

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

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
Кафедра математичного моделювання та статистики**

**Освітньо-професійна програма
Галузь знань
Спеціальність**

**Економічна кібернетика
05 Соціальні та поведінкові науки
051 Економіка**

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему **«Моделювання ефективності контакту з клієнтом банку у продажі депозитного продукту»**

(назва теми)

здобувача Красноусової Поліни Олегівни

(ПІБ, підпис)

Науковий керівник: доктор економічних наук,
професор, Матвійчук А. В.
(науковий ступінь, учене звання, ПІБ)

(підпис)

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

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

(підпис)

Київ 2023

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

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
Кафедра математичного моделювання та статистики**

Освітньо-професійна програма
Галузь знань
Спеціальність

Економічна кібернетика
05 Соціальні та поведінкові науки
051 Економіка

ПОГОДЖЕНО

Керівник проектної групи (гарант)
освітньо-професійної програми
О. М. Притоманова

(підпис)

20__ р.

ЗАТВЕРДЖУЮ

Завідувач кафедри математичного
моделювання та статистики
Г. І. Великоіваненко

(підпис)

20__ р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

Здобувача вищої освіти Красноусової Поліни Олегівни

очної (денної) форми навчання

на підготовку кваліфікаційної бакалаврської роботи

**на тему «Моделювання ефективності контакту з клієнтом банку у продажі
депозитного продукту»**

Темі затверджено наказом ректора Університету від "20" березня 2023 р. №

**Кваліфікаційна бакалаврська робота виконується на матеріалах
з відкритих джерел UCI Learning Repository**

План кваліфікаційної бакалаврської роботи

Розділ 1	Дослідження поточного стану питання аналізу та підвищення ефективності маркетингових кампаній банківських установ
Розділ 2	Формування інформаційної бази та попередня обробка даних

Розділ 3	Комплекс моделей оцінювання ефективності контакту з клієнтами банку під час маркетингової кампанії з продажу депозитного продукту
Об'єкт дослідження:	Маркетингова кампанія комерційного банку з продажу депозитного продукту.
Предмет дослідження:	Моделювання ефективності контакту з клієнтом банку з продажу депозитного продукту.
Мета кваліфікаційної бакалаврської роботи:	Розробка математичного інструментарію передбачення ефективності контакту з клієнтами банку задля підвищення прибутковості маркетингових кампаній з продажу депозитних продуктів.

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1	Дослідити поточний стан маркетингових кампаній в банківських установ та проаналізувати методи підвищення їх ефективності.
У розділі 2	Провести збір даних щодо об'єкта дослідження, їх попередню обробку, детальний дослідницький аналіз та категоризацію (біннінг).
У розділі 3	Розробити математичний інструментарій для передбачення ефективності контакту з клієнтами банку під час маркетингової кампанії з продажу депозитних продуктів.

**Завдання підготував
науковий керівник**

(підпис)

А. В. Матвійчук

(ініціали, прізвище)

« ____ » _____ 20__ р.

**Завдання одержав
здобувач**

(підпис)

П. О. Красноусова

(ініціали, прізвище)

« ____ » _____ 20__ р.

РЕФЕРАТ

Кваліфікаційна бакалаврська робота містить 77 сторінки, 3 формули, 42 рисунки, список використаних джерел з 34 найменувань, додатки.

«Моделювання ефективності контакту з клієнтом банку у продажі депозитного продукту»

Відомо, що залучення нових клієнтів та утримання старих – є одним із найважливіших завдань, що з'являються у роботі кожної організації, компанії, установи. Різні великі організації прагнуть до максимізації прибутку та оптимізації своєї діяльності. Задоволення клієнтських потреб, надання якісного продукту та якісного сервісу – це запорука високих фінансових показників та стабільності будь-якої установи. За останнє століття банківська галузь на глобальному рівні зазнала значних трансформацій.

У даній дипломній роботі доведена актуальність та важливість аналізу клієнтів, розуміння їх особливостей, потреб та намірів. Дослідження проведено на прикладі португальської банківської установи під час маркетингової кампанії з продажу депозитних продуктів. Актуальність праці полягає у вивченні закономірностей та трендів різних клієнтських сегментів з метою підвищення ефективності роботи банку та мінімізації маркетингових витрат. Банки витрачають великі суми коштів на створення та запуск кампаній з метою продажі продуктів, тому надзвичайно важливо розуміти, на яку групу клієнтів зосереджувати більшу частину ресурсів установи з метою уникнення марнування часу та фінансів.

Метою даної роботи є розробка математичного інструментарію передбачення ефективності контакту з клієнтами банку задля підвищення прибутковості маркетингових кампаній з продажу депозитних продуктів.

Досягнення поставленої мети обумовлює необхідність вирішення *таких завдань:*

- дослідити сучасний стан ринку банківських послуг;

- здійснити критичний аналіз та узагальнення основних підходів до проведення маркетингових кампаній в банківській сфері, виявити їх недоліки та переваги;
- дослідити математичний інструментарій побудови моделей аналізу поведінки споживачів;
- запропонувати методологічний підхід до моделювання ефективності контакту з клієнтом банку;
- побудувати комплекс економіко-математичних моделей в рамках розробленого методологічного підходу;
- реалізувати побудовані моделі у програмному середовищі та провести експериментальне дослідження їх адекватності на основі реальних даних;
- сформулювати висновки та пропозиції щодо ефективності різних етапів побудови моделей та обґрунтувати вибір найкращих з них.

Об'єктом дослідження є маркетингова кампанія комерційного банку з продажу депозитних продуктів.

Предметом роботи є моделювання ефективності контакту з клієнтом банку у продажі депозитних продуктів.

Для досягнення мети будуть використані такі методи, як: балансування вибірки, відбір значущих факторів, детальний дослідницький аналіз, категоризація даних (біннінг) та логістична регресія.

Рік виконання кваліфікаційної бакалаврської роботи – 2023.

Рік захисту роботи – 2023.

Ключові слова: маркетингова кампанія, банківська установа, ефективність контакту, балансування вибірки, категоризація (біннінг), відбір значущих факторів, логістична регресія.

В і д г у к

про кваліфікаційну бакалаврську роботу
здобувача навчально-наукового інституту /факультету
освітньо-професійної програми « Економічна кібернетика»

Красноусової П.О

на тему: «Модельовання ефективності контакту з клієнтом банку у продажі
депозитного продукту»

1. Актуальність теми полягає у вивченні закономірностей та трендів різних клієнтських сегментів з метою підвищення ефективності роботи банку та мінімізації маркетингових витрат. Банки витрачають великі суми коштів на створення та запуск кампаній з метою продажі продуктів, тому надзвичайно важливо розуміти, на яку групу клієнтів зосереджувати більшу частину ресурсів установи з метою уникнення марнування часу та фінансів.

2. До позитивних рис кваліфікаційної бакалаврської роботи Красноусової П.О. можна віднести, зокрема те, що автором здійснено досить глибокий аналіз сучасного стану витрат банківських установ на маркетингові кампанії. Було досліджено різні види банківських організацій, а також проаналізована специфіка маркетингових комунікацій в банках. Красноусова П.О. зробила детальний дослідницький аналіз та обробку бази даних, що допомогло краще зрозуміти основні тенденції серед клієнтської поведінки. Проведено порівняльний аналіз восьми логістичних регресій, що були побудовані із застосуванням різних комбінацій таких методів: балансування вибірки, біннінгу даних, відбору значущих факторів та ін. Даний аналіз дозволив обрати найбільш адекватний інструментарій для вирішення поставленої задачі.

3. Наявність самостійних розробок автора. Автором бакалаврської роботи проведено категоризацію (біннінг) бази даних португальської банківської установи, проведено балансування вибірки для покращення результатів роботи логістичної регресії. Методом покрокового відбору значущих факторів визначено змінні, які необхідно включити до моделі. Практична реалізація моделей була здійснена за допомогою сучасних інструментів, а саме мови програмування Python

та спеціалізованих бібліотек для обробки даних та моделювання – Pandas, Sckit-learn тощо. Розроблений комплекс моделей оцінювання ефективності контакту з клієнтами банку успішно пройшов тестування на реальних даних.

4. Цінність теоретичних висновків та практичних рекомендацій визначається проведеним детальним дослідницьким аналізом, розробленим методологічним підходом до моделювання ефективності контакту з клієнтом банку та побудованим комплексом моделей на основі логістичної регресії. Визначено основні характеристики клієнтів банку (вік, освіта, кредитна історія, фактори, що пов'язані з минулими маркетинговими кампаніями), які визначають високу ймовірність успішності контакту з клієнтом. У результаті цих досліджень надані рекомендації щодо наступних маркетингових кампаній, сегментів клієнтів, на які треба зосереджувати більшу частину ресурсів банку, тощо.

