку кластеризацію та базується на оцінюванні ймовірностей.

GMM описує розподіл даних як суміш:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k),$$

де π_k — ваги компонент (ймовірність вибору k-го кластера);

μ_k — вектор середніх значень;

Σ_k — коваріаційна матриця;

N — багатовимірний нормальний розподіл.

На відміну від KMeans, який використовує лише центр μ_k, модель GMM враховує форму кластера, орієнтацію та розмір.

Це особливо важливо у банківських даних, де спостереження розташовані нерівномірно. Параметри моделі оцінюються алгоритмом

Expectation-Maximization. Обчислення ймовірності, що спостереження x належить кластеру k:

$$\gamma_k(x\ \ ) = \frac{\pi_k N(x|\mu,\Sigma_k)}{\sum_{j=1}^{k} \pi_j N\left(x|\mu_j\Sigma_j\right)}$$

Оновлення параметрів на основі отриманих ймовірностей:

$$\mu_k = \frac{\Sigma_i \gamma_k(x_i) x_i}{\Sigma_i \gamma_k(x_i)}$$

$$\Sigma = \frac{\Sigma_i \gamma_k(x_i)(x_i - \mu_k)(x_i - \mu_k)^T}{\Sigma_i \gamma_k(x_i)}$$

$$\pi_k = \frac{1}{N}\sum_i \gamma_k(x_i)$$

Переваги GMM для банківської сегментації:

1) дозволяє визначати не лише сегмент, а й ступінь належності до кожного кластера;
2) підтримує еліптичні, витягнуті, перекриті групи;
3) кожен кластер може мати власну дисперсію й кореляції між ознаками;
4) можна аналізувати всю структуру ймовірностей, а не лише призначення кластерів;
5) у реальних банківських вибірках GMM відтворює природну структуру клієнтів значно краще за KMeans.

Перехід від KMeans до Gaussian Mixture Models дозволив підвищити точність кластеризації, отримати реалістичні сегменти, що краще відповідають поведінковим моделям клієнтів, врахувати різні масштаби, дисперсії та форми кластерів та формувати ймовірнісні профілі клієнтів, що є більш корисним для DSS-систем.

Отже, GMM є більш теоретично обґрунтованим та практично ефективним методом сегментації для задач банківської аналітики.

2.2 Архітектура інтелектуальної системи прогнозування клієнтського портфелю

Архітектура інтелектуальної системи прогнозування клієнтського портфелю побудована відповідно до принципів системного підходу та модульності, що забезпечує гнучкість, масштабованість і можливість подальшого розширення функціоналу. Система розглядається як сукупність взаємопов’язаних компонентів, кожен з яких виконує окрему функцію у загальному процесі аналізу та прогнозування клієнтської поведінки (рис. 2.1).

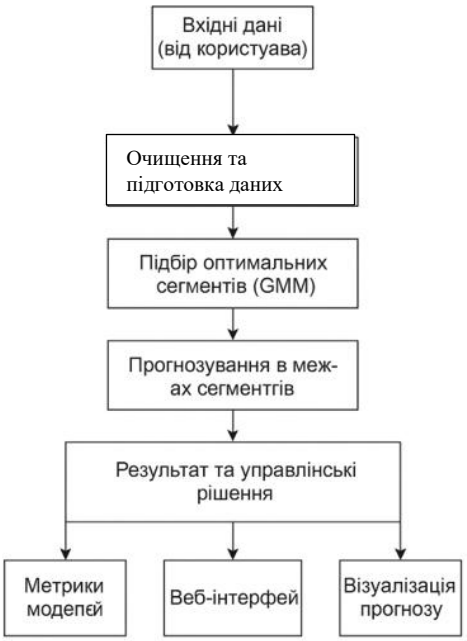


Рисунок 2.1 — Блок схема архітектури інтелектуільної системи

Основною метою архітектури є забезпечення наскрізного аналітичного конвеєра — від отримання вхідних даних до формування прогнозів і підсумкових управлінських результатів. Усі етапи обробки реалізовані послідовно та логічно пов’язані між собою, що дозволяє мінімізувати помилки, уникати витоку інформації та забезпечувати відтворюваність результатів.

Модуль вхідних даних забезпечує завантаження клієнтської інформації, яка використовується для подальшого аналізу та прогнозування. До таких

даних належать демографічні характеристики клієнтів, фінансові показники, транзакційна активність та поведінкові ознаки. Інформація може надходити з різних джерел, проте у межах системи вона приводиться до єдиного структурованого формату, що забезпечує узгодженість та коректність подальшої обробки.

Наступним етапом є модуль очищення та підготовки даних, який відповідає за попередню обробку інформації. На цьому етапі здійснюється усунення або імпутація пропущених значень, обробка аномалій та шумів у даних, нормалізація і масштабування числових ознак, а також кодування категоріальних змінних. Додатково виключаються ознаки, які можуть призводити до витоку інформації та спотворення результатів моделювання. Результатом роботи даного модуля є підготовлена матриця ознак, придатна для подальшої сегментації та прогнозного аналізу.

На основі очищених і підготовлених даних функціонує модуль сегментації клієнтів, у межах якого застосовується метод Gaussian Mixture Models. Даний модуль забезпечує підбір оптимальної кількості сегментів, оцінювання параметрів гауссових компонент та визначення ймовірностей належності кожного клієнта до відповідних сегментів. Сформовані сегменти відображають реальну структуру клієнтської бази та слугують основою для більш точного та диференційованого аналізу.

Подальшим етапом є модуль прогнозування для окремих сегментів, у якому для кожної сформованої групи клієнтів застосовуються відповідні моделі машинного навчання. Такий підхід дозволяє враховувати специфіку поведінки клієнтів у межах окремих сегментів та підвищити точність прогнозів. У межах цього модуля здійснюється оцінка ймовірності відкриття депозиту, залучення кредиту або виконання іншої цільової дії, що є важливою для банківських управлінських рішень.

Завершальним елементом архітектури є модуль формування результатів та інтерпретації, який агрегує результати сегментації та прогнозування. У цьому модулі формуються підсумкові таблиці, аналітичні звіти та індикатори,

що забезпечують підтримку управлінських рішень. Отримані результати подаються у зручному для інтерпретації вигляді та можуть бути використані для реалізації маркетингових, кредитних або стратегічних задач банку.

2.3 Формування та опис вибірки даних

Формування вибірки даних є одним з ключових етапів побудови інтелектуальної системи прогнозування клієнтського портфелю, оскільки саме структура та якість вихідних даних визначають коректність сегментації та точність прогнозних моделей. У межах даного дослідження використовуються клієнтські дані, що відображають демографічні, фінансові та поведінкові характеристики фізичних осіб, які взаємодіють з банком.

Вхідний набір даних містить інформацію про вік клієнта, рівень доходу, сімейний стан, зайнятість, а також показники фінансової активності, зокрема наявність кредитів і депозитів, обсяги транзакцій та частоту користування банківськими продуктами. Додатково використовуються поведінкові ознаки, які характеризують активність клієнтів у цифрових каналах обслуговування. Така структура даних дозволяє комплексно описати профіль клієнта та врахувати різні аспекти його взаємодії з банком.

Перед початком моделювання здійснюється первинний аналіз вибірки, який включає перевірку повноти даних, виявлення пропущених значень та аналіз розподілу основних ознак. На цьому етапі також проводиться контроль коректності даних і усунення очевидних помилок, що можуть виникати внаслідок технічних збоїв або особливостей збору інформації. Особливу увагу приділено забезпеченню анонімності клієнтів шляхом використання ідентифікаторів, що не дозволяють прямо ідентифікувати особу.

Для побудови прогнозних моделей формується цільова змінна, яка відображає настання певної події, зокрема відкриття депозиту або залучення кредиту. Формування цільових змінних здійснюється на основі історичних даних із дотриманням часової послідовності, що дозволяє уникнути витоку інформації та забезпечити реалістичність прогнозування. У разі наявності

кількох цільових показників для різних продуктів вони розглядаються окремо в межах відповідних моделей.

Після формування фінального набору ознак вибірка поділяється на навчальну та тестову частини. Навчальна вибірка використовується для навчання моделей сегментації та прогнозування, тоді як тестова — для оцінювання їх узагальнювальної здатності та перевірки якості результатів. Такий підхід забезпечує об’єктивну оцінку ефективності побудованої системи та її придатність для практичного використання.

Отже, сформована вибірка даних є репрезентативною та структурованою з урахуванням вимог до побудови інтелектуальних систем у банківській сфері. Вона створює надійну основу для реалізації сегментації клієнтів за допомогою Gaussian Mixture Models та подальшого прогнозування поведінки клієнтів у межах кожного сегмента.

2.4 Модуль попередньої обробки даних

Модуль попередньої обробки даних є обов’язковим елементом інтелектуальної системи прогнозування клієнтського портфелю, оскільки забезпечує підготовку вхідної інформації до подальшої сегментації та прогнозного моделювання. Банківські дані, як правило, характеризуються неоднорідністю, наявністю пропусків, аномальних значень і різними масштабами ознак, що унеможливлює безпосереднє застосування алгоритмів машинного навчання без попередньої обробки.

На першому етапі здійснюється аналіз повноти даних та обробка пропущених значень. Пропуски можуть виникати внаслідок відсутності взаємодії клієнта з окремими продуктами або технічних обмежень збору інформації. Для числових ознак застосовуються методи імпутації, що базуються на середніх або медіанних значеннях, тоді як для категоріальних змінних використовуються найбільш частотні значення. Такий підхід дозволяє зберегти структуру вибірки та мінімізувати втрату інформації.

Наступним етапом є виявлення та обробка аномальних значень, які можуть суттєво впливати на результати моделювання. Аномалії виникають через помилки введення даних або нетипову поведінку окремих клієнтів. Для їх обробки використовуються статистичні методи, що дозволяють обмежити вплив екстремальних значень на загальний розподіл ознак. Це сприяє підвищенню стабільності сегментації та прогнозних моделей.

Важливою складовою попередньої обробки є нормалізація та масштабування числових ознак, оскільки різні показники можуть мати суттєво відмінні діапазони значень. Масштабування забезпечує приведення ознак до порівнянного масштабу, що є критично важливим для алгоритмів, які використовують відстані або ймовірнісні оцінки, зокрема Gaussian Mixture Models. Завдяки цьому зменшується домінування окремих змінних у процесі моделювання.

Категоріальні змінні на даному етапі перетворюються у числовий формат, придатний для подальшого аналізу. Кодування дозволяє зберегти інформацію про належність клієнтів до певних категорій та інтегрувати ці ознаки у загальну матрицю даних. Така трансформація є необхідною умовою коректної роботи моделей машинного навчання.

Окрему увагу у модулі попередньої обробки приділено запобіганню витоку інформації. Для цього з вибірки вилучаються ознаки, які можуть прямо або опосередковано відображати цільову змінну або майбутні події, що не були відомі на момент прийняття рішення. Дотримання цього принципу забезпечує реалістичність оцінки моделей та їх придатність для практичного використання.

Результатом роботи модуля попередньої обробки є сформована та очищена матриця ознак, яка відображає релевантні характеристики клієнтів і може бути безпосередньо використана для сегментації за допомогою Gaussian Mixture Models та подальшого прогнозування поведінки клієнтів. Таким чином, попередня обробка даних виступає фундаментом усієї інтелектуальної системи та визначає якість кінцевих результатів аналізу.

2.5 Валідація, інтерпретація та профілювання клієнтських сегментів

Після побудови сегментаційної моделі важливим етапом є оцінювання якості отриманих сегментів та їх інтерпретація з точки зору практичного застосування у банківській сфері. Навіть математично коректна кластеризація не має прикладної цінності, якщо сформовані сегменти не є стабільними, відокремленими та зрозумілими для подальшого використання у системі підтримки прийняття рішень.

Одним з ключових завдань на даному етапі є визначення оптимальної кількості сегментів. Для цього у роботі застосовуються метод ліктя та коефіцієнт силуету, які дозволяють кількісно оцінити якість сегментації та обґрунтувати вибір кількості кластерів.

Метод ліктя базується на аналізі залежності внутрі кластерної дисперсії від кількості сегментів. Для кожного значення кількості кластерів K обчислюється сумарна квадратична помилка, яка визначається як:

$$SSE(K) = \sum_{k=1}^{K} \sum_{x \in c_k} \|x - \mu_k\|^2,$$

де c_k — множина об'єктів k-го сегмента;

μ_k — центр відповідного сегмента.

Ідея методу полягає у тому, що зі збільшенням кількості сегментів значення SSE монотонно зменшується, однак після певного значення K темп зменшення істотно сповільнюється. Точка, у якій спостерігається злам кривої, інтерпретується як оптимальна кількість сегментів, оскільки подальше збільшення K не призводить до суттєвого покращення якості кластеризації.

Однак метод ліктя має певну суб’єктивність, оскільки точка зламу не завжди є чітко вираженою. Тому для додаткової валідації сегментації у роботі використовується коефіцієнт силуету, який дозволяє оцінити ступінь відокремленості сегментів та їх внутрішню однорідність.

Коефіцієнт силуету для окремого об’єкта і визначається як:

$$s(\mathrm{i}) = \frac{b(\mathrm{i}) - a(\mathrm{i})}{max\{a(\mathrm{i}), b(\mathrm{i})\}},$$

де $a(\mathrm{i})$ є середньою відстанню між об'єктом і та іншими об'єктами того ж сегмента;

$b(\mathrm{i})$ є мінімальною середньою відстанню між об'єктом і та об'єктами найближчого альтернативного сегмента.

Значення коефіцієнта силуету знаходиться у межах від -1 до 1. Чим ближче значення до 1, тим краще об'єкт відповідає своєму сегменту і тим чіткіше відокремлені кластери.

Для оцінювання якості сегментації використовується середнє значення коефіцієнта силуету по всій вибірці:

$$s = \frac{1}{N} \sum_{i=1}^{N} s\,(i),$$

де N — кількість об'єктів. Оптимальною вважається така кількість сегментів, для якої середнє значення коефіцієнта силуету є максимальним.

Після вибору кількості сегментів здійснюється інтерпретація та профілювання отриманих груп. Для кожного сегмента аналізуються середні значення числових ознак та характерні поведінкові патерни, що дозволяє сформувати узагальнений портрет клієнта. Такий підхід дає змогу пов'язати результати сегментації з реальними бізнес-сценаріями банку, зокрема маркетинговими кампаніями, кредитними рішеннями та управлінням ризиками.

Отримані сегменти інтегруються у подальший етап прогнозування, де належність клієнта до певного сегмента та відповідні ймовірності використовуються як додаткові інформативні ознаки. Таким чином, сегментація виконує не лише описову, а й функціональну роль у загальній архітектурі інтелектуальної системи підтримки прийняття рішень.

2.6 Формалізація моделей прогнозування поведінки клієнтів

Після формування та валідації клієнтських сегментів наступним етапом побудови інтелектуальної системи є прогнозування поведінки клієнтів у

межах кожного сегмента. У роботі прогнозування розглядається як задача бінарної класифікації, де цільовою змінною є настання або ненастання певної події, зокрема відкриття депозиту або залучення кредиту. Сегментно-орієнтований підхід дозволяє враховувати відмінності у поведінкових патернах клієнтів та підвищити точність прогнозів.

Базовою моделлю прогнозування використовується логістична регресія, яка широко застосовується у банківській сфері завдяки своїй простоті та високій інтерпретованості. Логістична регресія моделює ймовірність настання події як сигмоїдну функцію лінійної комбінації вхідних ознак і має вигляд:

$$P(y = 1|x) = \frac{1}{1+e^{-\left(\beta_0+\sum_j^p \beta_j x_j\right)}}$$

Дана модель дозволяє оцінити вплив кожної ознаки на ймовірність цільової події, що є важливим для аналізу факторів поведінки клієнтів та прийняття управлінських рішень.