5. Наявність недоліків: незначна кількість граматичних та синтаксичних помилок, що, тим не менш, не знижує наукову та практичну цінність виконаної роботи.

6. Загальна оцінка кваліфікаційної бакалаврської роботи та її допущення до захисту перед ЕК: Вважаю, що Красноусова П.О. в результаті виконання кваліфікаційної роботи продемонструвала високий рівень самостійності та відповідальності, глибокі знання з обраної спеціалізації, вміння аналізувати та вирішувати поставлені завдання. Кваліфікаційна робота заслуговує високої оцінки, а Красноусова П.О. – присвоєння кваліфікації «бакалавра» з економічної кібернетики.

Науковий керівник:

Професор кафедри математичного моделювання та статистики,
д.е.н., професор Матвійчук А. В.

“ ____ ” _____ 20__ р.

ЗМІСТ

ВСТУП	3
РОЗДІЛ 1. ДОСЛІДЖЕННЯ ПОТОЧНОГО СТАНУ ПИТАННЯ АНАЛІЗУ ТА ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ МАРКЕТИНГОВИХ КАМПАНІЙ БАНКІВСЬКИХ УСТАНОВ.....	7
1.1 Банківська система: поняття, види та функції. Банківський депозит	7
1.2 Специфіка маркетингових комунікацій в банківських установах	14
1.2.1 Маркетингові комунікації	14
1.2.2 Аналіз маркетингових витрат українських банків	15
1.3 Методологічний підхід до моделювання ефективності контакту з клієнтом банку	18
РОЗДІЛ 2. ФОРМУВАННЯ ІНФОРМАЦІЙНОЇ БАЗИ ТА ПОПЕРЕДНЯ ОБРОБКА ДАНИХ	30
2.1 Підготовка та опис даних	30
2.2 Детальний дослідницький аналіз даних	32
2.3 Категоризація даних (біннінг).....	45
РОЗДІЛ 3. КОМПЛЕКС МОДЕЛЕЙ ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ КОНТАКТУ З КЛІЄНТАМИ БАНКУ ПІД ЧАС МАРКЕТИНГОВОЇ КАМПАНІЇ З ПРОДАЖУ ДЕПОЗИТНОГО ПРОДУКТУ	50
3.1 Побудова логістичної регресії на «сирих» даних.....	50
3.2 Логістична регресія на збалансованій вибірці.....	55
3.3 Побудова логістичних регресій із застосуванням процедур відбору значущих факторів ..	59
3.3.1 Backward Feature Selection із збалансованою вибіркою.....	59
3.3.2 Forward Feature Selection із збалансованою вибіркою	61
3.4 Оптимізація біннінгу даних та вибір найкращої моделі	62
3.4.1 Покращення біннінгу даних за допомогою параметра «cat_cutoff»	62
3.4.2 Узагальнення результатів та вибір найкращої моделі	64
ВИСНОВКИ	66
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	69
ДОДАТКИ	73

ВСТУП

Відомо, що залучення нових клієнтів та утримання старих – є одним із найважливіших завдань, що з'являються у роботі кожної організації, компанії, установи. Різні великі організації прагнуть до максимізації прибутку та оптимізації своєї діяльності. Задоволення клієнтських потреб, надання якісного продукту та якісного сервісу – це запорука високих фінансових показників та стабільності будь-якої установи.

За останнє століття банківська галузь на глобальному рівні зазнала значних трансформацій. Ці зміни включають реструктуризацію, модернізацію та автоматизацію банківських операцій, а також злиття банків і створення великих банківських установ з акцентом на маркетинг. У даній дипломній роботі доведена актуальність та важливість аналізу клієнтів, розуміння їх особливостей, потреб та намірів. Дослідження проведено на прикладі португальської банківської установи під час маркетингової кампанії з продажу депозитних продуктів. Актуальність праці полягає у вивченні закономірностей та трендів різних клієнтських сегментів з метою підвищення ефективності роботи банку та мінімізації маркетингових витрат. Банки витрачають великі суми коштів на створення та запуск кампаній з метою продажі продуктів, тому надзвичайно важливо розуміти, на яку групу клієнтів зосереджувати більшу частину ресурсів установи з метою уникнення марнування часу та фінансів.

Наразі основною проблемою у сфері маркетингу є обмежена ефективність і брак відповідальності. Шет і Сісодія у своїй роботі [1] пропонують стратегії для підвищення продуктивності маркетингу. У статті вони вказують, що глобальною проблемою маркетингу на сьогоднішній день є низька продуктивність та нестача підзвітності. Автори наголошують на два способи підвищення ефективності маркетингових кампаній: по-перше, зміщення фокусу з сукупних ринків на більш індивідуальних клієнтів; по-друге, розглядати маркетинг як виробничу функцію, а не як щорічні витрати.

Автори Кулі, Грінлі, Кадоган та Фейхі у статті [2] фокусуються на ролі маркетингових ресурсів у створенні конкурентних переваг та підвищенні загальної ефективності компанії. Результати їхнього дослідження демонструють, що маркетингові ресурси мають опосередкований вплив на фінансові показники, сприяючи задоволеності, лояльності та вищій ринковій ефективності.

У статті «Оцінка кількісних показників маркетингової діяльності в банківському секторі» [3] розглядається ситуація на словацькому банківському ринку. Автори цієї статті виокремили основну ціль свого дослідження: підкреслити важливість інновацій та адаптації традиційних підходів до управління маркетингом у відповідь на турбулентні події та зростаючу глобальну конкуренцію. У науковій роботі також наголошується на постійному управлінні маркетингом з трьох точок зору: економічної, соціальної та екологічної. Розглядаючи економічну сторону питання, предметом дослідження є зосередження ресурсів на ефективному маркетинговому управлінні задля стабільності позицій банківської установи на конкурентному ринку. Метою аналізу показників словацького банку є оцінка ефективності маркетингових кампаній, використовуючи кількісні показники, спрямованих на стабілізації позицій банку на конкурентному ринку, а також підвищення лояльності клієнтів.

Питанням дослідження ефективності маркетингових кампаній займався ряд закордонних та вітчизняних вчених, такі як: Кризанова, Гаянова, Клістікова [4], Кумар та Міттал [5] та інші. Незважаючи на це, залишилось недослідженим питання оцінювання ефективності контакту з клієнтами банку під час маркетингової кампанії з продажу депозитного продукту.

Метою даної роботи є розробка математичного інструментарію передбачення ефективності контакту з клієнтами банку задля підвищення прибутковості маркетингових кампаній з продажу депозитних продуктів.

Досягнення поставленої мети обумовлює необхідність вирішення таких завдань:

- дослідити сучасний стан ринку банківських послуг;

- здійснити критичний аналіз та узагальнення основних підходів до проведення маркетингових кампаній в банківській сфері, виявити їх недоліки та переваги;
- дослідити математичний інструментарій побудови моделей аналізу поведінки споживачів;
- запропонувати методологічний підхід до моделювання ефективності контакту з клієнтом банку;
- побудувати комплекс економіко-математичних моделей в рамках розробленого методологічного підходу;
- реалізувати побудовані моделі у програмному середовищі та провести експериментальне дослідження їх адекватності на основі реальних даних;
- сформулювати висновки та пропозиції щодо ефективності різних етапів побудови моделей та обґрунтувати вибір найкращих з них.

Об'єктом дослідження є маркетингова кампанія комерційного банку з продажу депозитних продуктів.

Предметом роботи є моделювання ефективності контакту з клієнтом банку у продажі депозитних продуктів.

Можна виокремити такі методи дослідження:

- Метод аналізу використовувався для розуміння актуальності та проблематики теми роботи, а також вивчення клієнтської бази з метою глибшого розуміння потреб та цінностей клієнтів;
- Метод детального дослідницького аналізу використовувався для глибоко розуміння бази даних, визначення основних закономірностей серед поведінки клієнтів банку та аналізу тенденцій серед продажів депозитного продукту;
- Логістична регресія використовувалась для побудови моделі класифікації клієнтів, які більш схильні до купівлі запропонованого депозитного продукту, та клієнтів, які скоріше відмовляться від даної пропозиції;

- Метод біннінгу, відбору факторів до моделі та балансування вибірки використовувались для покращення якості класифікаційної моделі та підвищення ефективності результатів.

У даній роботі досліджуються комбінації факторів, що відповідають за профіль кожного клієнта. На основі відбору факторів, що найпомітніше впливають на результати ефективності, будується модель. Побудована модель виступає в ролі класифікатора та передбачає позитивний або негативний (продаж товару клієнту або ні) результат для досліджуваної банківської установи.