Для підвищення якості прогнозування у роботі застосовується ансамблевий метод Random Forest, який ґрунтується на побудові множини дерев рішень та агрегуванні їх результатів. Прогноз для окремого клієнта визначається як усереднене значення прогнозів усіх дерев ансамблю:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x),$$

де $h_t(x)$ — прогноз t-го дерева, а T — кількість дерев у моделі. Використання Random Forest дозволяє зменшити ризик переобучення та враховувати складні нелінійні залежності між ознаками.

Ще одним методом прогнозування є Gradient Boosting, який реалізує послідовну побудову ансамблю моделей, де кожна наступна модель коригує помилки попередніх. Загальний прогноз Gradient Boosting формується як зважена сума окремих базових моделей:

$$\hat{y}(x) = \sum_{m=1}^{M} \gamma_m h_m(x),$$

де $h_m(x)$ — базова модель, а γ_m — її вага.

Такий підхід дозволяє досягти високої точності прогнозування за рахунок поетапного уточнення результатів.

У межах інтелектуальної системи прогнозування моделі навчаються окремо для кожного клієнтського сегмента, що дозволяє адаптувати параметри моделей до специфіки поведінки клієнтів у межах конкретних груп. Отримані ймовірнісні оцінки використовуються для формування персоналізованих рекомендацій та підтримки управлінських рішень у банківській діяльності.

2.7 Оцінювання якості прогнозних моделей

Оцінювання якості прогнозних моделей є обов'язковим етапом побудови інтелектуальної системи прогнозування клієнтського портфелю, оскільки дозволяє визначити надійність, стабільність та практичну придатність отриманих результатів. У банківській сфері особливе значення має не лише загальна точність прогнозів, а й здатність моделі коректно ідентифікувати клієнтів, схильних до цільової дії, зокрема відкриття депозиту або залучення кредиту.

Для оцінювання якості моделей у роботі використовується матриця помилок, яка відображає кількість правильних і помилкових класифікацій. На її основі розраховуються основні метрики бінарної класифікації, що дозволяють оцінити модель з різних позицій. Точність класифікації визначається як частка правильних прогнозів від загальної кількості спостережень і обчислюється за формулою:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN},$$

де TP та TN відповідно кількість правильно передбачених позитивних і негативних випадків, а FP та FN — помилкові класифікації. Однак у задачах банківської аналітики дана метрика не завжди є достатньо інформативною, особливо у випадку незбалансованих класів.

Для детальнішого аналізу якості моделей використовується метрика точності позитивного прогнозу (Precision), яка показує, яка частка передбачених позитивних випадків є дійсно коректною:

$$Precision = \frac{TP}{TP + FP}$$

Ця метрика є важливою при оцінюванні ефективності маркетингових кампаній, оскільки дозволяє зменшити кількість хибних рекомендацій клієнтам.

Повнота (Recall) характеризує здатність моделі знаходити всі фактичні позитивні випадки та визначається як:

$$F1 = 2 * \frac{Precision * Recacll}{Precision + Recall}$$

Окрім зазначених метрик, у роботі застосовується ROC-крива, яка відображає залежність між часткою правильно ідентифікованих позитивних випадків та часткою хибнопозитивних прогнозів при різних порогових значеннях. Інтегральним показником якості моделі є площа під ROC-кривою (ROC-AUC), яка характеризує здатність моделі розрізняти класи незалежно від вибору порогу. Чим ближче значення ROC-AUC до 1, тим вищою є якість прогнозування.

З метою підвищення об’єктивності оцінювання всі метрики розраховуються на тестовій вибірці, яка не використовувалася під час навчання моделей. Такий підхід дозволяє оцінити узагальнювальну здатність моделей та їх готовність до практичного застосування.

Таким чином, комплексне використання декількох метрик оцінювання забезпечує багатосторонній аналіз якості прогнозних моделей і дозволяє обґрунтовано обрати оптимальні моделі для інтеграції в інтелектуальну систему підтримки прийняття рішень у банківській сфері.

2.8 Інтеграція сегментації та прогнозування в систему підтримки прийняття рішень

Інтеграція результатів сегментації та прогнозного моделювання є завершальним етапом побудови інтелектуальної системи прогнозування клієнтського портфелю. На цьому етапі аналітичні результати перетворюються на інструмент підтримки прийняття управлінських рішень, що дозволяє використовувати їх у практичній діяльності банку.

Результати сегментації клієнтів, отримані за допомогою Gaussian Mixture Models, використовуються як основа для формування сегментно-орієнтованих прогнозів. Для кожного клієнта система визначає ймовірність належності до відповідних сегментів, а також узагальнений сегментний профіль. Ця інформація інтегрується у модулі прогнозування як додатковий контекст, що дозволяє підвищити точність оцінювання ймовірності цільових дій.

Прогнозні моделі формують імовірнісні оцінки для кожного клієнта з урахуванням його сегментної приналежності. Таким чином, система забезпечує не лише загальний прогноз, а й сегментно-адаптований результат, який відображає реальні поведінкові особливості клієнтів. Це дозволяє уникнути універсального підходу та реалізувати більш точні персоналізовані сценарії взаємодії.

У межах системи підтримки прийняття рішень результати прогнозування агрегуються у зручні аналітичні представлення, що включають ранжування клієнтів за рівнем ймовірності цільової дії, формування списків пріоритетних клієнтів та зведених показників ефективності. Такі результати можуть бути використані для планування маркетингових кампаній, оптимізації кредитних пропозицій або управління ризиками.

Важливою перевагою інтегрованого підходу є можливість пояснення результатів прогнозування. Завдяки поєднанню сегментації та моделей прогнозування система надає не лише числову оцінку ймовірності, а й контекст її формування, зокрема сегмент, до якого належить клієнт, та характерні для нього ознаки. Це підвищує довіру до результатів системи та спрощує їх використання у прийнятті рішень.

Таким чином, інтеграція сегментації та прогнозного моделювання дозволяє сформувати цілісну інтелектуальну систему підтримки прийняття рішень, яка забезпечує перехід від аналізу даних до практичних управлінських дій. Такий підхід підвищує ефективність роботи банку, сприяє персоналізації сервісу та оптимізації використання ресурсів.

2.9 Висновки до розділу 2

У другому розділі було розглянуто системний підхід до побудови інструментарію прогнозування клієнтського портфелю банку. Запропонована архітектура інтелектуальної системи охоплює повний цикл аналітичної обробки даних — від формування та підготовки вибірки до сегментації, прогнозування та інтеграції результатів у систему підтримки прийняття рішень.

У ході розробки інструментарію обґрунтовано вибір методів сегментації та прогнозного моделювання. Показано, що застосування Gaussian Mixture Models дозволяє отримати гнучку та імовірнісну сегментацію клієнтів, яка краще відображає реальну структуру клієнтської бази порівняно з класичними центричними методами. Додаткове використання методів ліктя та коефіцієнта силуету забезпечує обґрунтований вибір кількості сегментів і підвищує якість кластеризації.

Також у розділі було формалізовано процес попередньої обробки даних, що включає очищення, трансформацію та запобігання витоку інформації. Наголошено на важливості якісної підготовки даних як фундаменту для коректної роботи моделей машинного навчання.

Окрему увагу приділено формалізації моделей прогнозування поведінки клієнтів. Показано доцільність використання як інтерпретованих статистичних моделей, так і ансамблевих методів машинного навчання, що дозволяє досягти балансу між точністю прогнозів та їх пояснюваністю. Оцінювання якості моделей здійснюється на основі комплексу метрик, що дозволяє всебічно оцінити їх ефективність.

Завершальним етапом розділу стала інтеграція результатів сегментації та прогнозування в систему підтримки прийняття рішень, що забезпечує перехід від аналітичних розрахунків до практичних управлінських дій. Таким чином, у другому розділі сформовано цілісний інструментарій, який створює основу для експериментального дослідження та аналізу результатів, представлених у наступному розділі роботи.

РОЗДІЛ 3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНТ ЗА ТЕМОЮ МД

3.1 Проблема формування клієнтського портфелю

Формування якісного клієнтського портфелю є одним із ключових завдань сучасних банків, адже саме від нього залежить рівень прибутковості, конкурентоспроможності та фінансової стабільності установи. Для побудови адекватної моделі портфелю необхідно визначити перелік параметрів, які найбільш суттєво впливають на поведінку клієнтів. Серед них — соціально-демографічні характеристики (вік, сімейний стан, рівень доходів), параметри банківських продуктів (категорія картки, кредитний ліміт, наявність депозитів), а також динамічні показники активності клієнтів: частота транзакцій, використання кредитних лімітів, взаємодія з банком у різні періоди часу.

У процесі роботи з реальними даними банк стикається з кількома проблемами. По-перше, через великі обсяги інформації необхідно здійснювати складні запити до баз даних (SQL-запити або аналітичні пайплайни в Python), що потребує не лише значних обчислювальних ресурсів, а й високої кваліфікації аналітиків. По-друге, критично важливим є дотримання принципів безпеки та конфіденційності даних. Усі унікальні ідентифікатори клієнтів та чутлива інформація мають бути захешованими або анонімізованими для подальшої безпечної роботи з ними.

У реалізованому пайплайні для цього передбачено створення додаткової колонки, де унікальний клієнтський ідентифікатор відображається у захищеному вигляді: відкритими залишаються лише перші три цифри та останні чотири, а проміжні символи замінюються на *. Такий підхід дозволяє зберегти можливість ідентифікації клієнта у рамках дослідження без розкриття його персональних даних.

Ще однією суттєвою проблемою є врахування ризиків, пов’язаних зі зміною поведінки клієнтів під впливом зовнішніх факторів: економічних криз,

коливань доходів населення, зміни ринкових умов чи пропозицій конкурентів. Тому банку недостатньо лише зафіксувати поточний стан клієнтської бази — необхідно також прогнозувати її динаміку. Для цього у практичній частині роботи застосовуються сучасні методи аналізу даних та алгоритми машинного навчання, що дають змогу виявляти приховані закономірності, оцінювати ризики неплатоспроможності та прогнозувати ймовірність відтоку клієнтів.

Таким чином, головна проблема формування клієнтського портфелю полягає у необхідності поєднання трьох ключових аспектів: якісної аналітики великих даних, дотримання вимог безпеки й конфіденційності та застосування інноваційних прогнозних моделей. Її вирішення сприятиме підвищенню ефективності управління ресурсами банку й забезпеченню його сталого розвитку у довгостроковій перспективі.

3.2 Опис вхідних даних та їх обробка

Вхідними даними для дослідження є CSV-файл, що містить анкетні та транзакційні характеристики клієнтів банку. База формується з відкритих джерел (наприклад, Kaggle чи GitHub), що дозволяє працювати з максимально реалістичною вибіркою без ризику використання чутливої інформації. До основних змінних відносяться: унікальний ідентифікатор клієнта (CLIENTNUM), вік, стать, кількість утриманців, рівень освіти, сімейний стан, категорія картки, кредитний ліміт, кількість та сума транзакцій, інші показники, що відображають фінансову активність клієнтів. Таким чином, дані охоплюють демографічні, соціальні та фінансові аспекти, необхідні для побудови прогнозних моделей.

Завантаження інформації реалізується за допомогою бібліотеки pandas через команду:

apart_df = pd.read_csv(URL)

Для забезпечення стійкості до помилок під час імпорту передбачено обробку винятків: FileNotFoundError використовується, якщо файл відсутній,

pd.errors.EmptyDataError — якщо він не містить даних, pd.errors.ParserError — у разі пошкодженої структури, а загальний блок Exception відловлює непередбачені ситуації. Це дозволяє уникнути зупинки роботи пайплайну навіть при збої.

Після зчитування даних виконується первинна перевірка структури за допомогою, що дає змогу візуально оцінити правильність формату та типів колонок (рис. 3.1). Для моделювання умов, близьких до реальних, до бази додається згенерований унікальний десятизначний ідентифікатор: apart_df[CLIENT_ID] = np.random.randint(10**9, 10**10, size=len(apart_df)) Це створює аналог внутрішнього коду клієнта у банківській системі.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown

Рисунок 3.1 — Приклад вивантаження бази

Оскільки робота з персональними даними потребує дотримання принципів конфіденційності, у пайплайні реалізовано хешування ідентифікаторів. Для цього використовується функція:

```
def mask_client_id(client_id):
    client_id_str = str(client_id)
    return client_id_str[:3] + **(len(client_id_str)-7) + client_id_str[-4:]
```

Вона формує нову колонку, де відображаються лише перші три та останні чотири символи, а проміжні замінюються на *. Таким чином, наприклад, значення 1234567890 перетворюється на 123***7890 (рис. 3.2). Це дозволяє зберегти унікальність клієнта у межах вибірки та водночас захистити його персональні дані.

	client_id	client_id_masked	client_hash
0	7965604437	796****4437	d7989b2a9584c28c61d2042479bd6c6eabbabfda37bae3...
1	4949905957	494****5957	c4d876911be02289178417a5b560f9b6f5815f393fc12a...
2	8727381279	872****1279	d3081848cf8df606bfd078b603d14d9cc94205ad24d106...
3	7276312261	727****2261	8619a3bdc2542f23f24d04937488d42e5d32c040857402...
4	1847596130	184****6130	767def8e74957e34e17d06c75cdc01960d4201c7d99ddf...

Рисунок 3.2 — Приклад хешування персонального ідентифікатора клієнта

Додатково передбачені прості перевірки правильності хешування: виведення перших кількох значень, контроль довжини ідентифікатора та зіставлення вихідних і оброблених даних. Це підтверджує, що унікальність клієнтів не порушена, а конфіденційність надійно захищена.

На фінальному етапі формується матриця ознак для моделі машинного навчання. До неї входять демографічні, поведінкові та фінансові характеристики, а цільовою змінною є наявність або відсутність прогнозованої взаємодії з банком (отримання кредиту чи депозиту). Таким чином, реалізовано повний цикл підготовки: від завантаження даних із перевірками, через їх анонімізацію до готовності використання у прогнозних алгоритмах.

3.3 Формування набору ознак та моделювання

На цьому етапі реалізовано формування набору ознак для моделювання та побудову препроцесорів, які забезпечують правильну обробку як числових, так і категоріальних змінних. Головне завдання — уникнути витоків інформації та підготувати дані у форматі, придатному для подальшого навчання моделей.

Спочатку з датафрейму виключаються службові стовпці (client_id, client_id_masked, client_hash). Також жорстко вилучається ознака duration, оскільки вона стає відомою лише після завершення контакту з клієнтом і призвела б до витоку інформації при навчанні.

Далі формуються два піднабори ознак. Для задачі прогнозування взяття депозиту використовуються всі відібрані колонки. Для задачі прогнозування

взяття кредиту додатково виключаються поля loan та housing, оскільки вони безпосередньо відображають поточний кредитний стан клієнта і можуть викликати витік.

Після цього ознаки розділяються на числові та категоріальні. Критерієм слугує тип даних: якщо стовпець має тип object, він відноситься до категоріальних, усі інші — до числових.

Для кожного типу ознак створюється окремий конвеєр обробки. Числові змінні проходять медіанне імпутування пропусків та стандартизацію, щоб привести їх до єдиного масштабу. Категоріальні змінні імпутуються за найбільш поширеним значенням, а потім перетворюються через one-hot кодування. При цьому використовується параметр handle_unknown=ignore, який дозволяє моделі коректно працювати з новими категоріями, що можуть з’явитися у майбутніх даних, а також max_categories=80, щоб уникнути надмірного розростання кількості ознак.

Обидві гілки поєднуються у ColumnTransformer, що дозволяє створити єдиний препроцесор. У підсумку отримано два окремих препроцесори: prep_B для прогнозування депозитів і prep_C для прогнозування кредитів.