Інформаційною базою дослідження слугували дані португальської банківської установи. База даних містить різні фактори, що описують клієнтів банку, наприклад вік, освіта, сімейний стан, наявність кредитів та інші. Джерела інформаційної бази наведені у розділі «Список використаних джерел», що знаходиться в кінці даної дипломної роботи і включають перелік наукових статей, інтернет-джерел, баз даних тощо.

Дана кваліфікаційна робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків. Усі розрахунки здійснювались на мові програмування Python у середовищі Jupyter Notebook.

РОЗДІЛ 1

ДОСЛІДЖЕННЯ ПОТОЧНОГО СТАНУ ПИТАННЯ АНАЛІЗУ ТА ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ МАРКЕТИНГОВИХ КАМПАНІЙ БАНКІВСЬКИХ УСТАНОВ

1.1 Банківська система: поняття, види та функції. Банківський депозит

Термін "банк" походить від французького слова "Banco", що означає "лава" або "стіл для обміну грошей". Згідно з тлумаченням Великої української енциклопедії: «Банківська система — сукупність різних за організаційно-правовою формою та спеціалізацією банків, інститутів, кредитних установ, а також структурних зв'язків і організаційно-правових відносин, що діють у межах єдиної фінансової системи та загальнодержавного грошово-кредитного механізму». Банківська система є складною та самостійною господарською структурою, яка регулюється законодавством і має внутрішню організацію [6, 7].

Дана система має загальні та специфічні характеристики, які відображають загальну природу системи та відмінності банківської діяльності. Загальні характеристики включають внутрішню складність та структурованість банківської системи, її постійний розвиток і адаптацію до змін у економічній ситуації та вимог ринкової економіки, зміну методів та інструментів банківської діяльності, розширення кола банківських операцій тощо. Банківська система також характеризується закритістю та саморегуляцією [7].

Банк - це установа, яка пропонує клієнтам широкий спектр фінансових послуг. Він виступає фінансовим посередником, з'єднуючи вкладників, які надають капітал, і позичальників, яким потрібен капітал. Приймаючи депозити, банк збільшує свої зобов'язання, що робить його боржником. І навпаки, коли він надає кредити, його активи збільшуються, що робить його кредитором.

Банки працюють у сфері фінансових послуг і відіграють центральну роль у фінансовій системі в кожній країні. Вони зберігають гроші та цінності, надають позики, кредити та платіжні послуги, такі як грошові перекази та чеки. Крім того, банки пропонують інвестиційні та страхові продукти. З інтеграцією фінансових індустрій традиційні межі між банками, страховими компаніями та компаніями, що займаються цінними паперами, стають менш помітними, що призводить до появи універсальних банків, які надають різноманітні послуги під одним дахом. Незважаючи на ці зміни, банки все ще зберігають свою основну функцію - приймати депозити та надавати кредити за рахунок цих депозитів [6].

Існують такі види банківських організацій:

- Кредитні спілки (Credit Unions): кредитна спілка - це неприбуткова установа, яка надає однакові базові послуги своїм членам.
- Інвестиційні банки (Investment banks): дані банки обслуговують багато великих корпорацій, фірм і купують акції для перепродажу інвесторам і навіть урядам.
- Комерційні банки (Commercial banks): комерційні банки - це фінансові організації, які приймають депозити, забезпечують безпеку рахунків і надають кредити.
- Роздрібні банки (Retail banks): роздрібний банк - це банк, який надає кредити лише малим підприємствам, компаніям та споживачам.
- Ощадно-позичкові асоціації (Savings and loan associations): ощадно-позичкові асоціації - це установи, які в основному допомагають фізичним особам з іпотечними кредитами на житло або нерухомість.
- Банки розвитку громад (Community development banks): метою банків розвитку громад є надання допомоги фізичним особам, які живуть у несприятливих соціально-економічних умовах.
- Інтернет- та необанки (Online and neobanks): інтернет- та необанки - це онлайн-банки без фізичних відділень. Вони популярні тому, що все, що завгодно, можна зробити онлайн, і це зручно і без зайвих клопотів.

Розглянемо інформацію про кожний вид банківської установи детальніше:

- Кредитні установи

Кредитні спілки - це неприбуткові установи, що належать та управляються своїми членами, які об'єднують свої фінансові ресурси. Вони надають послуги, подібні до банківських, але зазвичай пропонують їх за більш низькими відсотковими ставками. Щоб скористатися перевагами кредитної спілки, необхідно стати її членом. Члени вносять свої кошти для підтримки кредитування та інших послуг, які надає кредитна спілка. Надлишок коштів інвестується з метою отримання прибутку [8].

Кредитні спілки пропонують такі переваги, як:

- високі ставки збережень;
- низькі процентні ставки для позик;
- мінімальні комісійні збори.

Ці переваги пояснюються тим, що кредитні спілки не прагнуть отримувати прибуток від заощаджень своїх членів, але пропонують їм вигідні відсоткові ставки для позик. Однак кредитні спілки мають певні ризики, такі як:

- кредитний ризик;
- процентний ризик;
- ризик ліквідності;
- стратегічний ризик.

Наприклад, якщо один з членів не виконує свої зобов'язання щодо погашення боргу, це може суттєво вплинути на активи кредитної спілки.

Незважаючи на ці ризики, багато людей готові їх прийняти, оскільки отримують користь від членства в спілці, якщо вони дотримуються правил та умов, встановлених установою [8].

- Інвестиційні банки

Наступний вид банківської установи – це інвестиційні банки. Вони спеціалізуються на наданні комплексної фінансової допомоги великим корпораціям, урядам і фірмам з метою сприяння зростанню їх бізнесу. Вони також виступають посередниками між акціонерами та компаніями. У відмінну від

звичайних банків, інвестиційні банки не приймають депозити, оскільки їх головна мета полягає в торгівлі та продажу корпоративних акцій інвесторам. Вони виявляють потенційних інвесторів, які можуть бути зацікавлені у придбанні акцій компаній. Послуги інвестиційних банків включають управління фінансами та активами, проведення досліджень компаній, які можуть купити акції, а також допомогу у купівлі-продажу акцій від імені клієнтів [8, 10].

Незважаючи на різні переваги, пов'язані з інвестиційними банками, важливо враховувати ймовірні ризики. Один зі значних ризиків у галузі інвестиційного банкінгу - це ринковий ризик та ризик ліквідності. Ринковий ризик означає можливість збитків для клієнтів через коливання ринкових цін. З іншого боку, ризик ліквідності пов'язаний з проблемами продажу акцій за розумну ціну або навіть неможливістю їх продажу. Незважаючи на ці ризики, інвестиційні банки продовжують надавати цінні послуги своїм клієнтам, які часто переважають потенційні негаразди [8, 9].

- Комерційні банки

Комерційні банки - це фінансові установи, які забезпечують банківські послуги фізичним особам та підприємствам. Вони надають можливість вкласти гроші, відкривати та управляти рахунками, а також отримувати кредити для фінансової підтримки. Комерційні банки отримують прибуток шляхом оплати клієнтських комісій, щомісячної плати за обслуговування, відсотків з кредитів та оплати за користування сейфовими ячейками.

Функції комерційних банків:

- надають можливість приватним особам безпечно зберігати та відстежувати свої кошти;
- надають важливі кредити підприємствам;
- забезпечують онлайн-банкінг та зручність операцій з дебетовими картками через банкомати.

Комерційні банки мають багато філій, тому доступні для всіх. Однак користування послугами комерційного банку пов'язане з ризиками. Ці ризики включають кредитний ризик, операційний ризик та ризик ліквідності. Незважаючи

на це, банки застосовують методи управління ризиками для зниження ймовірності виникнення цих ризиків [8, 11].

- Роздрібні банки

Роздрібні банки функціонують подібно до традиційних банків, але обслуговують виключно індивідуальних клієнтів. Дані банки можуть пропонувати такі види послуг, як:

- різні види депозитів (строкові або фіксовані);
- рахунки в іноземній валюті;
- видача дебетових карток;
- видача кредитних карток;
- та інші.

Запровадження клієнтам кредитними та дебетовими картками допомагає підвищити кредитний рейтинг користувачів, що дозволяє мати змогу користуватись більш вигідними умовами запозичень.

Основним ризиком, пов'язаним з використанням роздрібного банку, є кредитний ризик, оскільки банк пропонує численні кредити, які потенційно можуть призвести до надмірної заборгованості. Крім того, операційні ризики можуть виникати як для працівників банку, так і для клієнтів. Основним стимулом для фізичних осіб користуватися послугами роздрібного банку є широкий спектр пропозицій, які він пропонує [8].

- Ощадно-позичкова асоціація

Це є фінансовою установою, що надає необхідну фінансову допомогу фізичним особам. Головна її мета полягає у допомозі клієнтам отримати іпотечні кредити на придбання нерухомості. Крім цього, асоціація приймає депозити на ощадні рахунки, які використовуються як інвестиції для відповідального кредитування позичальників [8, 12].

- Банки розвитку громад

Банки розвитку громад є приватними установами, які відрізняються від роботи інших банківських установ своєю функціональністю, оскільки вони не приймають грошові вклади від клієнтів. Замість цього, їх основний фокус полягає

на наданні допомоги, як особистої, так і для бізнесу, з метою надання низькопроцентних кредитів економічно вразливим районам у певному регіоні [8, 13].

- Інтернет-банкінг та необанки

Інтернет-банкінг та необанки - це два типи банків, які працюють в онлайн форматі. Однак, вони відрізняються у тому, що при онлайн-банкінгу клієнти можуть управляти своїми рахунками онлайн, не відвідуючи фізичні відділення банку, тоді як необанки - це банки, які повністю працюють онлайн та не мають фізичних відділень. Необанки можуть значно зменшити витрати та надати більший спектр продуктів, завдяки роботі онлайн. Відкриття рахунку в них також є простим і зазвичай не вимагає паперової документації.

Обидва типи банків пропонують різні послуги, такі як:

- рахунки збережень;
- поточні рахунки;
- збереження коштів;
- інструменти фінансової освіти;
- допомога в складанні бюджету.

Існує ризик, якщо системи онлайн-банкінгу та необанку не захищені Федеральною корпорацією страхування депозитів (FDIC). Тому важливо, щоб користувачі уважно читали умови та угоди перед відкриттям рахунку. Онлайн-банкінг є безпечним, якщо він працює під захистом FDIC [8, 14].

Існують багато різних функцій банку, знання яких допомагає краще розуміти такий вид системи та бути досвідченим користувачем.

Депозит - це укладення заощаджень клієнта у банку під відсоток, згідно з визначенням Національного банку України [15]. За цією угодою банк, який є однією стороною, приймає від вкладника, що є другою стороною, певну суму коштів (вклад), зобов'язуючись повністю виплатити вкладникові цю суму разом з відсотками або доходом в іншій формі, згідно з умовами та порядком, встановленими угодою. Термін "депозит" походить від латинського слова "depositum", що означає переданий на зберігання.

Виділимо найголовніші функції банківської системи:

- Страхові депозити: банки запроваджують можливість безпечно зберігати кошти клієнтів та заробляти певні відсотки на цих активах.
- Відсотки за депозитами: комерційні банки виплачують відсотки за депозитами. Розмір відсотка залежить від типу рахунку. Поточні рахунки зазвичай пропонують нижчі відсоткові ставки, у порівнянні з ощадними рахунками.
- Позики: одним із значних джерел фінансів для банку – є надання грошей у позику.
- Створення кредитів: банки мають змогу регулювати грошову масу, а також створювати та керувати кредитами, використовуючи клієнтські депозити для надання позик із дотриманням регуляторних вимог.
- Інші послуги: банкомати, міжнародні грошові перекази, страхування та сейфи для зберігання матеріальних цінностей, фінансові консультації та інші [16].

1.2 Специфіка маркетингових комунікацій в банківських установах

1.2.1 Маркетингові комунікації

Із розвитком нових технологій ведення бізнесу змінюється для кожної компанії, організації чи установи. Дуже важливо адаптуватися до нового бізнес-середовища, залишатися в курсі подій і навіть випереджати зміни, щоб якість продукту та сервісу не знижувала своїх стандартів. Кожен аспект діяльності компанії зазнає трансформацій, які необхідно подолати та використати на власну користь. Маркетингові стратегії допомагають не тільки підтримувати привабливість вже виготовлених продуктів, а й шляхом аналізу та дослідження,

виявити цільову аудиторію та підлаштувати створення нового продукту під дану аудиторію [17].

Важливо розуміти цінність та вплив якісної маркетингової кампанії та маркетингу в цілому. Можемо виділити чотири способи розуміння важливості маркетингових комунікацій, такі як:

- Маркетингові комунікації слугують зв'язком між різними учасниками ринку, тому обов'язковою мірою є встановлення міцного контакту між цільовими аудиторіями за допомогою різних каналів, такі як реклама (банери, контекстна, пряма, нативна (природна), емейл-розсилки і тд) стимулювання збуту, зв'язки із громадськістю тощо.
- Також маркетингові комунікації можуть виступати процесом обміну інформацією між суб'єктами ринку. Дана точка зору на маркетинг пов'язана із попередньою, проте його процесно-орієнтований підхід в першу чергу фокусується на передачі та отриманні інформації, нехтуючи при цьому процесом встановлення зв'язків.
- Зі структурної точки зору, маркетинг – це сукупність елементів, що передаються між суб'єктами. З даного боку маркетингові комунікації виступають сигналами, що спрямовані на споживача, з метою привернути його увагу та створити обізнаність про продукт або об'єкт управління в цілому.
- З іншого боку, інтегрований підхід розглядає маркетингові комунікації як комплексний набір заходів, що застосовуються компанією, організацією чи установою для впливу на споживачів. Ця перспектива фокусується на визначенні основних стратегій і тактик, що використовуються установою для просування товарів або послуг і максимізації прибутку [18].

Банківські маркетингові комунікації можна описати як систему взаємовідносин між банком та його клієнтами з урахуванням дій конкурентів, основною метою яких є сприяння співпраці та досягнення економічних і соціальних вигод для всіх залучених сторін. Банківські маркетингові кампанії

підтримують стабільну лояльність клієнтів до установи, а також укріплюють довіру та свої показники у конкурентному середовищі [19].

1.2.2 Аналіз маркетингових витрат українських банків

Банківським установам, так само як й іншим компаніям та організаціям, слід приділяти чималу частку свого бюджету на рекламу та просування продуктів, адже дані дії можуть значно підвищити пізнаваність банку та зосередити увагу потенційних клієнтів. Аналізуючи статтю Богдана Вальда [20] у 2021 році українські банки збільшили бюджет, що виділений на маркетингові кампанії, на 44%. У 2021 році банківські установи в Україні витратили на рекламу понад 1,36 млрд грн, а у 2020 році - 947 млн грн. На рис. 1.1 наведені витрати українських банків та рекламу та маркетинг і частка цих витрат у загальних видатках у період з 2016 по 2021 роки.

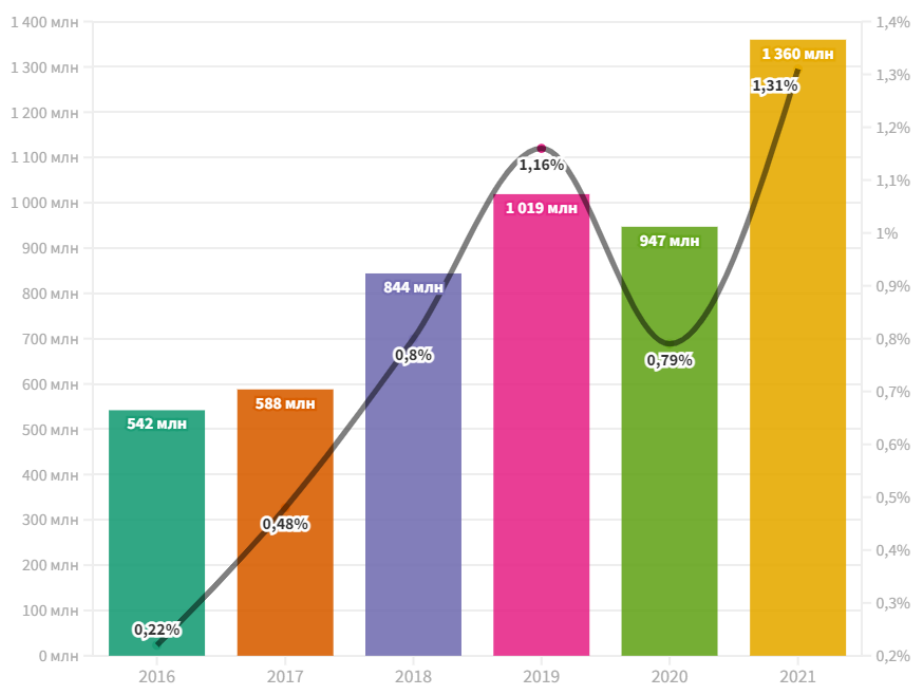


Рисунок 1.1 - Витрати банків на рекламу і маркетинг та частка цих витрат у загальних видатках, грн та %

Джерело: розроблено автором за даними [20]

Починаючи з 2016 року витрати банків зросли з 542 мільйонів гривень (2016 рік) до одного мільярда та 360 мільйонів гривень (2021 рік), тобто у два з половиною рази. У останній допандемічний рік (2019) витрати досягли одного мільярда, а після початку коронавірусної пандемії знизились на сімдесят мільйонів гривень впродовж 2020 року.