Такий підхід дозволяє:

1) уникнути витоків інформації (виключення duration, loan, housing);
2) забезпечити стабільну роботу з пропущеними даними;
3) підготувати числові ознаки до використання у моделях, чутливих до масштабу;
4) конвертувати категоріальні змінні у зручний числовий формат без втрати інформації.

У результаті система формує чистий та збалансований набір ознак, готовий до подальшого моделювання, при цьому всі перетворення інтегровані у конвеєр, що гарантує відтворюваність експериментів.

3.4 Сегментація клієнтської бази

Сегментація клієнтів є одним з ключових етапів управління клієнтським портфелем банку, оскільки дозволяє виявити групи клієнтів із подібними характеристиками та поведінковими патернами. Це створює підґрунтя для формування цільових маркетингових стратегій, оптимізації пропозицій банківських продуктів і підвищення ефективності управління ризиками. У даному дослідженні для сегментації клієнтів використано модель Gaussian Mixture Models (GMM), яка є імовірнісним методом кластеризації та добре підходить для аналізу складних багатовимірних даних.

На відміну від центричних методів кластеризації, GMM не виконує жорсткого поділу об’єктів на кластери на основі мінімізації відстані до центроїда. Замість цього модель припускає, що розподіл клієнтів у просторі ознак може бути поданий як суміш кількох гауссових розподілів. Кожен сегмент описується власними параметрами розподілу, зокрема середнім значенням та коваріаційною матрицею, а кожному клієнту відповідає ймовірність належності до кожного сегмента. Такий підхід дозволяє враховувати перекриття між сегментами та більш точно відображати реальну структуру клієнтської бази.

У контексті банківських клієнтів це означає, що система формує сегменти не як жорстко відокремлені групи, а як імовірнісні профілі. Наприклад, окремий клієнт може з високою ймовірністю належати до сегмента молодих активних користувачів цифрових каналів, водночас частково відповідати характеристикам сегмента клієнтів із потенціалом до кредитування. Такий підхід є більш гнучким і відповідає реальним сценаріям поведінки клієнтів у банківській сфері.

Для автоматизації процесу сегментації у роботі реалізовано окрему функцію fit_gmm_and_profile, яка забезпечує повний цикл побудови та аналізу сегментів. На вхід функції подається датафрейм із клієнтськими даними та перелік ознак, що використовуються для сегментації. Перед побудовою моделі

застосовується попередньо сформований препроцесор, який виконує стандартизацію числових змінних та кодування категоріальних ознак, забезпечуючи коректність подальшого аналізу.

Далі здійснюється перебір кількості сегментів у заданому діапазоні значень з метою визначення оптимальної структури кластеризації. Для кожного варіанта кількості сегментів обчислюються показники якості кластеризації, зокрема коефіцієнт силуету, який оцінює ступінь відокремленості сегментів та внутрішню узгодженість об’єктів у межах кожної групи. Оптимальна кількість сегментів обирається на основі максимального значення цієї метрики, що дозволяє забезпечити баланс між деталізацією сегментації та її стабільністю.

Після визначення оптимальної кількості сегментів виконується остаточне навчання моделі GMM, за результатами якого для кожного клієнта обчислюються ймовірності належності до відповідних сегментів. На основі цих результатів до набору даних додається нова змінна, що відображає домінуючий сегмент клієнта, а також формується сегментний профіль.

Профілювання сегментів здійснюється шляхом обчислення середніх значень числових ознак та найбільш характерних значень категоріальних змінних у межах кожного сегмента. Це дозволяє сформувати узагальнений опис клієнтів кожної групи та забезпечити інтерпретованість результатів сегментації з точки зору бізнес-логіки банку.

Для візуального аналізу результатів сегментації додатково застосовується метод зниження розмірності Principal Component Analysis, який дозволяє відобразити клієнтів та сформовані сегменти у двовимірному просторі. Така візуалізація спрощує аналіз структури сегментів і наочно демонструє ступінь їх відокремленості. Вкажемо фрагмент реалізації функції (скорочено):

```
def fit_gmm_and_profile(df, feature_cols, preprocessor,
k_range=range(2,11), random_state=42):
    X = df[feature_cols]
```

```
    Xs = preprocessor.fit_transform(X)
    sse, sil = [], []
    for k in k_range:
        km = Gmm(n_clusters=k, random_state=random_state, n_init=auto)
        labels = km.fit_predict(Xs)
        sse.append(km.inertia_)
        sil.append(silhouette_score(Xs, labels))
    best_k = k_range[np.argmax(sil)]
    gmm = Gmm(n_clusters=best_k, random_state=random_state, n_init=auto)
    clusters = gmm.fit_predict(Xs)
    seg_df = df.copy()
    seg_df[cluster] = clusters
    return seg_df, prof_num, prof_cat, X_pca, clusters, best_k, (sse, sil)
```

Визначення оптимального числа кластерів є критичним завданням. Використано два підходи.

1. Метод ліктя (Elbow method) — ґрунтується на аналізі графіка залежності значення SSE від числа кластерів. Зазвичай оптимальним вважається значення k, у якому спостерігається злам кривої (рис. 3.3).
2. Коефіцієнт силуету — більш формальна метрика, яка дозволяє обрати таке число кластерів, при якому внутрішня щільність і зовнішня відокремленість груп максимальні (рис. 3.4).

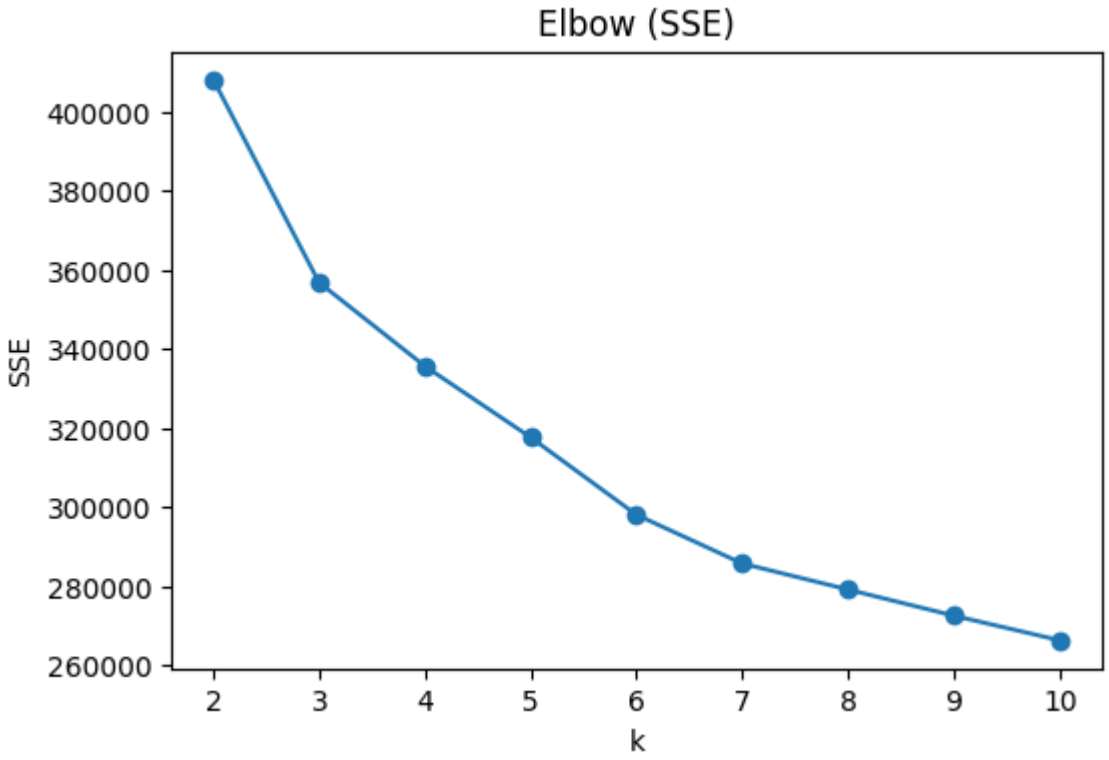


Рисунок 3.3 — Зображення залежності суми квадратів відстаней (SSE) від кількості кластерів. Точка зламу графіка відповідає оптимальному значенню k.

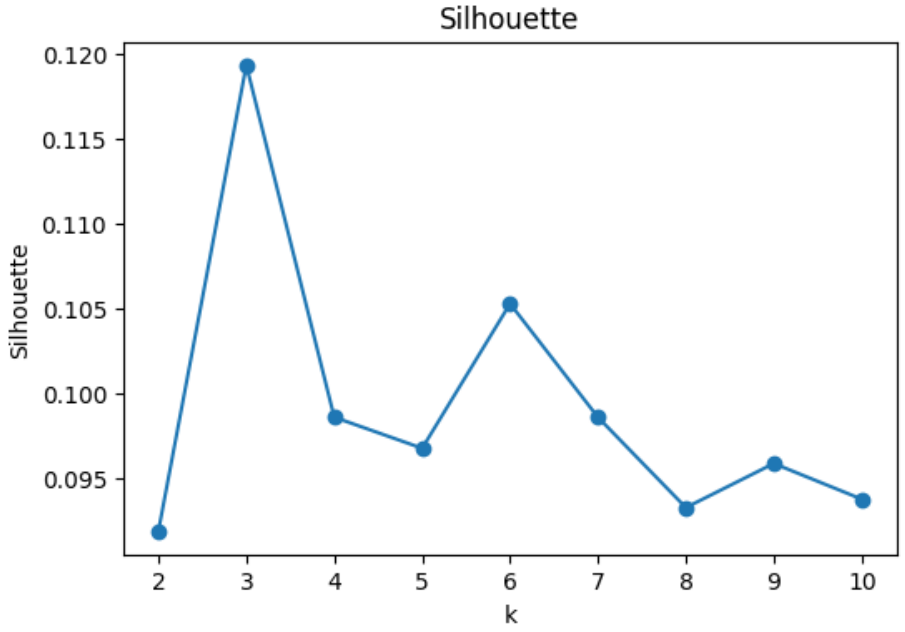


Рисунок 3.4 — Графік залежності коефіцієнта силуету від кількості кластерів.

У результаті оптимальним виявилось значення k= best_k = 3, яке дало найкращі значення коефіцієнта силуету. Для наочності було застосовано метод головних компонент (PCA), що дозволив знизити розмірність даних до двох

компонент. Це дає можливість відобразити клієнтів на площині та візуально оцінити межі кластерів (рис. 3.5).

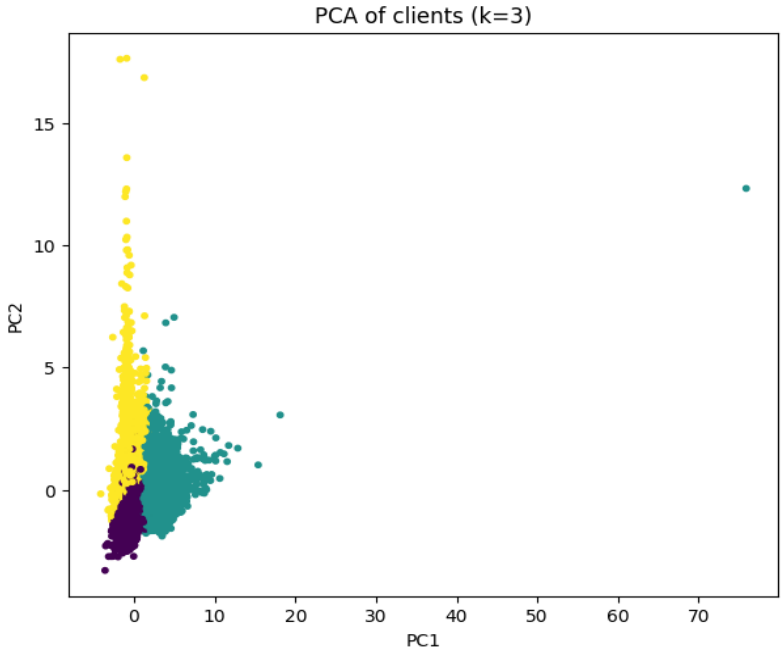


Рисунок 3.5 — Візуалізація клієнтів у просторі двох головних компонент.

Після завершення кластеризації кожен кластер було проаналізовано з точки зору його середніх характеристик. Для числових змінних розраховано середні значення, для категоріальних — визначено найбільш поширені категорії. Це дозволяє скласти портрет кожного сегмента клієнтів.

У результаті отримано дві таблиці:

1) середні значення числових змінних за кластерами (наприклад, середній вік, середній кредитний ліміт, середня кількість транзакцій);
2) найчастіші категоріальні характеристики за кластерами (тип картки, сімейний стан, рівень освіти тощо).

Отримані кластери можна інтерпретувати як окремі сегменти клієнтів із власними особливостями. Охарактеризуємо далі всі три сегменти.

Характеристика сегменту 0:

1) вік молодший за середній (переважно 20–35 років);
2) рівень доходу та кредитний ліміт невисокий, середні залишки на рахунках;

3) транзакційна активність середня, але зростаюча — клієнти активно використовують картки для щоденних витрат;
4) категорії карток — переважно стандартні або базові (Classic);
5) сімейний стан — більшість неодружені або молоді сім'ї з невеликою кількістю утриманців.

Цей сегмент можна визначити як молоді клієнти з потенціалом зростання. Вони ще не є головними споживачами кредитних продуктів, але демонструють високий інтерес до повсякденних банківських послуг. Для них доцільно пропонувати програми лояльності, молодіжні пакети та продукти для накопичення.

Характеристика сегменту 1:

1) вік середній (35–50 років);
2) фінансові параметри — стабільні доходи, вищий за середній кредитний ліміт, помірні залишки на рахунках;
3) транзакційна активність висока — часто користуються як кредитними, так і дебетовими картками;
4) категорії карток Gold або Platinum;
5) сімейний стан переважно одружені, мають утриманців.

Цей сегмент — основна робоча аудиторія банку, яка активно користується продуктами й формує основну частину прибутку. Вони зацікавлені у кредитуванні (іпотека, споживчі кредити), але також часто відкривають депозити. Це найбільш прибутковий сегмент, для якого актуальні персоналізовані пропозиції та преміальні пакети послуг.

Характеристика сегменту 2:

1) вік старший (50+ років);
2) фінансові параметри високі залишки на рахунках, але невисока потреба у кредитах;
3) транзакційна активність низька — рідкісні операції, переважно пов'язані зі зняттям готівки або депозитними операціями;
4) категорії карток Standard або Classic, рідше преміальні;

5) сімейний стан стабільний, велика частка клієнтів без утриманців (дорослі діти).

Цей сегмент — консервативні клієнти, які переважно орієнтовані на заощадження й депозити. Вони не є цільовою групою для активних кредитних продуктів, проте формують значну базу депозитів, що є важливим джерелом ліквідності для банку. Для них актуальні стабільні депозитні програми, зручність обслуговування та збереження довіри до банку.

Подібне профілювання дозволяє банку краще зрозуміти структуру клієнтської бази, ідентифікувати цінні сегменти, а також визначити групи з підвищеним ризиком відтоку або неплатоспроможності.

3.5 Побудова моделей

У цьому підрозділі розглядається етап побудови та оцінки моделей машинного навчання, спрямованих на прогнозування поведінки клієнтів банку. Ключовою особливістю є використання реалістичного підходу, коли з набору ознак усунуто змінні, які можуть призводити до витоку інформації (зокрема, поле duration). Це дозволяє змоделювати реальні умови прийняття рішень, коли банк не має доступу до післяподієвих змінних у момент ухвалення рішення.