Частка витрат у загальних видатків під час цього періоду знизилась на 0,37%. Також на рис. 1.1 ми можемо спостерігати різке зростання маркетингових витрат українських банків на 413 мільйонів гривень (2021 рік). Частка цих витрат досягла піку та зросла на 0,52 відсотки.

Наприклад, банківська установа «ПУМБ» у 2021 році витратила найбільшу кількість грошей із свого бюджету на рекламу і досягла рекордні 252 млн гривень. Друге місце посідає «Альфа-Банк», що витратив 191 мільйон гривень на своє просування на конкурентному ринку. На третьому місці - «Райффайзен Банк», що витратив 118 мільйонів гривень. Детальніше про маркетингові витрати різних банків у 2021 році можемо подивитись на рис. 1.2.

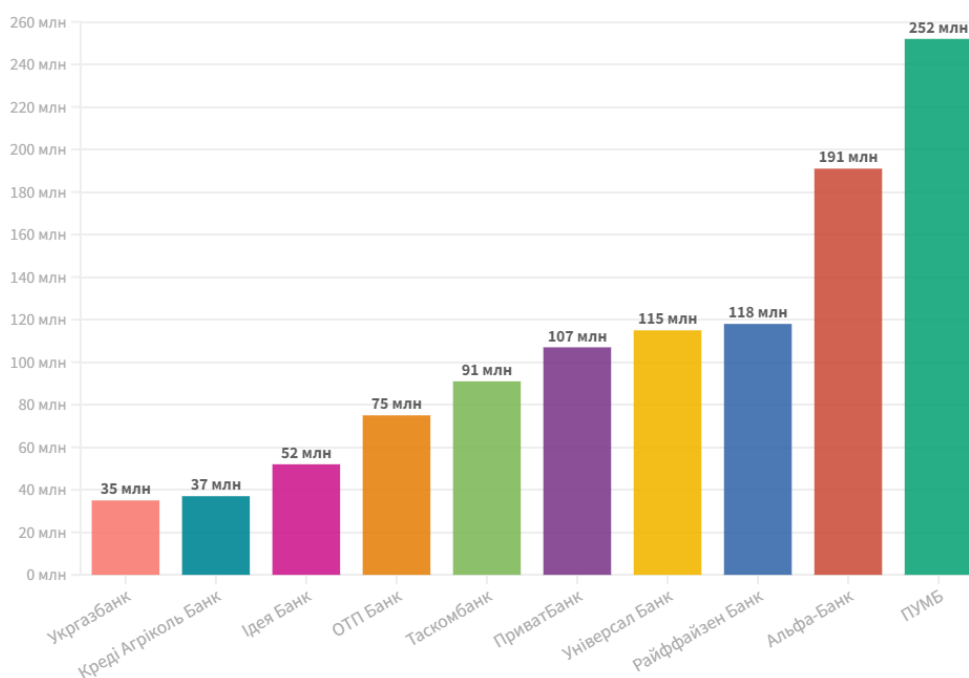


Рисунок 1.2 - Банки з найбільшими витратами на маркетинг і рекламу за 2021 рік

Джерело: розроблено автором за даними [20]

Серед усіх розглянутих випадків банківська установа «ПУМБ» відрізняється найбільше. Різниця між витратами із «Альфа-Банком» сягнула 61 млн грн.

Подивимось, яка частка від загальних витрат банківської установи йде саме на маркетингові витрати (див. рис. 1.3).

Очолює рейтинг банк «Таскомбанк», що витрачає майже 5% зі своїх витрат на рекламу. На другому місці – «УкрБудІнвестБанк» з частотою витрат 4,2%.

У банківських установах «ПУМБ» та «Ідея Банк» 3,7% та 3,4% відповідно йдуть на погашення рекламних рахунків та спроби залучити якомога більше нових клієнтів з метою підвищення прибутковості (рис. 1.3).

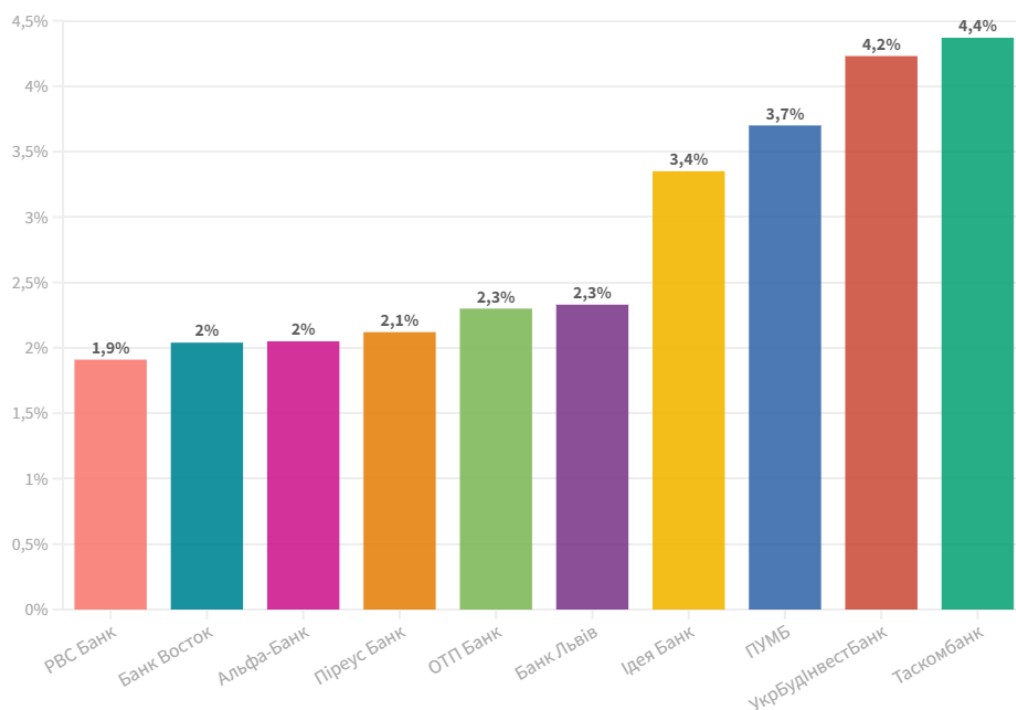


Рисунок 1.3 - Банки з найбільшою часткою витрат на маркетинг і рекламу в структурі загальних витрат за 2021 рік

Джерело: розроблено автором за даними [20]

Отже, аналізуючи витрати найвідоміших українських банківських установ, можна зробити висновок, що дана тема дипломної роботи є актуальною, адже суми витрат є істотною складовою бюджетів банків та дозволяють суттєво збільшити прибутковість та загальну ефективність банківських установ. Оптимізація

маркетингової діяльності допоможе знизити марнотратство ресурсів банків, а також зрозуміти, на який сегмент клієнтів слід зосередити більшу частину свого бюджету та як правильно влаштувати роботу маркетингових кампаній. Згідно зі статистикою по українським банкам, наше дослідження та метод покращення ситуації найбільше знадобиться «Таскомбанку» та «УкрБудІнвестБанку», адже саме в цих банках майже 5% витрат йдуть на маркетингові кампанії.

1.3 Методологічний підхід до моделювання ефективності контакту з клієнтом банку

У нашій бакалаврській роботі результуюча змінна «у», що вказує на подію покупки запропонованого депозитного продукту, набуває бінарного вигляду («0», коли клієнт відмовився від пропозиції, «1» - клієнт купив депозитний продукт). Для вирішення поставленої в роботі задачі пропонується методологічний підхід до моделювання ефективності контакту з клієнтом банку у продажі депозитних продуктів, основу якого складають такі складові етапи:

- Детальний аналіз даних;
- Категоризація даних (біннінг);
- Балансування вибірки;
- Логістична регресія;
- Stepwise Feature Selection для логістичної регресії.

Кожний етап реалізується на основі відповідного методу, сутність яких розглянемо детальніше.

○ Детальний дослідницький аналіз даних (EDA)

Розвідковий або детальний дослідницький аналіз даних (EDA) використовується аналітиками для дослідження наборів даних і узагальнення їхніх основних характеристик, часто із застосуванням методів візуалізації (графіків, діаграм, схем і т.д.). Даний аналіз допомагає краще зрозуміти структуру даних,

особливості, закономірності, виявити аномалії та зробити перевірку гіпотез та припущень.