Процес моделювання складається з кількох етапів. Формування вибірок для двох основних задач:

1) задача 1 — прогнозування того, чи клієнт відкриє депозит (term_deposit);
2) задача 2 — прогнозування ймовірності того, що клієнт візьме кредит (will_take_credit).

Побудова конвеєра обробки даних, що включає:

1) препроцесинг числових і категоріальних змінних (описаний у попередньому розділі);
2) вибір моделі класифікації.

Навчання кількох моделей та їх порівняння за ключовими метриками. Для оцінки якості прогнозів застосовано три класифікатори:

1) logistic regression — базова лінійна модель, що дозволяє інтерпретувати внесок окремих ознак;
2) random forest — ансамблевий метод на основі великої кількості дерев рішень, здатний працювати з нелінійними залежностями;
3) gradient boosting — алгоритм послідовного навчання слабких моделей, який зазвичай показує високу точність на табличних даних.

Вибір цих моделей пояснюється тим, що вони представляють три різні підходи до класифікації: лінійний, ансамблевий та бустинговий.

Для зручності було створено функцію evaluate_models, яка:

1) розділяє дані на тренувальну та тестову вибірки (75% / 25%) з використанням стратифікації, щоб зберегти баланс класів;
2) навчає кожну модель у складі конвеєра (preprocessor + model);
3) робить прогнози та оцінює їх за низкою метрик.

Серед використаних метрик:

1) accuracy — частка правильних відповідей;
2) F1-score — збалансована міра точності й повноти;
3) precision (точність) та Recall (повнота);
4) ROC-AUC — площа під кривою ROC, що характеризує якість ранжування;
5) confusion Matrix — матриця помилок, що дозволяє оцінити кількість правильних і неправильних класифікацій для кожного класу.

Для спрощення повторного використання створено функцію run_task, яка викликає оцінку моделей для конкретної задачі, використовуючи відповідні ознаки та препроцесор:

1) res_B = run_task(B_term_deposit, y_term, features_B, prep_B);
2) res_C = run_task(C_will_take_credit, will_take_credit, features_C, prep_C).

У результаті формується підсумкова таблиця з усіма метриками для обох задач, що дозволяє порівняти якість моделей (рис. 3.6).

	task	model	accuracy	f1	precision	recall	roc_auc	y_test_mean
0	B_term_deposit (realistic_no_duration)	LogReg	0.899343	0.346926	0.436447	0.287879	0.721723	0.092871
1	B_term_deposit (realistic_no_duration)	RandomForest	0.910413	0.145031	0.637795	0.081818	0.737273	0.092871
2	B_term_deposit (realistic_no_duration)	GradBoost	0.911257	0.185886	0.627907	0.109091	0.747294	0.092871
3	C_will_take_credit (realistic_no_duration)	LogReg	0.772045	0.821874	0.825140	0.818633	0.827045	0.642402
4	C_will_take_credit (realistic_no_duration)	RandomForest	0.785272	0.839896	0.806014	0.876752	0.840031	0.642402
5	C_will_take_credit (realistic_no_duration)	GradBoost	0.788368	0.843312	0.804106	0.886536	0.849219	0.642402

Рисунок 3.6 — Таблиця порівняння результатів якості моделей

Результати показують, що:

1) логістична регресія слугує хорошою базовою моделлю, проте її точність обмежена у випадках, коли залежності між ознаками є нелінійними;
2) random forest зазвичай демонструє вищу точність і кращі значення F1, оскільки враховує складні взаємозв’язки між ознаками;
3) gradient boosting часто забезпечує найкращий результат за метрикою ROC-AUC, що свідчить про високу здатність до розрізнення класів.

Результати моделювання мають безпосереднє практичне застосування.

1. Високі значення Recall у задачі прогнозування взяття депозиту свідчать про здатність моделі виявляти клієнтів, які з більшою ймовірністю відкриють депозит, що дає можливість банку спрямувати маркетингові зусилля саме на ці групи.
2. У задачі прогнозування взяття кредиту кращі значення Precision дозволяють точніше визначати потенційних позичальників, що знижує ризики відмови від кредитних продуктів і підвищує ефективність кредитної політики.

Таким чином, обидві задачі доповнюють одна одну: одна орієнтована на залучення нових депозитів, інша — на прогнозування кредитного попиту. Використання ансамблевих моделей дає змогу банку підвищити точність прогнозування та приймати більш обґрунтовані рішення.

3.6 Результати сегментації: прогноз депозитів і кредитів

Завершальним етапом побудованого пайплайну є інтеграція моделей прогнозування з результатами сегментації клієнтів. Це дозволяє не лише отримати прогноз імовірності відкриття депозиту чи залучення кредиту на рівні кожного клієнта, але й здійснити узагальнення на рівні сегментів. Таким чином банк отримує можливість не тільки ідентифікувати окремих клієнтів, які з великою ймовірністю виявлять певну фінансову поведінку, а й аналізувати, які групи клієнтів найбільш схильні до депозитних або кредитних продуктів.

Для формування кінцевих прогнозів було обрано Random Forest Classifier як основну модель. Вона була навчена на повному обсязі доступних даних, що дозволило підвищити стабільність прогнозів. Оскільки попередні етапи показали високу ефективність цього алгоритму, його використання на завершальному кроці є виправданим.

Для кожної з задач побудовано окремий конвеєр:

1) депозити — pipe_B = Pipeline(steps=[(prep, prep_B), (model, rf_B)]);
2) кредити — pipe_C = Pipeline(steps=[(prep, prep_C), (model, rf_C)]).

Кожна з моделей повертає два типи результатів:

1) ймовірність відкриття депозиту чи залучення (значення від 0 до 1);
2) фінальне передбачення (0 або 1) залежно від порогу прийняття рішення.

Щоб забезпечити зручність аналізу, усі результати були приєднані до датафрейму із сегментацією клієнтів. До нього додано нові колонки:

1) proba_deposit — прогнозована ймовірність відкриття депозиту;
2) pred_deposit — фінальне рішення (1 = прогнозується відкриття депозиту);
3) proba_credit — прогнозована ймовірність отримання кредиту;
4) pred_credit — фінальне рішення (1 = прогнозується отримання кредиту).

Відображаються лише безпечні для демонстрації колонки: client_id_masked, cluster, proba_deposit, pred_deposit, proba_credit, pred_credit.

Таким чином зберігається анонімність клієнтів та водночас залишається можливість аналізувати їх поведінку.

На наступному кроці було проведено групування за кластерами. Для кожного сегмента клієнтів виводяться ті, у кого прогнозується позитивна відповідь за депозитом чи кредитом (рис. 3.7, 3.8). Це дозволяє:

1) визначити кількість клієнтів у кожному сегменті, які з високою ймовірністю відкриють депозит;
2) аналогічно оцінити частку клієнтів, які виявлять попит на кредит;
3) порівняти сегменти між собою і виділити найбільш перспективні групи.

```
=== SEGMENT 0 ===
Deposit predicted positives (count=1971):
```

	client_id_masked	proba_deposit
34017	217****5931	0.9475
31283	385****0760	0.9475
33953	798****0249	0.9350
34047	653****6346	0.9275
33907	363****5578	0.9250
39763	965****4927	0.9225
42282	455****6759	0.9225
34164	996****5693	0.9225
40163	650****0213	0.9200
39937	967****7489	0.9200
40087	722****9377	0.9175
39835	988****4867	0.9175
33765	818****1231	0.9150
39787	572****0736	0.9150
32035	349****4769	0.9150
34009	476****8691	0.9125
31277	808****8752	0.9125
34035	971****9538	0.9125
34142	638****4081	0.9100
41589	377****9768	0.9050

Рисунок 3.7 — Приклад виведення результатів прогнозування взяття депозиту для сегменту 0.

Credit predicted positives (count=4680):

	client_id_masked	proba_credit
38428	817****8630	1.0
39341	821****0150	1.0
38370	207****4646	1.0
39279	816****9843	1.0
38234	326****4728	1.0
36541	748****1341	1.0
38117	906****3855	1.0
36438	266****5510	1.0
38093	307****3316	1.0
36733	436****4491	1.0
38164	766****4923	1.0
38161	943****4471	1.0
36440	664****0995	1.0
36383	579****4669	1.0
37798	729****8501	1.0
36114	650****4206	1.0
37630	981****5274	1.0
32214	705****3113	1.0
35772	183****9879	1.0
32346	238****9112	1.0

Рисунок 3.8 — Приклад виведення результатів прогнозування взяття кредиту для сегменту 1.

Опис функції:

```
for cl in unique_clusters:
    seg_slice = seg_view[seg_view[cluster] == cl].copy()
    dep_pos = seg_slice[seg_slice[pred_deposit] == 1]
    cred_pos = seg_slice[seg_slice[pred_credit] == 1]
    print(fSegment {cl}: deposits={len(dep_pos)}, credits={len(cred_pos)})
```

3.7 Висновки до розділу 3

У даному розділі було реалізовано повний цикл побудови системи прогнозування та сегментації клієнтського портфелю банку на основі сучасних методів аналізу даних і машинного навчання. Спочатку проведено підготовку та очищення даних: сформовано унікальні ідентифікатори клієнтів, реалізовано їх анонімізацію шляхом хешування, а також налаштовано механізми обробки пропусків та кодування категоріальних ознак. Це забезпечило якісну основу для подальшого аналізу та збереження конфіденційності персональних даних.

Далі була проведена сегментація клієнтів методом GMM, що дало змогу виділити три ключові групи з відмінними соціально-демографічними та фінансовими характеристиками. За допомогою профілювання для кожного сегмента сформовано узагальнені портрети клієнтів, які відрізняються за віком, рівнем доходів, активністю транзакцій та перевагами у використанні банківських продуктів. Це створило підґрунтя для розробки цільових стратегій роботи з кожною групою.

На етапі моделювання було розглянуто кілька алгоритмів класифікації — логістичну регресію, випадковий ліс і градієнтний бустинг. Оцінювання моделей проводилося за ключовими метриками якості (Accuracy, F1, Precision, Recall, ROC-AUC). Порівняльний аналіз показав, що ансамблеві методи (Random Forest і Gradient Boosting) забезпечують більш високу точність прогнозування як у задачі відкриття депозитів, так і у задачі прогнозування кредитів.

Завершальним етапом стала інтеграція результатів прогнозування з сегментацією клієнтів. Для кожного клієнта було визначено ймовірність та фінальне рішення щодо відкриття депозиту і отримання кредиту, після чого ці дані були агреговані на рівні сегментів. Це дозволило визначити, які групи клієнтів є найбільш перспективними для депозитних продуктів, а які — для кредитних.

Таким чином, проведений аналіз довів ефективність поєднання сегментації та прогнозного моделювання для управління клієнтським портфелем банку. Запропонований підхід забезпечує можливість не лише класифікувати клієнтів за поведінковими ознаками, але й передбачати їх майбутні дії, що створює основу для персоналізованих маркетингових стратегій, підвищення прибутковості та зменшення ризиків у банківській діяльності.

РОЗДІЛ 4 РОЗРОБКА ВЛАСНОГО СТАРТАП ПРОЕКТУ

У сучасних умовах банківська сфера перебуває на етапі масштабних змін, спричинених активною цифровізацією та посиленням конкуренції між фінансовими установами. Одним із головних викликів стає потреба у створенні точних і гнучких систем прогнозування клієнтського портфелю. Йдеться не лише про визначення ризикових клієнтів чи ймовірності відтоку, а й про сегментацію, виявлення потенційно прибуткових груп та формування індивідуальних стратегій взаємодії з ними.

Традиційні статистичні підходи та класичні скорингові моделі вже не здатні повною мірою задовольнити ці потреби. Робота з великими масивами клієнтських даних вимагає не лише значних обчислювальних ресурсів, а й використання складних алгоритмів аналізу, які забезпечують високу точність та адаптивність. Саме тому зростає інтерес до інтелектуальних систем прогнозування, що базуються на методах машинного навчання та штучного інтелекту.

Застосування таких систем дозволяє оптимізувати управління клієнтською базою, зменшувати кредитні ризики, підвищувати рівень лояльності клієнтів та ефективніше розподіляти маркетингові бюджети. У результаті банки отримують можливість не лише зберігати конкурентоспроможність, а й формувати нові джерела прибутку.

Таким чином, проблема, яку вирішує запропонований стартап, є надзвичайно актуальною. Він поєднує прогнозування поведінки клієнтів, автоматизовану сегментацію та підтримку прийняття рішень на основі інтелектуальних алгоритмів. Це робить проєкт цінним інструментом для фінансового сектору, здатним забезпечити конкурентні переваги та стати ключовим рішенням для управління клієнтським портфелем у сучасних банківських умовах.

4.1 План розробки стартапу та масштабування його на ринок

Розробка стартапу MishBank передбачає комплексний підхід, що охоплює всі етапи — від початкового аналізу ринку до комерціалізації та масштабування продукту. Такий підхід забезпечує не лише створення якісного інструменту для автоматизованого аналізу текстових даних, але й формування стійкої бізнес-моделі для його подальшого розвитку.

На старті важливим є проведення маркетингових досліджень, які включають такі ключові напрями.

1. Конкурентний аналіз — вивчення існуючих рішень на ринку, методів та технологій обробки текстових даних, їхніх можливостей та обмежень. Це дозволяє визначити конкурентні переваги стартапу та шляхи його унікалізації.
2. Формування ідеї проєкту та визначення цільової аудиторії — розробка ключових функцій продукту та його відмінних характеристик. Основною аудиторією визначено дослідників та аналітиків у сферах R&D, маркетингу, управління інноваціями та високотехнологічних компаній, які потребують інструментів для автоматизованого виявлення взаємозв’язків між технологіями.
3. Розробка стратегії виходу на ринок — створення плану запуску з урахуванням основних каналів просування, а також визначення способів залучення потенційних користувачів і клієнтів.

Другим кроком є організація процесу розробки стартапу:

1) побудова детального плану та таймлайну реалізації проєкту, включно зі створенням MVP (мінімально життєздатного продукту), тестуванням та запуском;
2) оцінка необхідних ресурсів – фінансових, технічних і кадрових;
3) розрахунок витрат, до яких належать витрати на оплату праці фахівців, інфраструктуру (сервери, офіс), програмне забезпечення та інші операційні потреби.

Далі формується економічна модель проєкту:

1) визначається обсяг інвестиційних витрат для старту;
2) розраховуються ключові фінансово-економічні показники – собівартість, ціна реалізації, податкове навантаження, рівень прибутку;
3) аналізується інвестиційна привабливість проєкту за показниками рентабельності та терміном окупності;
4) ідентифікуються потенційні ризики (технічні, фінансові, ринкові) та розробляються заходи для їх мінімізації – створення резервного фонду, диверсифікація джерел фінансування тощо.

Фінальним етапом є підготовка заходів для комерціалізації та масштабування:

1) проведення дослідження інтересів потенційних інвесторів та партнерів;
2) формування інвестиційної пропозиції з описом продукту, його функціоналу, поточних результатів та перспектив зростання;
3) вибір оптимальних каналів комунікації з інвесторами та клієнтами з метою ефективного просування продукту й формування довготривалих зв’язків із цільовою аудиторією.

4.2 Опис ідеї стартап-проекту

Стартап-проєкт спрямований на вирішення задачі аналізу та прогнозування поведінки клієнтів банку з використанням сучасних методів машинного навчання та сегментаційних алгоритмів. Ідея полягає у створенні інструменту, який на основі обробки великих масивів клієнтських даних дозволить автоматизовано визначати ключові сегменти клієнтів, прогнозувати ймовірність відкриття депозитів чи залучення кредитів, а також виявляти приховані закономірності у фінансовій поведінці. Це рішення передбачає інтеграцію алгоритмів кластеризації (Gmm), моделей прогнозування (Random

Forest, Gradient Boosting) та аналітичних візуалізацій для наочного представлення результатів. Проєкт буде особливо корисним для банківських аналітиків, фінансових менеджерів та маркетологів, адже надає інструменти для прийняття обґрунтованих управлінських рішень, підвищення прибутковості, зниження ризиків та ефективнішої взаємодії з клієнтами. У таблиці 4.1 наведена інформаційна карта стартапу.