EDA також може допомогти визначити, чи підходять статистичні методи, які розглядаються для аналізу даних. Поняття введено математиком Джоном Тьюкі, який сформулював мету даного аналізу: максимальне «проникнення» в дані, виявлення основних структур, вибір найбільш важливих змінних, виявлення відхилень і аномалій, перевірка основних гіпотез, розробка початкових моделей [20].

У науковій статті «Дослідницький підхід до аналізу даних фінансового обліку з використанням машинного навчання» автори використовували дослідницький аналіз даних для перевірки запропонованих гіпотез, а також пояснення закономірностей, що виникали у наборі даних [22].

Автор наукової роботи «Дослідження факторів, що впливають на впровадження мобільного банкінгу в Бангладеш» використовував детальний дослідницький аналіз даних (EDA) для з'ясування чинників, що впливають на впровадження телефонного банкінгу в Бангладеш. Для цього автор зібрав інформацію 630 респондентів методом опитування. На основі зібраних та посортованих даних, вияснилось, що ризик, зручність використання, простота доступу, вартість та порівняльні переваги є важливими факторами, що впливають на використання мобільного банкінгу в Бангладеш. Ризик став найвпливовішим фактором із всіх розглянутих, що впливає на розвиток телефонного банкінгу у цій країні [23].

Зважаючи на розмір та тип бази даних, ми можемо виділити такі види дослідницького аналізу даних [24]:

1. Одновимірний аналіз – це найпростіша форма дослідницького аналізу даних, оскільки цей аналіз фокусується на дослідженні або аналізі тільки однієї змінної величини. Одновимірний аналіз не розглядає причини або взаємозв'язки, і його основна мета - описати одну змінну та виявити закономірності в значеннях, яких вона набуває.

2. Двовимірний аналіз включає дві різні змінні. Цей тип аналізу досліджує причини та взаємозв'язки між двома факторами та його метою є відстеження та розуміння цих зв'язків.
3. Багатовимірний аналіз застосовується, коли існує три та більше змінних в наборі даних, що дозволяє прослідкувати складні закономірності та зв'язки між цими факторами.

- Категоризація даних (біннінг)

При категоризації даних (біннінгу, англ. binning) кількісні змінні (числові) розбиваються на інтервали, а якісні (текстові) переводяться у числову форму. У даній науковій роботі проводиться біннінг даних відповідно до бінарної (значення якої набувають тільки 0 або 1) таргетованої залежної змінної. Ця процедура проводиться з метою підвищення стійкості моделі до випадкових збурень та аномальних випадків в базі даних. Одночасно категоризація змінних дозволяє підвищити ефективність та якість класифікації, адже розбиття даних на категорії відносно бінарної результативної змінної зменшує ймовірність негативного впливу екстремальних викидів у кожній змінній, перетворюючи їх в оцінку систематичного впливу категорії на бінарну таргетовану змінну [25, 26].

Створення якісного алгоритму біннінгу полягає у вирішенні питання визначення оптимальної кількості категорій та їх діапазонів для кожної з числових пояснюючих змінних. Для ефективної категоризації існує правило, яке полягає у тому, що кожна категорія певної змінної повинна об'єднати групу сусідніх значень змінної, яким властиві подібні характеристики щодо впливу на результуючу змінну (у нашому випадку «у»). Цій темі було присвячено декілька вітчизняних і закордонних публікацій, про які надається короткий аналіз нижче.

У літературному джерелі [27] автори детально описують процес біннінгу при побудові моделі. Аналіз якості категорій відбувається за допомогою показників вагомості ознаки WOE та інформаційної значимості IV. Під час категоризації автори вирішують:

- Максимізувати показники IV (інформаційної значимості) як критерію оптимальності біннінгу;
- Розбити вибірку змінних так, щоб показник WOE (вагомість змінної) мав зростаючий або спадаючий тренд при зміні категорії;
- Об'єднати категорії з майже однаковими значеннями показника вагомості змінної WOE в одну категорію, з метою посилення зростаючого або спадного тренду;
- Прослідкувати, щоб у різних категоріях показник вагомості змінної різко різнився;
- Встановити ліміт по кількості категорій до 50.

Також про важливість значної різниці між показниками вагомості змінної WOE в категоріях було описано в джерелі [26], де автор виділив три способи біннінгу:

1. Встановити однакові довжини в інтервалах категорій на множині можливих значень певної змінної;
2. Виділити в категорії однакові кількості спостережень;
3. Посилити різницю в значеннях величини вагомості змінної WOE між сусідніми категоріями.

Автор книги Н. Сіддікі [28] наголошує на такі рекомендації щодо біннінгу даних:

1. Усі пропущенні значення певного показника записувати як окрему категорію;
2. Кожна категорія змінної має містити не менше, ніж 5% вибірки даних;

Найпоширенішими показниками, за якими прийнято оцінювати категоризацію даних – є вагомість змінної WOE (Weight of Evidence) та інформаційної значимості IV (Information Value).

Показник WOE розраховує для кожної підгрупи (категорії) узагальнену кількісну оцінку поведінки того чи іншого спостереження. Вагомість змінної вимагає обчислення часток негативних подій (при яких таргетована бінарна змінна дорівнюватиме нулю) та позитивних подій (при яких таргетована бінарна змінна

дорівнюватиме одиниці) за кожною категорією показника відносно загальної кількості негативних і позитивних випадків. Формула WOE виглядає таким чином:

$$WOE_i = \ln \left(\frac{B_i}{G_i} \right), i = \overline{1, k}, \quad (1.1)$$

де B_i — відношення кількості негативних подій у i -й категорії до загального числа негативних подій у вибірці;

G_i — частка позитивних подій за i -ю категорією відносно їх загальної кількості;

k — кількість категорій певної змінної.

Автор публікації Дж. Германа [26] приводить такі аргументи щодо використання показника вагомості змінної WOE (Weight of Evidence):

- Даний показник допомагає якісно обробити пропущенні значення у певній змінній;
- Ефективне усунення проблем із екстремальними випадками у змінній, що в подальшому впливало б на якість класифікації;
- Створення усіх змінних в одній розмірності.

Також для оцінювання ефективності біннінгу бази даних та якості прогностичної сили використовують інформаційної значимості IV (Information Value) [27]:

$$IV = \sum_{i=1}^k (B_i - G_i) \cdot WOE_i. \quad (1.2)$$

Чим більше інформаційна значимість змінної приближається до одиниці, тим сильнішим є вплив даної змінної на таргетовану результуючу змінну «у». Інтерпретувати результати ми можемо таким чином:

- Якщо IV менше, ніж 0,02 – змінна не має прогностичної сили;
- Якщо IV більше або дорівнює 0,02, але менше, ніж 0,1 – змінна має слабку прогностичну силу;
- Якщо IV більше або дорівнює 0,1, але менше, ніж 0,3 – змінна має середню прогностичну силу;

- Якщо IV більше або дорівнює 0,3, але менше, ніж 0,5 – змінна має високу прогностичну силу;
- Якщо IV більше або дорівнює 0,5 – змінна має відмінно прогностичну силу.

Проте у даній дипломній роботі оцінювалась результативність кожної категорії (біну) за допомогою показника оцінки події Event Rate. Даний показник відповідає за відсоток подій для кожного біну (категорії), тобто це відношення позитивної події, при якій бінарна результуюча змінна $y=1$, до загальної кількості подій у цій змінній (негативні + позитивні). Завдяки цьому показнику ми можемо подивитись, наскільки позитивно та чи інша категорія впливає на результуючу змінну. Чим вища оцінка події Event Rate, тим певна категорія (бін) має більше подій із позитивним результатом ($y=1$).

Отже, біннінг даних (категоризація) допомагає значно покращити якість класифікації та моделі в цілому, позбавляючись екстремальних значень у кожній змінній та приводячи якісні та кількісні дані до числового формату у стандартизованому вигляді.

○ Балансування вибірки

Балансування вибірки використовуються, коли присутній сильний дисбаланс у позитивних випадках (коли $y = 1$) та негативних (коли $y = 0$) у наборі даних. Доречність та важливість балансування вибірки описані в науковому джерелі [29]. Дисбаланс у подіях сильно впливатиме на якість класифікації, так як побудувати ефективну модель на тренінгових даних та перевірити якість на тестових даних стає набагато складніше, коли вибірка має велику різницю нулів та одиниць [30].