Таблиця 4.1 – Інформаційна карта стартап-проекту

Назва проекту	MishBank
Автори проекту	Міщенко Антон Сергійович
Коротка анотація	Інструмент для аналізу та прогнозування поведінки клієнтів банку, що поєднує методи машинного навчання та сегментаційні алгоритми. Система автоматизовано визначає ключові сегменти клієнтів, прогнозує ймовірність відкриття депозитів чи залучення кредитів та формує профілі груп для підтримки управлінських рішень.
Термін реалізації проекту	12 місяців
Необхідні ресурси	Приміщення для роботи команди, комп’ютерне обладнання з доступом до Інтернету та хмарних обчислювальних потужностей, програмне забезпечення для аналітики та машинного навчання, фінансові кошти на оплату праці, , а також хмарні сховища для зберігання великих масивів даних.

Продовження таблиці 4.1

Опис проблеми, яку вирішує проект	Ручний аналіз клієнтських даних є надзвичайно трудомістким і часто не дозволяє вчасно виявити закономірності у поведінці клієнтів. Запропонований продукт вирішує проблему автоматизованого прогнозування та сегментації клієнтського портфелю, дозволяючи банку точніше оцінювати ймовірність відкриття продуктів, а також формувати персоналізовані пропозиції. Це підвищує ефективність і точність прийняття управлінських рішень.
Головні цілі та завдання проекту	1. Розробити систему, здатну обробляти великі обсяги клієнтських даних у реалістичних умовах. 2. Автоматично ідентифікувати сегменти клієнтів за допомогою алгоритмів кластеризації. 3. Прогнозувати поведінку клієнтів (депозити, кредити) на основі моделей машинного навчання.
Очікувані результати	Створення інструменту, який стане незамінним для банківських установ, фінансових аналітиків та маркетологів, що працюють з великими обсягами даних. Система дозволить підвищити точність прогнозування, знизити ризики, оптимізувати управління клієнтським портфелем та забезпечити ефективнішу взаємодію з клієнтами завдяки персоналізованим підходам.

4.3 Технологічний аудит ідеї проекту

У таблиці 4.2 наведено опис основних технічних аспектів та напрямків застосування технологій у проєкті.

Таблиця 4.2 – Опис ідеї стартапу

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Розробка інструменту для автоматизованого аналізу клієнтських даних, сегментації та прогнозування фінансової поведінки клієнтів банку із застосуванням методів машинного навчання та кластеризаційних алгоритмів.	Управління клієнтським портфелем банку, маркетинг та персоналізація банківських продуктів, аналіз ризиків та прогнозування ймовірності відтоку клієнтів, підтримка прийняття рішень на основі даних (Data-Driven Decision Making).	Інструмент дозволяє швидко та ефективно сегментувати клієнтів, виявляти приховані закономірності у їхній поведінці, прогнозувати ймовірність відкриття депозитів чи залучення кредитів.

Далі проведемо порівняльний аналіз конкурентів проекту та наведемо результати у таблиці 4.3.

Таблиця 4.3 – Порівняльний аналіз конкурентів проекту

№	Техніко-економічні характеристики ідеї	потенційні товари/концепції конкурентів				W	N	S
		Власний проект	SAS Customer Intelligence	IBM SPSS Modeler	Palantir Foundry			
1	Якість аналізу зв’язків	Висока точність завдяки інтеграції кластеризації (Gmm) та моделей ML (Random Forest, Gradient Boosting)	Висока для задач маркетингової аналітики	Висока, але потребує складного налаштування	Висока, але орієнтована на корпоративні масштаби			+
2	Доступність по ціні	Гнучка цінова політика, можливість створення MVP безкоштовно	Висока вартість для корпоративних клієнтів	Висока ціна ліцензій	Дуже висока, розрахована на великі корпорації		+	
3	Персоналізація та адаптація під потреби користувача	Гнучке налаштування під потреби конкретного банку (депозити, кредити, сегменти клієнтів)	Є галузеві шаблони, але вимагають адаптації	Середня - потребує додаткових інтеграцій	Висока, але потребує значних ресурсів			+

Продовження таблиці 4.3

4	Простота у використанні	Інтуїтивний інтерфейс з візуалізаціями (сегменти, прогнози)	Потребує навчання користувачів	Простий інтерфейс, підходить для базових завдань	Складний для звичайних користувачів, орієнтований на експертів			+
5	Унікальність	Поєднання сегментації клієнтів і прогнозного моделювання у єдиній системі	Аналіз маркетингових кампаній	Фокус на статистичному моделюванні	Потужний інструмент, але без акценту на банківську сегментацію		+	
6	Варіанти використання	Управління клієнтським портфелем, маркетинг, прогнозування відтоку клієнтів, кредитний скоринг	Бізнес-аналітика, маркетингова ефективність	Основні задачі статистичного аналізу, огляд даних	Складні корпоративні проєкти з аналітики та безпеки			+

Далі аналізуємо реальність технічно здійснити ідею проекту (таблиця 4.4).

Таблиця 4.4 – Технологічна здійсненність продукту

№	Ідея проекту	Технології і реалізації	Наявність технологій	Доступність технологій
1	Використання методів машинного навчання для прогнозування поведінки клієнтів (депозити, кредити)	Random Forest, Gradient Boosting, Logistic Regression (sklearn, xgboost)	Наявні	Доступні як бібліотеки з відкритим кодом
2	Сегментація клієнтів для формування портфелю	Gmm (sklearn), PCA для візуалізації	Наявні	Доступні як модулі Python
3	Обробка та підготовка клієнтських даних	pandas, NumPy, SimpleImputer, StandardScaler, OneHotEncoder	Наявні	Доступні у вільному д source)
4	Візуалізація результатів сегментації та прогнозування	Matplotlib, Seaborn, Plotly	Наявні	Доступні як бібліотеки Python
5	Хмарне зберігання та обробка даних	Хмарне зберігання та обробка даних	Наявні	Доступні за підпискою
Обрана технологія реалізації ідеї проекту: Python				

4.4 Аналіз ринкових можливостей запуску стартап-проекту

Далі проведемо попередній аналіз ринку для запуску стартап-проекту (таблиця 4.5).

Таблиця 4.5 – Попередня характеристика потенційного ринку стартап-проекту

№	Показники ринку (найменування)	Характеристика
1	Кількість основних гравців	5–7 великих компаній та консалтингових агентств, що працюють у сфері фінансової аналітики, CRM та управління клієнтськими портфелями
2	Загальний обсяг ринку	$3.5 млрд з потенційним зростанням до $6 млрд у найближчі 5 років завдяки цифровізації банківського сектору
3	Динаміка ринку	Позитивна, очікується зростання 18–22% щороку за рахунок впровадження AI та Data Science у банках
4	Наявність обмежень для входу	Низькі бар’єри входу для MVP-рішень, однак високі вимоги до обробки персональних даних, кібербезпеки та відповідності регуляторним нормам (GDPR, НБУ)
5	Специфічні вимоги до стандартизації та сертифікації	Відсутні обов’язкові сертифікації, але важливо забезпечити відповідність політикам безпеки та стандартам захисту клієнтських даних
6	Середня норма рентабельності в галузі	15%–25%, залежно від рівня автоматизації та масштабу впровадження систем прогнозної аналітики

Далі розглянемо потенційних клієнтів, які можуть бути зацікавлені у використанні продукту MishBank. З огляду на широкий спектр можливостей інструменту (сегментація клієнтів, прогнозування депозитів та кредитів, управління ризиками, автоматизація аналітики), цільова аудиторія охоплює як банківські установи, так і суміжні організації, що працюють із фінансовими даними. (таблиця 4.6).

Таблиця 4.6 – Характеристика потенційних клієнтів стартап-проекту

№	Потреби, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Вимоги споживачів до товару
1	Аналіз клієнтської поведінки та прогноз депозитів/кредитів	Комерційні банки, фінансові установи	Висока точність прогнозів, автоматизація процесу
2	Дослідження конкурентного середовища та оптимізація маркетингу	Маркетингові агентства, банківські аналітики	Доступність, простота використання, інтеграція з CRM
3	Виявлення ризикових клієнтів і зниження кредитних втрат	Державні установи, регулятори, небанківські фінансові компанії	Надійність, зручний інтерфейс, дотримання вимог безпеки даних (GDPR, НБУ)
4	Спрощення управління клієнтським портфелем і персоналізація продуктів	Банківські R&D-відділи, фінтех-компанії	Інтерактивність, можливість адаптації під специфічні бізнес-процеси

SWOT-аналіз ринкових можливостей проєкту MishBank допоможе виявити сильні та слабкі сторони продукту, можливості для розширення, а також основні загрози, з якими може зіткнутися стартап.

Обрахуємо фактори загроз (таблиця 4.7) та можливостей (таблиця 4.8). Проаналізуємо загрози, щоб зрозуміти потенційні перешкоди при запуску

продукту на ринок. Фактори можливостей необхідно оцінити, щоб визначити сприятливі умови для розвитку та максимально ефективно їх використати.

Таблиця 4.7 – Фактори загроз

№	Фактор	Зміст загрози	Можлива реакція компанії
1	Конкуренція	Присутність великих компаній і банків, що вже мають власні системи аналітики та значні ресурси для розвитку подібних продуктів	Створити унікальні функції (сегментація, персоналізовані прогнози), які виділять MishBank на фоні конкурентів; інвестувати в маркетинг
2	Кібербезпека	Ризик витоку клієнтських даних під час обробки великих обсягів інформації та персональних даних	Забезпечити високий рівень кібербезпеки, сертифікацію, шифрування та надійний захист даних
3	Технічна складність	Високі вимоги до обчислювальних ресурсів при роботі з великими клієнтськими базами	Інвестувати в оптимізацію алгоритмів, використання хмарних потужностей для масштабованості
4	Вартість Інтеграції для клієнтів	Висока вартість впровадження рішення для малих банків та фінансових компаній	Запропонувати гнучку цінову політику: безкоштовний доступ до MVP, різні рівні підписки
5	Зміна регуляторних норм	Можливі зміни у вимогах до захисту фінансових даних і персональних даних (НБУ, GDPR)	Відслідковувати регуляторні зміни та адаптувати продукт до нових стандартів безпеки

Продовження таблиці 4.7

6	Низький рівень довіри нових користувачів	Відсутність сильної репутації на ринку може знизити довіру до продукту	Інвестувати в PR та маркетинг, надавати безкоштовні демо-версії для ознайомлення з можливостями MishBank

Таблиця 4.8 – Фактори можливостей

№	Фактор	Зміст загрози	Можлива реакція компанії
1	Зростання попиту на автоматизацію фінансової аналітики	Підвищений інтерес банків до рішень, що дозволяють автоматизувати сегментацію клієнтів та прогнозування поведінки	Розширити функціонал продукту, додавши нові модулі для кредитного скорингу та управління ризиками
2	Універсальність застосування	Можливість використання продукту у різних сферах фінансового сектору: комерційні банки, страхові компанії, фінтех	Орієнтувати маркетингову стратегію на кілька сегментів ринку для збільшення охоплення
3	Зростання ринку фінансових технологій (FinTech)	Ринок фінансової аналітики та прогнозних систем швидко зростає	Інвестувати у розвиток продукту та підтримку новітніх ML-алгоритмів для збереження конкурентоспроможності

Продовження таблиці 4.8

4	Технологічний прогрес у ML/AI	Постійні покращення у методах машинного навчання, прогнозуванні та обробці великих даних	Залучати інновації та регулярно оновлювати алгоритми для підвищення точності та швидкодії
5	Можливості інтеграції з іншими системами	Потенціал інтеграції MishBank з CRM, системами управління ризиками та банківськими платформами	Впровадити API для легкої інтеграції з існуючою інфраструктурою банку
6	Поширення цифровізації банківських послуг	Більше установ переходять на цифрові канали обслуговування клієнтів, зростає попит на онлайн-аналітику	Розвивати продукт як хмарне рішення, забезпечуючи доступність та простоту використання з будь-якого пристрою

Далі розглянемо питання конкуренції, а саме визначимо її тип та рівень (таблиця 4.9).

Таблиця 4.9 – Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	У чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
Тип конкуренції	Недосконала конкуренція, зростання кількості фінтех-рішень для аналітики клієнтів	Розробка унікальних функцій (прогноз депозитів і кредитів, сегментація клієнтів) для виділення на ринку
Рівень конкурентної боротьби	Міжнародний рівень, ринок представлений великими гравцями (SAS, IBM, Palantir, Google)	Вихід на глобальний ринок із фокусом на вузькі сегменти (банківські установи), адаптація продукту під локальні потреби
Галузева ознака	Внутрішньогалузева конкуренція з іншими системами бізнес-аналітики та CRM	Підвищення якості сервісу, надання спеціалізованих можливостей для банківської сфери
Конкуренція за видами товарів	Товарно-родова конкуренція з продуктами для фінансової аналітики та кредитного скорингу	Позиціонування MishBank як комплексного рішення: сегментація + прогнозування + візуалізація можливостей
Характер конкурентних переваг	Нецінова конкуренція: головний акцент робиться на функціоналі, точності прогнозів та інтеграції	Підтримка високої якості прогнозів, використання інноваційних алгоритмів та зручних інтерфейсів

Продовження таблиці 4.9

Інтенсивність конкуренції	Марочна конкуренція: головні гравці мають відомі бренди та сформовану клієнтську базу	Формування сильної брендової ідентичності MishBank, акцент на унікальній цінності для банківських клієнтів

Далі виконаємо аналіз конкуренції за моделлю 5 сил конкуренції Майкла Портера (таблиця 4.10).

Таблиця 4.10 – Аналіз конкуренції в галузі за М. Портером

Складові аналізу	Прямі конкуренти у галузі	Потенційні конкуренти	Постачальники	Клієнти	Товаровиробники
	Великі компанії, такі як SAS Customer Intelligence, IBM SPSS Modeler, Palantir Foundry, що вже пропонують схожі рішення для фінансової аналітики та клієнтського скорингу	Зростаючий попит на автоматизовану фінансову аналітику може залучити нових гравців, зокрема фінтех-стартапи, які працюють з ML/AI	Постачальники обчислювальних ресурсів та хмарних сервісів (AWS, Google Cloud, Microsoft Azure)	Банки, небанківські фінансові установи, маркетингові агентства, державні регулятори; усі вони мають високі вимоги до точності прогнозів і безпеки даних	Традиційні методи аналізу даних у банках (SQL-звіти, статистичні моделі) або базові BI-інструменти без прогнозної аналітики

Продовження таблиці 4.10

Вплив на діяльність проєкту MishBank	Необхідність створення унікальних функцій і конкурентних переваг, таких як інтеграція сегментації з прогнозуванням та візуалізацією клієнтських даних	Фокус на швидкому масштабуванні та вдосконаленні функціоналу для випередження нових гравців	Вибір надійних постачальників і оптимізація витрат на обчислювальні ресурси для забезпечення рентабельності	Підтримка високої якості, гнучка цінова політика, можливість персоналізації продукту під конкретний банк	Позиціонування MishBank як більш ефективного рішення, що поєднує сегментацію, прогнозування та візуалізацію для комплексного аналізу клієнтів

Маючи результати аналізу конкуренції (таблиця 4.10), характеристики ідеї стартап-проекту (таблиця 4.5), характеристики потенційних клієнтів і їх вимоги до продукту (таблиця 4.6) та фактори ринкового середовища (таблиці 4.7 і 4.8), сформулюємо та обґрунтуємо перелік факторів конкурентоспроможності (таблиця 4.11).