Балансування вибірки можна виконати таким чином:

- Проаналізувати вибірку даних та зрозуміти які події (негативні або позитивні) переважають;
- Частіше переважають негативні події (тобто бінарна результуюча змінна дорівнює 0, $y = 0$). Слід подивитись, яка кількість позитивних результатів ($y = 1$) присутня у вибірці;

- Якщо балансування вибірки робити 50% на 50%, то слід випадковим чином обрати кількість негативних результатів ($y = 0$), що дорівнюватиме точній кількості позитивних результатів ($y = 1$).
- У залежності від вимог, можна корегувати випадковий вибір негативних або позитивних подій.
- Логістична регресія (Logistic Regression)

На сьогоднішній день існує багато методів оцінки бінарної класифікації. Незважаючи на це, використання логістичної регресії при класифікації та дослідженні факторів, що відображаються як результат бінарної таргетованої змінної, залишається одним з найоптимальніших методів [31].

Логістична регресія - це метод статистичного аналізу, який дозволяє передбачити бінарний результат, наприклад, "так" чи "ні", на основі попередніх спостережень за набором даних. Модель логістичної регресії прогнозує залежну змінну (y), аналізуючи залежність від однієї або кількох існуючими незалежними змінними.

Існують такі види логістичної регресії:

- Бінарна (двофакторна) логістична регресія: вихідна результуюча змінна набуває тільки двох значень – 0 або 1;
- Багатофакторна логістична регресія: вихідна результуюча змінна набуватиме трьох або більше значень без упорядкування;
- Порядкова логістична регресія: вихідна результуюча змінна набуватиме трьох або більше значень (категорій) із певним упорядкуванням (рейтингом).

Основні припущення бінарної логістичної регресії [32]:

- Вихідна результуюча змінна повинна бути дихотомічного характеру (наприклад, позитивна або негативна подія);
- У наборі даних не повинно бути екстремальних викидів, які б сильно впливали на роботу моделі класифікації;

- Кореляція (взаємозв'язок) між предикторами повинен бути відсутній. Цю характеристику можна перевірити за допомогою кореляційної матриці між незалежними змінними.

Задачею логістичної регресії є оцінка логарифмічних шансів події (позитивної або негативної). Із сторони математики логістична регресія оцінює функцію множинної лінійної регресії, яка розраховується за формулою [28]:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \dots + \epsilon, \quad (1.3)$$

де y_i – залежна результуюча змінна;

x_i – незалежні змінні набору даних;

β_0 – стала;

β_p – коефіцієнти нахилу для кожної незалежної змінної;

ϵ – вільний член (похибка моделі).

Виділяють такі показники ефективності моделі (у тому числі логістичної регресії, призначеної для вирішення задачі бінарної класифікації), як [28]:

- Ассурасу – це загальна точність моделі за певного рівня розмежовування (за допомогою параметру cut-off); Чим більший даний показник, тим якість моделі є вищою;
- Коефіцієнт Gini (Джіні) – інтегральний показник якості класифікації моделі, набуває значення від 0 до 1. Чим вищий даний показник, тим якість класифікації є кращою.
- AUROC – площа під ROC-кривою (Receiver Operating Characteristic). Даний показник також є інтегральною характеристикою якості роботи моделі класифікації. Крива ROC дозволяє візуально оцінити точність бінарної класифікації. По осі ординат відображається частка правильно визначених позитивних результатів TPR (True Positive Rate або чутливість — Sensitivity), а по осі абсцис — частка помилково діагностованих позитивних результатів FPR (False Positive Rate) при варіюванні порога відсікання.

У даній дипломній роботі якість роботи моделі класифікації аналізується за показником AUROC, адже він дозволяє оцінити ефективність моделі в цілому, а не тільки частину, яка залишилась до відсікання. Чим вищий даний показник, тим якість класифікації є кращою.

○ Stepwise Feature Selection для логістичної регресії

Іноді велика кількість змінних при побудові моделі може призвести до вагомого погіршення якості регресії. Деякі змінні можуть не нести в собі цінної інформації і ніяк не впливати на результуючу залежну змінну і навпаки призводити до перенавчання моделі. Задля уникнення проблем із ефективністю моделі часто використовують метод Stepwise Feature Selection. Нижче будуть описані два види: Forward Feature Selection та Backward Feature Selection.

○ Forward Feature Selection

Прямий покроковий вибір змінних – це метод відбору змінних, який:

1. Починається з моделі, яка не містить змінних (називається «нульовою моделлю» (Null Model));
2. Потім починає по одній додавати значущі змінні до моделі;
3. Зупиняється, коли буде виконане заздалегідь задане правило зупинки, або коли всі змінні, що розглядаються, будуть включені до моделі.

Forward Feature Selection обирає значущу змінну, яку додавати до моделі, такими методами:

1. Змінна має найменше значення p-value;
2. Змінна має найвищий приріст R^2 ;
3. Змінна має найбільшу суму залишків квадратів, порівняно з іншими предикторами, що розглядаються.

Правило зупинки для Forward Feature Selection виконується, коли усі розглянуті змінні матимуть значення p-value вище за порогове, тобто $p\text{-value} > 0,05$. У такому випадку Forward Feature Selection зупиниться і отримає результат зі

змінними, p -value яких буде меншим за порогове значення, тобто $p\text{-value} < 0,05$. У результаті ми отримаємо модель, що містить тільки значущі змінні.

- Backward Feature Selection

Зворотній покроковий відбір змінних – це метод вибору змінних, який:

1. Починається з моделі, що містить усі запропоновані змінні (називається «повною моделлю» (Full Model));
2. Видаляє найменш значущі змінні одна за одною;
3. Закінчує свою роботу, коли буде досягнуто заздалегідь встановлене правило зупинки, або коли модель не матиме жодної змінної.

Backward Feature Selection обирає найменш значущу змінну, яку виключати з моделі, такими методами:

1. Змінна має найбільше значення p -value;
2. Змінна має найменший приріст R^2 ;
3. Змінна має найменшу суму залишків квадратів, порівняно з іншими предикторами, що розглядаються.

Правило зупинки Backward Feature Selection виконується, коли всі змінні, що залишилися в моделі, матимуть p -value менші за певне попередньо визначене порогове значення, тобто $p\text{-value} < 0,05$. У такому разі зворотній покроковий відбір змінних припинить роботу та поверне модель поточного кроку [33].

РОЗДІЛ 2

ФОРМУВАННЯ ІНФОРМАЦІЙНОЇ БАЗИ ТА ПОПЕРЕДНЯ ОБРОБКА ДАНИХ

2.1 Підготовка та опис даних

У рамках виконання даного дослідження, побудови математичних моделей та проведення експериментів було застосовано дані прямих маркетингових кампаній португальської банківської установи. Ці маркетингові кампанії базувалися на телефонних дзвінках. У деяких випадках вимагалось робити більше одного контакту з тим самим клієнтом задля отримання відповіді, чи підпишеться клієнт на продукт (банківський строковий депозит).

Датасет складається із 16 змінних та 4119 спостережень. Джерело даних – це репозитарій машинного навчання, де зберігається велика кількість баз даних, спрямованих на дослідження різних явищ [34]. Переглянути частину датасету з цього серверу можна на рис. А.1 Додатку А.

Перед початком аналізу бази даних португальської банківської установи, ми провели роботу над даними, а саме:

- У змінній «prdays»: змінили значення 999 на «по», щоб не допустити трактування відсутності попередніх контактів як час у днях, що минули з моменту останнього контакту;
- У результуючій залежній змінній «у»: змінили значення «по» на «0», а «yes» на «1». Тепер дана змінна є бінарною, що набуває значення нуля або одиниці.

Змінений вигляд датасету можна переглянути на рис. А.2 Додатку А. Проаналізуємо змінні отриманої бази даних.

Дані клієнтів банку:

- Вік (age): вік певного клієнта банку – числова змінна;

- Професія (job): тип роботи (адміністратор, робітник, підприємець, домогосподарка, менеджмент, пенсіонер, фрілансер, сфера послуг, студент, технік, безробітний, невідомо) – категоріальна змінна;
- Сімейний стан (marital): розлучений, одружений, неодружений, невідомо – категоріальна змінна;
- Освіта (education): базова 4 роки, базова 6 років, базова 9 років, середня школа, неписьменний, професійні курси, університетський ступінь, невідомо – категоріальна змінна;
- Дефолт (default): чи має кредитний дефолт (ні, так, невідомо) – категоріальна змінна;
- Житло (housing): чи має кредит на житло? (ні, так, невідомо) – категоріальна змінна;
- Кредит (loan): чи має особистий кредит? (ні, так, невідомо) – категоріальна змінна.

Дані, пов'язані з останнім контактом поточної маркетингової кампанії:

- Контакт (contact): тип зв'язку з контактом (мобільний телефон (telephone), стільниковий (cellular) – категоріальна змінна;
- Місяць (month): останній контактний місяць року (січень, лютий, березень, ..., листопад, грудень) – категоріальна змінна;
- День тижня (day_of_week): останній контактний день тижня ('mon', 'tue', 'wed', 'thu', 'fri') – категоріальна змінна;
- Тривалість контакту (duration): тривалість останнього контакту в секундах; числова змінна.