Таблиця 4.11 – Обґрунтування факторів конкурентоспроможності

№	Фактор конкурентоспроможності	Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)
1	Унікальне поєднання сегментації та прогнозування	MishBank пропонує інтегровану систему, яка поєднує сегментацію клієнтів (кластеризація) та прогнозування їхньої поведінки (депозити/кредити). Це забезпечує глибший аналіз, ніж звичайні CRM чи BI-інструменти, які зазвичай працюють лише з ретроспективними даними.
2	Гнучка цінова політика та доступність MVP	На відміну від великих конкурентів (SAS, IBM), MishBank орієнтується на малий та середній банківський бізнес, пропонуючи безкоштовний доступ до MVP та кілька рівнів підписки. Це дозволяє банкам поступово інтегрувати рішення без значних початкових витрат.
3	Інтуїтивно зрозумілий інтерфейс і простота використання	З урахуванням потреб потенційних користувачів, продукт забезпечує зрозумілу візуалізацію даних (сегменти, ймовірності депозитів/кредитів), що не вимагає від користувача глибоких технічних знань. Це створює перевагу над складними корпоративними рішеннями, які потребують тривалого навчання.
4	Персоналізація для різних фінансових установ	Завдяки гнучкій архітектурі, MishBank може налаштовуватися під специфіку банків, страхових компаній чи фінтех-компаній, що дозволяє адаптувати модель прогнозування під індивідуальні потреби.

Продовження таблиці 4.11

5	Підвищений рівень кібербезпеки та конфіденційності даних	В умовах високих вимог до захисту персональних і фінансових даних, продукт включає механізми хешування унікальних ідентифікаторів, захищену роботу з базами даних та відповідність регуляторним нормам (НБУ, GDPR).
6	Позиціонування як інноваційного рішення на ринку	Завдяки можливості поєднання сучасних методів машинного навчання, прогнозного моделювання та сегментаційного аналізу, продукт відрізняється від конкурентів і здатний зайняти нішу як комплексне рішення для управління клієнтським портфелем банку.

Тепер можна провести аналіз сильних та слабких сторін продукту (таблиця 4.12).

Таблиця 4.12 – Порівняльний аналіз сильних та слабких сторін продукту

№	Фактор конкурентоспроможності	Бали 1-20	Рейтинг товарів-конкурентів						
			-3	-2	-1	0	1	2	3
1	Унікальне поєднання сегментації та прогнозування	18	+						
2	Гнучка цінова політика та доступність MVP	17		+					
3	Інтуїтивно зрозумілий інтерфейс	15				+			

Продовження таблиці 4.12

4	Підтримка персоналізації для різних фінансових установ	16						+	
5	Підвищений рівень витоку інформації	13				+			
6	Позиціонування як інноваційного рішення	18	+						

Далі проведемо SWOT-аналіз продукту (таблиця 4.13).

Таблиця 4.13 – SWOT-аналіз стартап-проекту

Сильні сторони	Унікальне поєднання сегментації клієнтів та прогнозування їхньої поведінки (депозити/кредити); гнучка цінова політика та доступність MVP, що робить рішення привабливим для малих і середніх банків; інтуїтивно зрозумілий інтерфейс, який не потребує від користувачів глибоких технічних знань; високий рівень кібербезпеки та конфіденційності даних із дотриманням регуляторних вимог (НБУ, GDPR); можливість персоналізації продукту під специфіку різних фінансових установ.
Слабкі сторони	Відсутність сильної брендової репутації на ранніх етапах розвитку; значні вимоги до обчислювальних ресурсів при роботі з великими клієнтськими базами; залежність від постачальників хмарних сервісів та обчислювальних потужностей.
Можливості	Зростаючий попит на автоматизацію фінансової аналітики та прогнозування поведінки клієнтів; інтеграція з іншими бізнес-інструментами та банківськими CRM-системами; використання продукту у різних сферах фінансового сектору (банки, страхові компанії, фінтех); постійний розвиток технологій машинного навчання та аналітичних інструментів, що дає змогу вдосконалювати продукт.

Продовження таблиці 4.13

Загрози	Посилення конкуренції з боку великих гравців, які можуть швидко адаптувати подібні рішення (SAS, IBM, Palantir); ризики витоку чи кібератак при обробці конфіденційних фінансових даних; можливі зміни у регуляторних вимогах до захисту даних у різних країнах; недостатня довіра нових користувачів до молодого бренду на початковому етапі.

Завдяки проведенню SWOT-аналізу, ми змогли визначити сильні та слабкі сторони, можливості та загрози, пов'язані з конкуренцією та плануванням стартап-проекту. Далі спроектуємо альтернативну ринкову поведінку для інтеграції стартап-проекту на ринок та приблизний час реалізації системного комплексу, з урахуванням потенційних проектів, що можуть бути виведені на ринок та наведемо результати у таблиці 4.14.

Таблиця 4.14 – Альтернативи ринкового впровадження стартап-проекту

№	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	Запуск MVP з базовим функціоналом на безкоштовній основі для привернення уваги та залучення клієнтської бази	80%	2 місяці
2	Розробка платної підписки з розширеним функціоналом (додаткові модулі: управління ризиками, інтеграція з CRM, персоналізовані звіти)	65%	6 місяців

Продовження таблиці 4.14

3	Співпраця з великими фінансовими установами для інтеграції MishBank у їхні робочі процеси через API та персоналізовані рішення	50%	8 місяців
4	Вихід на міжнародні ринки через партнерські програми, локалізацію продукту та відповідність регуляторним вимогам (GDPR, PSD2)	40%	12 місяців

У даному пункті був проведений детальний аналіз ринку та продукту. Також відповідно до результатів проведеного конкурентного аналізу, визначених факторів ринку та його сприятливість, описання ідеї та характеристик стартап-проекту, робимо висновок висновок, що існують дуже сприятливі умови для виходу продукту на ринок.

4.5 Розроблення ринкової стратегії стартап-проекту

Для розробки ринкової стратегії продукту, у першу чергу, необхідно проаналізувати цільову аудиторію проекту (таблиця 4.15).

Таблиця 4.15 – Вибір цільових груп потенційних споживачів

№	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
1	Великі комерційні банки та фінансові корпорації, які активно впроваджують цифрові рішення для управління клієнтськими портфелями	Висока	35%	Висока	Середня
2	Маркетингові та консалтингові агентства, що аналізують банківські послуги та клієнтські дані для побудови стратегій	Середня	20%	Середня	Середня
3	Державні установи та регулятори, що здійснюють контроль за фінансовим ринком та прогнозування ризиків	Висока	15%	Низька	Висока

Продовження таблиці 4.15

4	Малі та середні банки, небанківські фінансові організації, що шукають доступні рішення для кредитного скорингу та прогнозування	Середня	20%	Середня	Середня
5	Академічні установи та дослідницькі центри, які вивчають фінансові тенденції та моделювання клієнтської поведінки	Низька	10%	Низька	Низька

Маючи аналіз цільових груп, далі визначимо базову стратегію розвитку продукту (таблиця 4.16).

Таблиця 4.16 – Визначення базової стратегії розвитку

№	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи
1	Випуск безкоштовної версії MVP для залучення користувачів	Стратегія проникнення на ринок	Доступність продукту, швидке формування клієнтської бази, зниження бар’єрів входу для малих та середніх банків

Продовження таблиці 4.16

2	Впровадження платної підписки з розширеним функціоналом (управління ризиками, інтеграція з CRM, розширені звіти)	Стратегія сегментування ринку	Персоналізація, підтримка різних галузевих потреб, гнучка цінова політика
3	Партнерство з великими банками та фінансовими корпораціями для інтеграції MishBank у їхні робочі процеси	Стратегія нішевого маркетингу	Висока якість і надійність продукту, можливість масштабної інтеграції через API та корпоративні модулі
4	Розширення на міжнародні ринки через локалізацію та партнерські програми	Стратегія диференційованого маркетингу	Адаптація продукту під регуляторні вимоги, культурні та мовні особливості, зміцнення глобальної присутності

Для роботи в обраних сегментах ринку сформовано базову стратегію розвитку (таблиці 4.17, 4.18).

Таблиця 4.17 Визначення базової стратегії конкурентної поведінки

Чи є проект першопрохідцем на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки
Ні. Ринок уже містить кількох значних гравців у фінансовій аналітиці та скорингу (корпоративні BI/ML-платформи). MishBank виходить не як перший, а як сфокусоване галузеве рішення для банків.	Так, у пріоритеті - залучення нових користувачів через безкоштовний MVP і гнучку підписку для МСБ-банків та небанківських установ. Паралельно — переманювання частини клієнтів конкурентів завдяки кращій персоналізації, прозорій економіці впровадження та швидкому POC.	Ні в лоб. Базові очікувані функції (дашборди, інтеграції через API, базові звіти) — так, але ключовий акцент на диференціації: вбудована сегментація + прогноз депозитів/кредитів, анти-лікідж препроцесинг, анонімізація клієнтських ID, готові галузеві пресети під банки.	Позиціонування як галузевого (banking-focused) інструменту сегментація + прогнозування; цінова гнучкість, швидкі пілоти; інтеграційна відкритість (API, підключення до CRM/даних банку) і акцент на безпеці (маскування ідентифікаторів, відповідність вимогам НБУ/GDPR); швидке оновлення моделей та бібліотеки сценаріїв (відтік, апсел, депозитні кампанії)

Таблиця 4.18 – Визначення стратегії позиціонування

Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспро можні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
Висока точність прогнозів, автоматизація обробки великих масивів клієнтських даних, інтуїтивно зрозумілий інтерфейс та можливість персоналізації під потреби різних банків і фінансових установ.потреби	Стратегія диференціації — створення унікальних можливостей продукту завдяки поєднанню сегментаційних моделей і прогнозної аналітики (депозити, кредити), інтеграції з CRM та високих стандартів безпеки.	Інноваційний функціонал для прогнозування та сегментації клієнтів, гнучка цінова політика (безкоштовний MVP, підписки різного рівня), висока якість кібербезпеки та конфіденційност і, інтеграційна відкритість (API, підключення до банківських систем).	Простота використання — інтерфейс, зрозумілий навіть без глибоких технічних знань; глибокий аналіз клієнтських даних; персоналізація — адаптація рішення під специфіку кожного банку чи фінансової організації.

4.6 Розроблення маркетингової програми стартап-проекту

Після проведеного комплексного аналізу, можемо повноцінно описати ключові переваги концепції потенційного товару (таблиця 4.19) та побудувати концепцію маркетингових комунікацій (таблиця 4.20).

Таблиця 4.19 – Ключові переваги концепції потенційного товару

№	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити
1	Висока точність та глибокий аналіз клієнтських даних	Надання точних прогнозів щодо депозитів/кредитів, автоматизований аналіз портфелю клієнтів, виявлення прихованих закономірностей	Поєднання сучасних ML-алгоритмів для сегментації та прогнозування, використання історичної поведінки клієнтів
2	Доступність та гнучка цінова політика	Безкоштовний MVP для тестування, різні рівні платних підписок для банків різного масштабу	Гнучкість у ціноутворенні, орієнтація на малий та середній бізнес, поступове масштабування
3	Простота використання	Інтуїтивно зрозумілий інтерфейс, що дозволяє швидко інтегрувати продукт у робочі процеси	Проста інтеграція з CRM та банківськими системами, відсутність потреби у високій технічній підготовці персоналу
4	Персоналізація під галузеві потреби	Адаптація функціоналу під конкретні сегменти: банківські відділення, страхові компанії, небанківські установи	Гнучкі налаштування, можливість створення кастомних сценаріїв (ризики, відтік клієнтів, апсел)
5	Забезпечення безпеки даних	Високий рівень кіберзахисту та конфіденційності клієнтських даних	Маскування ідентифікаторів клієнтів, відповідність регуляторним стандартам (НБУ, GDPR), захищені інтеграції через API

Таблиця 4.20 – Концепція маркетингових комунікацій

№	Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення
1	Пошук інноваційних рішень для аналізу клієнтських даних та прогнозування	Професійні онлайн-платформи (Finextra, BankingTech), галузеві конференції, фінансові видання	Унікальність продукту, глибокий аналіз клієнтської поведінки	Підкреслити інноваційний функціонал та конкурентні переваги
2	Орієнтація на доступність та простоту інтеграції	Соціальні мережі (LinkedIn, Facebook), таргетована реклама у фінансових спільнотах	Доступність, інтуїтивний інтерфейс	Показати легкість інтеграції з CRM та швидке впровадження у робочі процеси
3	Потреба в персоналізації під специфіку установи (банк, небанківська компанія, фінтех)	Прямі контакти, email-розсилки, презентації продукту	Персоналізація, можливість налаштування під конкретні потреби	Демонструвати адаптивність рішень для різних сегментів фінансового ринку

Продовження таблиці 4.20

4	Висока чутливість до кібербезпеки та захисту клієнтських даних	Спеціалізовані вебінари, тематичні форуми з кібербезпеки, сайти новин.	Безпека даних, відповідність стандартам (НБУ, GDPR)	Підкреслити високий рівень кіберзахисту, хешування ідентифікаторів та надійність продукту
5	Потреба у тестуванні продукту перед масштабним впровадженням	Безкоштовний доступ до MVP, відгуки та кейси від лідерів думок	Можливість спробувати продукт безкоштовно	Заклик протестувати MVP, сформувати довіру та лояльність користувачів

4.7 Висновки до розділу 4

Даний розділ був присвячений дослідженню стартап-проєкту MishBank, що орієнтований на сегментацію та прогнозування клієнтського портфелю банку. У рамках розділу було розглянуто стратегії виходу на ринок та маркетингові підходи, які забезпечать конкурентоспроможність проєкту. Ринок фінансової аналітики та прогнозування є динамічним і зростаючим, однак на ньому домінують великі гравці, які пропонують переважно універсальні рішення. На відміну від них, MishBank зосереджується на галузевій специфіці та забезпечує глибоку інтеграцію з банківськими процесами, що створює конкурентні переваги.

У дослідженні були детально опрацьовані сильні та слабкі сторони проєкту, проведений SWOT-аналіз, а також аналіз конкурентного середовища та цільової аудиторії. Було визначено ключові сегменти користувачів, їхні потреби та канали комунікацій. На основі цього сформовано концепт маркетингової стратегії, який враховує як залучення нових клієнтів через

MVP, так і подальший розвиток через платні підписки, партнерства та вихід на міжнародні ринки.

Таким чином, MishBank має всі передумови для успішного позиціонування на ринку фінансових технологій, оскільки поєднує інноваційність, доступність, персоналізацію та високий рівень безпеки даних.

ВИСНОВКИ

Узагальнення виконаного дослідження демонструє високий потенціал інтелектуальних аналітичних систем у задачах управління клієнтським портфелем банку та прогнозування поведінки клієнтів. У роботі розроблено цілісний підхід, що поєднує попередню обробку банківських даних, імовірнісну сегментацію клієнтів та сегментно-орієнтоване прогнозування цільових фінансових подій. Така інтеграція методів дозволяє перейти від описового аналізу клієнтської бази до формування обґрунтованих управлінських рішень на основі даних.

Основним результатом роботи є створення інтелектуального інструментарію, у якому модель Gaussian Mixture Models використовується для виявлення прихованої структури клієнтського портфеля, а прогнозні моделі машинного навчання застосовуються для оцінювання ймовірності відкриття депозитів і залучення кредитів у межах сформованих сегментів. Імовірнісний характер сегментації дозволяє враховувати розмитість меж між групами клієнтів та більш точно відображати реальні поведінкові патерни, що є суттєвою перевагою порівняно з традиційними центричними методами кластеризації.

Експериментальна частина роботи підтвердила ефективність запропонованого підходу. Проведені дослідження показали, що використання сегментно-орієнтованого прогнозування забезпечує підвищення якості моделей порівняно з універсальними підходами без сегментації. Оцінювання результатів за допомогою комплексу метрик бінарної класифікації продемонструвало стабільність моделей та їх здатність коректно ідентифікувати клієнтів з високою ймовірністю цільової дії. Отримані результати підтверджують доцільність поєднання сегментації та прогнозування у межах єдиного аналітичного контуру.

Розроблений інструментарій має практичну цінність для банківських установ, оскільки може бути використаний для підтримки маркетингових,

кредитних та стратегічних рішень. Система дозволяє формувати пріоритетні списки клієнтів, оптимізувати витрати на комунікації, підвищувати ефективність персоналізованих пропозицій та зменшувати ризики, пов’язані з прийняттям управлінських рішень. Використання сучасних методів машинного навчання та модульної архітектури створює можливості для масштабування та адаптації системи до різних бізнес-сценаріїв.

Таким чином, результати дослідження підтверджують доцільність застосування імовірнісної сегментації та сегментно-орієнтованого прогнозування у задачах управління клієнтським портфелем банку. У роботі закладено науково-методичну та практичну основу для подальшого розвитку інтелектуальних систем підтримки прийняття рішень у фінансовій сфері. Перспективними напрямами подальших досліджень є розширення переліку прогнозованих подій, інтеграція часових моделей, використання глибоких нейронних мереж, а також впровадження системи в режимі обробки даних у реальному часі та її комерціалізація у вигляді окремого програмного продукту.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Payne A., Frow P. Customer Relationship Management: Strategy and Implementation. London: Financial Times Press, 2017. 206 p.
2. Tan P., Steinbach M., Kumar V. Introduction to Data Mining. Pearson Education, 2019. 352 p.
3. Murphy K. P. Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press, 2012. 302 p.
4. Breiman L. Random forests. *Machine Learning*. 2001. Vol. 45, No. 3.P. 5–32.
5. Friedman J. H. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*. 2001. Vol. 29, No. 9. P. 1189–1232.
6. Pedregosa F. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. 2011. No. 6. P. 2825–2830.
7. Hosmer D. W., Lemeshow S., Sturdivant R. X. Applied Logistic Regression. 3rd ed. Wiley, 2013. 270 p.
8. Hair J. F., Black W. C., Babin B. J. Multivariate Data Analysis. 8th ed. Cengage Learning, 2019. 170 p.
9. Kaufman L., Rousseeuw P. J. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 2005. 266 p.
10. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987. No. 7. P. 53–65.
11. Box G. E. P., Jenkins G. M., Reinsel G. C. Time Series Analysis: Forecasting and Control. 5th ed. Wiley, 2016. 420 p.
12. Gujarati D. N., Porter D. C. Basic Econometrics. 5th ed. McGraw-Hill, 2009. 370 p.
13. Hand D. J., Mannila H., Smyth P. Principles of Data Mining. Cambridge, MA: MIT Press, 2001. 220 p.
14. Provost F., Fawcett T. Data Science for Business: What You Need to Know about

Data Mining and Data-Analytic Thinking. O'Reilly Media, 2013. 250 p.

15. Buttle F., Maklan S. Customer Relationship Management: Concepts and Technologies. 4th ed. London: Routledge, 2019. 338 p.
16. Berry M. J. A., Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 3rd ed. Hoboken: Wiley, 2011. 420 p.
17. Shmueli G., Bruce P. C., Yahav I. Data Mining for Business Analytics: Concepts, Techniques, and Applications. Hoboken: Wiley, 2020. 336 p.
18. Jain A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 2010. Vol. 31, No.10. P. 651–666.
19. Міщенко А.С., Гуськова В.Г. Інтелектуальні системи для прогнозування у банківській сфері. Системні науки та інформатика: збірник доповідей IV Всеукраїнської науково-практичної конференції «Системні науки та інформатика», 01–05 грудня 2025 року, Київ. К., НН ІПСА КПІ ім. Ігоря Сікорського, 2025, С. 39-44.

ДОДАТОК А. ЛІСТИНГ ПРОГРАМИ

```
import os, warnings

warnings.filterwarnings("ignore")


import numpy as np

import pandas as pd

import matplotlib.pyplot as plt


from sklearn.model_selection import train_test_split

from sklearn.preprocessing import OneHotEncoder, StandardScaler

from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline

from sklearn.impute import SimpleImputer

from sklearn.mixture import GaussianMixture

from sklearn.decomposition import PCA

from sklearn.metrics import silhouette_score, accuracy_score, roc_auc_score, f1_score, precision_score, recall_score, confusion_matrix, classification_report


from sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier


# Шлях до датасету з GitHub

DATA_URL = "https://raw.githubusercontent.com/Antonimmus/mag/refs/heads/main/BankCustomerData.csv"


# Завантаження даних з посилання

df = pd.read_csv(DATA_URL)
```

```
print("Dataset shape:", df.shape)

print(df.head())


np.random.seed(RANDOM_STATE)

os.makedirs("outputs", exist_ok=True)



df = pd.read_csv(DATA_URL)

print(df.shape)

display(df.head())

print(df.dtypes)


# %%


print("Missing values per column:")

display(df.isna().sum().sort_values(ascending=False))


print("\nBasic describe (numeric):")

display(df.describe().T)


print("\nTop categories (categorical):")

cat_cols = [c for c in df.columns if df[c].dtype == "O"]

for c in cat_cols[:10]:

    print(f"\n[{c}]")

    print(df[c].value_counts().head(10))



# Task B: Term deposit (Target A previously) -> binary 1/0

assert "term_deposit" in df.columns, "term_deposit column not found"
```

```
y_term = (df["term_deposit"].str.strip().str.lower() == "yes").astype(int)


# Task C: Credit uptake -> from 'housing' or 'loan' (yes -> 1)

for col in ["housing","loan"]:

    assert col in df.columns, f"{col} column not found"

will_take_credit        =        df[["housing","loan"]].apply(lambda        s:
s.str.strip().str.lower().isin(["yes","1","true"]).astype(int)).max(axis=1)


# Task A: Stay/Churn proxy (constructed):

# Heuristic: a client "will_stay" if any of these is true:

# - subscribed term deposit (term_deposit=yes), or

# - positive or above-median balance, or

# - successful previous campaign (poutcome='success'), or

# - has at least one bank credit product (housing/loan yes).

balance_med = df["balance"].median() if "balance" in df.columns else 0

poutcome_success = (df.get("poutcome","").astype(str).str.lower() == "success")

has_any_credit = will_take_credit.eq(1)


will_stay = (

    (y_term.eq(1)) |

    (df.get("balance", pd.Series([0]*len(df))).fillna(0) >= balance_med) |

    (poutcome_success) |

    (has_any_credit)

).astype(int)


# Save targets into frame

df_targets = pd.DataFrame({

    "will_stay": will_stay,

    "term_deposit_bin": y_term,
```

```
    "will_take_credit": will_take_credit
})

print(df_targets.mean().rename("positive_rate"))
display(df_targets.head())


# Separate features (exclude explicit targets and obvious ID-like fields if exist)
exclude_cols = {"term_deposit"}  # original categorical target
feature_cols_all = [c for c in df.columns if c not in exclude_cols]

# For realistic models, we EXCLUDE 'duration'
feature_cols_realistic = [c for c in feature_cols_all if c != "duration"]
feature_cols_benchmark = feature_cols_all[:]  # keep duration

print("Feature count realistic:", len(feature_cols_realistic))
print("Feature count benchmark:", len(feature_cols_benchmark))

# Identify categorical vs numeric for preprocessing
def split_cols(cols):
    cat = [c for c in cols if df[c].dtype == "O"]
    num = [c for c in cols if c not in cat]
    return num, cat

num_real, cat_real = split_cols(feature_cols_realistic)
num_bench, cat_bench = split_cols(feature_cols_benchmark)

def make_preprocessor(num_cols, cat_cols, max_categories=80):
```

```
    numeric_transformer = Pipeline(steps=[
        ("imputer", SimpleImputer(strategy="median")),
        ("scaler", StandardScaler())
    ])
    categorical_transformer = Pipeline(steps=[
        ("imputer", SimpleImputer(strategy="most_frequent")),
        ("onehot", OneHotEncoder(handle_unknown="ignore", sparse_output=False,
max_categories=max_categories))
    ])
    preprocessor = ColumnTransformer(
        transformers=[
            ("num", numeric_transformer, num_cols),
            ("cat", categorical_transformer, cat_cols)
        ]
    )
    return preprocessor


prep_real = make_preprocessor(num_real, cat_real)
prep_bench = make_preprocessor(num_bench, cat_bench)



def fit_gmm_and_profile(df, feature_cols, preprocessor, k_range=range(2,11),
random_state=42):
    X = df[feature_cols]
    Xs = preprocessor.fit_transform(X)

    sse, sil = [], []
    best_k = None
    best_sil = -1
    for k in k_range:
```

```
        gmm = GaussianMixture(n_components=k, covariance_type="full",
random_state=random_state)
        labels = gmm.fit_predict(Xs)
        sse.append(np.sum((Xs - gmm.means_[labels])**2))
        sc = silhouette_score(Xs, labels)
        sil.append(sc)
        if sc > best_sil:
            best_sil = sc
            best_k = k

    print("Silhouette by k:", dict(zip(list(k_range), np.round(sil,3))))
    print("Chosen k:", best_k)

    gmm = GaussianMixture(n_components=best_k, covariance_type="full",
random_state=random_state)
    clusters = gmm.fit_predict(Xs)

    # attach clusters
    seg_df = df.copy()
    seg_df["cluster"] = clusters

    # profile (numeric means, and top categories for a few key cat columns)
    num_cols = [c for c in feature_cols if df[c].dtype != "O"]
    prof_num = seg_df.groupby("cluster")[num_cols].mean()

    cat_cols = [c for c in feature_cols if df[c].dtype == "O"]
    top_cat = {}
    for c in cat_cols:
        top_cat[c] = seg_df.groupby("cluster")[c].agg(lambda s:
s.value_counts().index[0])
```

```
    # PCA for visualization
    pca = PCA(n_components=3, random_state=random_state)
    X_pca = pca.fit_transform(preprocessor.transform(df[feature_cols]))

    return seg_df, prof_num, pd.DataFrame(top_cat), X_pca, clusters, best_k, (sse, sil)

# Use REALISTIC features for segmentation
seg_df, prof_num, prof_cat, X_pca, clusters, best_k, (sse, sil) = fit_gmm_and_profile(
    df, feature_cols_realistic, prep_real, k_range=range(2,11), random_state=RANDOM_STATE
)

# Plots
plt.figure(figsize=(6,4))
plt.plot(range(2,11), sse, marker="o")
plt.title("Elbow (SSE)")
plt.xlabel("k"); plt.ylabel("SSE"); plt.show()

plt.figure(figsize=(6,4))
plt.plot(range(2,11), sil, marker="o")
plt.title("Silhouette")
plt.xlabel("k"); plt.ylabel("Silhouette"); plt.show()

plt.figure(figsize=(7,6))
plt.scatter(X_pca[:,0], X_pca[:,1], c=clusters, s=10)
plt.title(f"PCA of clients (k={best_k})")
plt.xlabel("PC1"); plt.ylabel("PC2"); plt.show()
```

```
print("\nNumeric profile by cluster (means):")

display(prof_num)


print("\nTop categorical values by cluster:")

display(prof_cat)


# Attach targets to compute per-segment conversion/credit/stay rates

seg_rates = seg_df.join(pd.DataFrame({

    "will_stay": ( (df["term_deposit"].str.lower()=="yes").astype(int) |

                   (df.get("balance",0) >= df["balance"].median()).astype(int)
|

                   (df.get("poutcome","").str.lower()=="success").astype(int)
|

                   ((df["housing"].str.lower()=="yes")                          |
(df["loan"].str.lower()=="yes")).astype(int) ).astype(int),

    "term_deposit_bin": (df["term_deposit"].str.lower()=="yes").astype(int),

    "will_take_credit":          ((df["housing"].str.lower()=="yes")          |
(df["loan"].str.lower()=="yes")).astype(int)

}))


segment_summary                                                              =
seg_rates.groupby("cluster")[["will_stay","term_deposit_bin","will_take_credit
"]].mean().rename(columns={


"will_stay":"rate_stay","term_deposit_bin":"rate_deposit","will_take_credit":"
rate_credit"

})

print("\nSegment conversion/credit/stay rates:")

display(segment_summary)


# Simple actionable hints per segment (heuristic text rules)
```

```
def segment_hint(row):
    hints = []
    if row["rate_deposit"] < 0.1:
        hints.append("Try awareness nudges; A/B test shorter scripts and benefits framing.")
    elif row["rate_deposit"] < 0.25:
        hints.append("Focus on incentives (fee waivers, small bonus rates); retarget warm leads.")
    else:
        hints.append("Double down on cross-sell; personalized rate offers and referral programs.")

    if row["rate_credit"] > 0.4:
        hints.append("Offer bundled credit+deposit packages; pre-approved credit lines.")
    elif row["rate_credit"] < 0.15:
        hints.append("Promote low-limit starter cards or education content; build trust first.")

    if row["rate_stay"] < 0.5:
        hints.append("Retention: proactive outreach, service recovery, check satisfaction; loyalty perks.")
    else:
        hints.append("Retention solid; prioritize CLV growth via add-on products.")
    return " ".join(hints)

actionable = segment_summary.copy()
actionable["hint"] = actionable.apply(segment_hint, axis=1)
print("\nActionable suggestions per segment:")
display(actionable)
```

```
# Save profiles
prof_num.to_csv("outputs/cluster_profile_numeric.csv")
prof_cat.to_csv("outputs/cluster_profile_categorical.csv")
segment_summary.to_csv("outputs/cluster_rates.csv")
actionable.to_csv("outputs/cluster_actionable.csv")




# %%


def evaluate_models(X, y, preprocessor, random_state=42, label=""):
    models = {
        "LogReg": LogisticRegression(max_iter=400),
        "RandomForest":               RandomForestClassifier(n_estimators=400,
random_state=random_state),
        "GradBoost": GradientBoostingClassifier(random_state=random_state)
    }
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,
random_state=random_state, stratify=y)
    rows = []
    for name, model in models.items():
        pipe = Pipeline(steps=[("prep", preprocessor), ("model", model)])
        pipe.fit(X_train, y_train)
        preds = pipe.predict(X_test)
        proba    =    pipe.predict_proba(X_test)[:,1]    if    hasattr(pipe,
"predict_proba") else None


        acc = accuracy_score(y_test, preds)
        f1 = f1_score(y_test, preds)
```

```
        prec = precision_score(y_test, preds)
        rec = recall_score(y_test, preds)
        auc = roc_auc_score(y_test, proba) if proba is not None else np.nan

        print(f"\n=== {label} :: {name} ===")
        print(classification_report(y_test, preds, digits=3))
        print("Confusion matrix:\n", confusion_matrix(y_test, preds))

        rows.append({
            "task": label, "model": name, "accuracy": acc, "f1": f1,
            "precision": prec, "recall": rec, "roc_auc": auc,
            "y_test_mean": float(np.mean(y_test))
        })
    return pd.DataFrame(rows)

def run_task(task_name, y, feature_cols, preprocessor, realistic=True):
    X = df[feature_cols]
    label = f"{task_name} ({'realistic' if realistic else 'benchmark_with_duration'})"
    res = evaluate_models(X, y, preprocessor, random_state=RANDOM_STATE, label=label)
    return res


results = []

results.append(run_task("A_will_stay", will_stay, feature_cols_realistic, prep_real, realistic=True))
results.append(run_task("A_will_stay", will_stay, feature_cols_benchmark, prep_bench, realistic=False))
```

```
results.append(run_task("B_term_deposit",    y_term,    feature_cols_realistic,
prep_real, realistic=True))

results.append(run_task("B_term_deposit",    y_term,    feature_cols_benchmark,
prep_bench, realistic=False))


results.append(run_task("C_will_take_credit",                will_take_credit,
feature_cols_realistic, prep_real, realistic=True))

results.append(run_task("C_will_take_credit",                will_take_credit,
feature_cols_benchmark, prep_bench, realistic=False))


res_all = pd.concat(results, ignore_index=True)

print("\nAll results (sorted by task, roc_auc desc):")

display(res_all.sort_values(["task","roc_auc"], ascending=[True,False]))


res_all.to_csv("outputs/model_results.csv", index=False)
```

ДОДАТОК Б. ПРЕЗЕНТАЦІЯ

01

ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ДЛЯ ПРОГНОЗУВАННЯ У БАНКІВСЬКІЙ СФЕРІ

ст. викладач кафедри ММСА,
д. філософії, Гуськова В.Г.
студень групи КА-41 МП
Міщенко Антон

02

МЕТА ТА МЕТОДИ ДОСЛІДЖЕННЯ

Мета дослідження – розробити інтелектуальну систему прогнозує ймовірність вибору банківських продукті автоматично сегментує клієнтів для підвищення ефективності маркетингових рішень.

Методи дослідження – машинне навчання (Random For KMeans), статистичний аналіз, попередня обробка методи зниження розмірності, а також візуалізація моделей за метриками (Silhouette Score, точність, ймовірнісні оцінки).

03

ОБ'ЄКТ ТА ПРЕДМЕТ ДОССЛІДЖЕННЯ

Об'єктом дослідження є процес підтримки прийн рішень у банківських установах під час прогнозу взяття ключових продуктів.

Предметом дослідження є методи машинного на сегментації клієнтів, оцінки їхніх ймовірностей вз банківськими продуктами та побудови персоналі рекомендацій.

04

АКТУАЛЬНІСТЬ ТЕМИ

Сучасний банківський сектор працює в умовах високої конкуренції, зростання ризиків та швидкої цифровізації. Застосування інтелектуальних систем та алгоритмів машинного навчання дозволяє прогнозувати поведінку клієнтів, визначати ризики, оптимізувати маркетингові стратегії та формувати персоналізовані пропозиції. Це підвищує якість управлінських рішень і забезпечує конкурентні переваги банківської установи.

05

АКТУАЛЬНІСТЬ ТЕМИ

Приблизна ціна залучення нового клієнта становить ~ 600-1500 грн.
Ціна комунікації з діючим клієнтом через застосунок, SMS повідомлення, Viber ~ 3 гривні.

06

АРХІТЕКТУРА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ПРОГНОЗУВАННЯ КЛІЄНТСЬКОГО ПОРТФЕЛЮ

07

ДАНІ ТА ЇХ ОБРОБКА

Для дослідження використано набір даних із соціальних, демографічних та фінансових характеристик клієнтів.

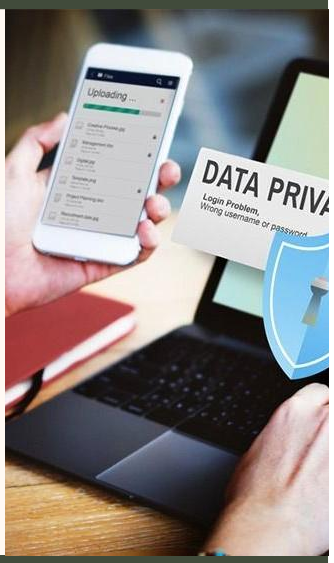
Реалізовано:

– Імпорт даних з обробкою помилок;
– Генерацію внутрішнього CLIENT_ID;
– Анонімізацію даних (маскування ID);
– Перевірку якості даних;
– Формування фінальної матриці ознак для моделювання.

Такий підхід забезпечує безпеку, коректність і відтворюваність аналітичного процесу.

08

ДАНІ ТА ЇХ ОБРОБКА



	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	term deposit
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

A
client_id_masked
796****4437
494****5957
872****1279
727****2261

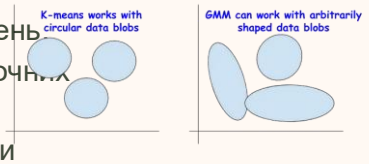
	count	mean	std	min	25%	50%	75%	max
age	42639.0	40.788808	10.200236	18.0	33.0	39.0	48.0	95.0
balance	42639.0	1331.863951	3011.537676	-8019.0	62.0	429.0	1381.5	102127.0
day	42639.0	15.854781	8.293901	1.0	8.0	16.0	21.0	31.0
duration	42639.0	255.957504	258.361368	0.0	101.0	177.0	315.0	4918.0

09

СЕГМЕНТАЦІЯ КЛІЄНТІВ

Для сегментації клієнтів використано алгоритм Gaussian mixture models:

- Ініціалізація параметрів моделі: початкові значення середніх векторів, коваріаційних матриць компонент.
- Обчислення ймовірностей належності: для кожного клієнта визначається ймовірність належності кожної гауссової компоненти.
- Оцінювання параметрів розподілів: уточнення середніх значень, коваріаційних матриць та ваг компонент на основі поточних ймовірностей належності.
- Повторення ітераційного процесу (Expectation-Maximization), доки параметри моделі не стабілізуються або не буде досягнуто критерію зупинки.



10

SILHOUETTE SCORE

Silhouette Score оцінює якість кластеризації, порівнюючи, наскільки точка близька до свого кластера та віддалена від інших.

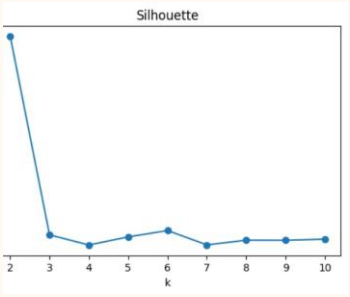
Для кожної точки x_i обчислюється **a(i)** – середня відстань до всіх точок у її кластері, та **b(i)** – мінімальна середня відстань до точок найближчого іншого кластера.

На основі цих величин формується показник силуету точки:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

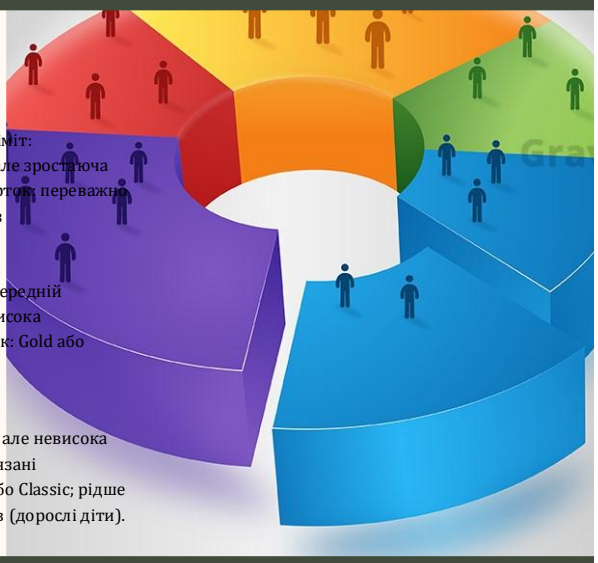
Значення **s(i)** близьке до **1** означає добре сформований кластер, близьке до **0** – перетин кластерів, а від'ємне – неправильне віднесення точки.

Загальний **Silhouette Score** – це середнє значення **s(i)** для всіх точок, яке служить критерієм вибору оптимальної кількості кластерів.



11

ПРИКЛАД СЕГМЕНТАЦІЇ



Вік: молодший за середній (переважно 20–35 років). • Рівень доходу та кредитний ліміт: невисокий, середні залишки на рахунках. • Транзакційна активність: середня, але зростаюча клієнти активно використовують картки для щоденних витрат. • Категорії карток: переважно стандартні або базові (Classic). • Сімейний стан: неодружені або молоді сім'ї з невеликою кількістю утриманців.

Вік: середній (30–50 років). • Фінансові параметри: стабільні доходи, вищий за середній кредитний ліміт, помірні залишки на рахунках. • Транзакційна активність: висока користуються як кредитними, так і дебетовими картками. • Категорії карток: Gold або Platinum. • Сімейний стан: переважно одружені, мають утриманців.

Вік: старший (50+ років). • Фінансові параметри: високі залишки на рахунках, але невисока потреба у кредитах. • Транзакційна активність: низька, переважно операції, пов'язані зі зняттям готівки або депозитними операціями. • Категорії карток: Standard або Classic; рідше преміальні. • Сімейний стан: стабільний, велика частка клієнтів без утриманців (дорослі діти).

12

ПОБУДОВА МОДЕЛЕЙ

У роботі були побудовані дві окремі моделі прогнозування:
—модель відкриття депозиту,
—модель взяття кредиту.
Для порівняння ефективності застосовувалися три алгоритми машинного навчання: Random Forest, Gradient Boosting та L
Regression. Random Forest був обраний як основна модель завдяки високій точності, стійкості до шуму та здатності відтво
нелінійні залежності. Gradient Boosting забезпечив альтернативний підхід із підвищеною чутливістю до складних патернів,
Regression виступала базовою інтерпретованою моделлю для формування порівняльного аналізу.

Отримані ймовірності використовувалися для визначення цільових груп у кожному сегменті та подальшого формування р

13

ПОБУДОВА МОДЕЛЕЙ

Архітектура побудованих моделей включає основний обробний конвеєр (pipeline), який складається з таких етапів:
Імпутація пропусків
— числові значення медіанна імпутація;
— категоріальні ознаки імпутація найчастішим значенням
Стандартизація числових ознак:

$$x^* = \frac{x - \mu}{\sigma}$$

Кодування категоріальних атрибутів:
— за допомогою One-Hot Encoding з обмеженням max_categories.
Навчання моделі:
— Random Forest / Gradient Boosting / Logistic Regression.
Побудова ймовірнісного прогнозу:
— вихід моделі — ймовірність P(y=1)P(y=1)P(y=1).

$$X \xrightarrow{\text{Preprocessing}} X^* \xrightarrow{\text{ML Model}} \hat{y}$$

Архітектура реалізована через конвеєр:

14

ПОРІВНЯННЯ МОДЕЛЕЙ

task	model	accuracy	f1	precision	recall	roc_auc	y_test_mean
B_term_deposit	RandomForest	0.918386	0.441592	0.605634	0.347475	0.928154	0.092871
B_term_deposit	GradBoost	0.920732	0.466877	0.621849	0.373737	0.927015	0.092789
B_term_deposit	LogReg	0.915103	0.375431	0.592593	0.274747	0.905942	0.09277
C_term_deposit	RandomForest	0.90985	0.2206	0.559671	0.137374	0.741186	0.092971
C_term_deposit	LogReg	0.908724	0.147239	0.556291	0.084848	0.721723	0.09251
C_term_deposit	GradBoost	0.911538	0.202874	0.621762	0.121212	0.749031	0.092971

15

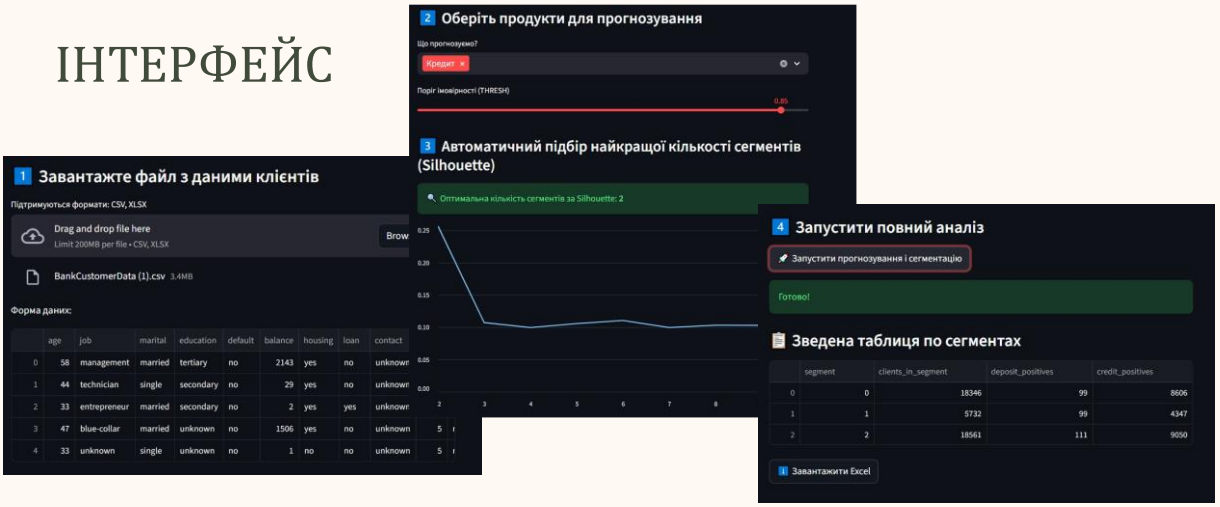
РЕЗУЛЬТАТ ПРОГНОЗУВАННЯ

	segment	clients_in_segment	deposit_positives	credit_positives
2	0	18346	99	8606
3	1	5732	99	4347
4	2	18561	111	9050

	client_id_masked	kmeans_segment	proba_deposit	pred_deposit	proba_credit	pred_credit
2	567****5114	0	0,0075	0	0,9525	1
3	794****4867	0	0,0175	0	0,9575	1
4	392****3003	0	0,01	0	0,925	1
5	604****5063	0	0,01	0	0,97	1
6	542****3008	0	0,005	0	0,99	1
7	825****2652	0	0,02	0	0,985	1
8	314****7964	0	0,005	0	0,9825	1
9	988****1230	0	0,005	0	0,9525	1
10	903****9034	0	0,03	0	0,925	1
11	514****7364	0	0,005	0	0,945	1
12	736****3291	0	0	0	0,9625	1
13	245****2120	0	0,0025	0	1	1

16

ІНТЕРФЕЙС



17

ВИСНОВКИ

Використання ансамблевих моделей дозволило значно підвищити точність прогнозування поведінки клієнтів.

Random Forest та Gradient Boosting показали найкраще співвідношення між точністю та узагальнюючою здатністю.

Комбінація моделей із сегментацією забезпечила комплексний підхід: від прогнозу до формування індивідуальних стратегій роботи з клієнтами.

Такий підхід підвищує ефективність маркетингових кампаній та дозволяє банку зосередитися на найбільш перспективних клієнтах.

18

НАУКОВА ТА ПРАКТИЧНА НОВИЗНА

У роботі ідля даного банківського кейсу реалізовано інтегровану систему, яка одночасно прогнозує ймовірність взяття депозиту/кредиту та автоматично сегментує клієнтів, поєднуючи обидва підходи в єдиному конвеєрі

Реалізовано механізм автоматичного визначення оптимальної кількості кластерів за метрикою Silhouette Score, що забезпечує стабільний і об'єктивний результат сегментації без ручного втручання.

Розроблено готовий Streamlit-додаток, який може використовуватися бізнес-підрозділами для швидкого аналізу клієнтської бази, формування їхніх сегментів та побудови таргетованих кампаній без участі аналітика

У моделюванні застосовано декілька алгоритмів (RandomForest, GradienBoosting, LogistiRegression) з порівнянням їхньої продуктивності та обґрунтованим вибором найефективнішої для банківської задачі

Запропоновано механізм автоматичної інтерпретації, стандартизації та обробки даних клієнтів із приховуванням персональних ідентифікаторів, що дозволяє масштабувати систему у реальні банківські процеси.

19

ДЯКУЮ

ЗА УВАГУ