Інші атрибути:

- Кампанія (campaign): кількість контактів, здійснених під час цієї кампанії та для цього клієнта;
- Попередні дні (pdays): кількість днів, що минули після останнього контакту з клієнтом з попередньої кампанії; числова змінна;

- Попередні контакти (previous): кількість контактів, здійснених до цієї кампанії та для цього клієнта; числова змінна;
- Результат попередньої кампанії (poutcome): результат попередньої маркетингової кампанії (провал, відсутній, успіх) – категоріальна змінна.

Вихідна (результуюча) змінна:

- у: факт оформлення клієнтом строкового депозиту (так, ні).

2.2 Детальний дослідницький аналіз даних

Завантажуємо датасет у простір Jupyter Notebook, де будемо виконувати дослідження даних (рис. 1.2).

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome
0	30	blue-collar	married	basic.9y	no	yes	no	cellular	may	fri	487	2	no	0	nonexistent
1	39	services	single	high.school	no	no	no	telephone	may	fri	346	4	no	0	nonexistent
2	25	services	married	high.school	no	yes	no	telephone	jun	wed	227	1	no	0	nonexistent
3	38	services	married	basic.9y	no	unknown	unknown	telephone	jun	fri	17	3	no	0	nonexistent
4	47	admin.	married	university.degree	no	yes	no	cellular	nov	mon	58	1	no	0	nonexistent
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4114	30	admin.	married	basic.6y	no	yes	yes	cellular	jul	thu	53	1	no	0	nonexistent
4115	39	admin.	married	high.school	no	yes	no	telephone	jul	fri	219	1	no	0	nonexistent
4116	27	student	single	high.school	no	no	no	cellular	may	mon	64	2	no	1	failure
4117	58	admin.	married	high.school	no	no	no	cellular	aug	fri	528	1	no	0	nonexistent
4118	34	management	single	high.school	no	yes	no	cellular	nov	wed	175	1	no	0	nonexistent

4119 rows × 16 columns

Рисунок 2.1 – Завантажена база даних

Джерело: розроблено автором

Подивимось на типи даних та на наявність пропущених значень (рис. 2.2).

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4119 entries, 0 to 4118
Data columns (total 16 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   age             4119 non-null   int64
1   job             4119 non-null   object
2   marital         4119 non-null   object
3   education       4119 non-null   object
4   default         4119 non-null   object
5   housing         4119 non-null   object
6   loan            4119 non-null   object
7   contact         4119 non-null   object
8   month           4119 non-null   object
9   day_of_week     4119 non-null   object
10  duration        4119 non-null   object
11  campaign        4119 non-null   int64
12  pdays          4119 non-null   object
13  previous        4119 non-null   int64
14  poutcome        4119 non-null   object
15  y               4119 non-null   int64
dtypes: int64(4), object(12)
memory usage: 515.0+ KB

```

Рисунок 2.2 – Типи даних

Джерело: розроблено автором

Одинадцять змінних мають тип даних `object`, адже вони є категоріальними. Пропущені значення відсутні. Змінна «*duration*», що відповідає за тривалість останнього контакту з клієнтом також визначилась, як змінна типу `object`, що не є коректним. Слід змінити на тип даних `int64`, що відповідає за числові змінні (рис. 2.3).

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4119 entries, 0 to 4118
Data columns (total 16 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   age             4119 non-null   int64
1   job             4119 non-null   object
2   marital         4119 non-null   object
3   education       4119 non-null   object
4   default         4119 non-null   object
5   housing         4119 non-null   object
6   loan            4119 non-null   object
7   contact         4119 non-null   object
8   month           4119 non-null   object
9   day_of_week     4119 non-null   object
10  duration        4118 non-null   float64
11  campaign        4119 non-null   int64
12  pdays          4119 non-null   object
13  previous        4119 non-null   int64
14  poutcome        4119 non-null   object
15  y               4119 non-null   int64
dtypes: float64(1), int64(4), object(11)
memory usage: 515.0+ KB

```

Рисунок 2.3 – Зміна типу даної «*duration*»*Джерело: розроблено автором*

Тепер змінна «*duration*», що відповідає за тривалість останнього контакту з клієнтом, набула числового характеру.

Проаналізуємо категоріальні змінні та подивимось, скільки факторів (категорій) міститься у кожній (рис. 2.4):

```

job          12
marital      4
education    8
default      3
housing      3
loan         3
contact      2
month       10
day_of_week  5
pdays      21
previous     7
poutcome     3
dtype: int64

```

Рисунок 2.4 – Кількість категорій у змінних

Джерело: розроблено автором

- Усього у датасеті по змінній *job* (професії клієнтів) налічується дванадцять різних категорій: адміністратор, робітник, підприємець, домогосподарка, менеджмент, пенсіонер, фрілансер, сфера послуг, студент, технік, безробітний, невідомо;
- Зміна «*marital*» (сімейний стан клієнта) містить 4 категорії: розлучений, одружений, неодружений або невідомо;
- Змінна «*education*» (освіта клієнта) має 8 категорій: базова освіта 4 роки (ступінь бакалавра), базова 6 років (магістр), базова 9 років, середня школа, неписьменний, професійні курси, університетський ступінь або немає інформації про освіту (тобто невідомо).
- Найменше категорій у змінній «*contact*», що відповідає за тип зв'язку (за допомогою мобільного телефону чи стільникового);
- І так далі.

Дана функція допомагає нам детальніше познайомитись з базою даних, а також зрозуміти, наскільки об'ємним є наш датасет.

Подивимось на загальну статистику числових даних (рис. 2.5).

	age	duration	campaign	previous	y
count	4119.000000	4118.000000	4119.000000	4119.000000	4119.000000
mean	40.113620	256.607819	2.537266	0.190337	0.108521
std	10.313362	254.471830	2.568159	0.541788	0.311076
min	18.000000	0.000000	1.000000	0.000000	0.000000
25%	32.000000	103.000000	1.000000	0.000000	0.000000
50%	38.000000	181.000000	2.000000	0.000000	0.000000
75%	47.000000	317.000000	3.000000	0.000000	0.000000
max	88.000000	3643.000000	35.000000	6.000000	1.000000

Рисунок 2.5 – Загальна статистика числових змінних датасету

Джерело: розроблено автором

Як видно з рис. 2.5:

- Середній вік клієнтів португальського банку – 40 років. Мінімальний – 18 років, а найстаршим клієнтам – 88 років. У даному банку набагато більше клієнтів, вік яких сягає 47 років і більше. Дані є цілком реальні, адже за законодавством Португалії людина, якій виповнилось вісімнадцять років, має право на відкриття банківського рахунку.
- Щодо поточної кампанії: середня тривалість контакту з клієнтом сягнула 257 секунд, тобто 4 хвилини та 17 секунд. Максимальна тривалість – година і 43 секунди.
- Розподіл змінної `campaign` каже про те, що мінімальна кількість контактів із клієнтом – це 1, а максимальна – 35. Можливо, таке відбувалось, коли клієнт не одразу виходив на зв'язок.

Розглянемо детальніше зв'язок між тривалістю контакту та результатом: чи клієнт придбав депозитний продукт (див. рис. 2.6-2.7). Подивимось на найдовші та найкоротші дзвінки із клієнтами, які в результаті погодились на пропозицію (рис. 2.6).

	duration	y
2455	63.0	1
143	63.0	1
947	64.0	1
3556	78.0	1
2501	85.0	1
...	...	...
262	1868.0	1
3266	1980.0	1
1685	2301.0	1
1392	2653.0	1
2231	3643.0	1

447 rows × 2 columns

Рисунок 2.6 – Розподіл тривалості контакту з успішним результатом ($y = 1$)*Джерело: розроблено автором*

Із досліджуваної вибірки (усього 4119 спостережень) 447 клієнтів погодились на придбання депозитного продукту. Найменша тривалість успішного контакту (тобто $y=1$) – 63 секунди (тобто 1 хвилина та 3 секунди). Найбільша – 3643 секунди (1 година та 43 секунди). Можемо припустити, що для успішного результату розмова повинна тривати не менше однієї хвилини. Чим більше триває зв'язок, тим вірогідніше, що клієнт придбає депозит.

Незважаючи на дане припущення, дослідження показало, що бувають і винятки. Розглянемо тривалості розмов та ситуацію, коли клієнт відмовився у пропозиції (рис. 2.7).

	duration	y
1546	3253.0	0
2530	1855.0	0
3256	1820.0	0
1651	1720.0	0
4059	1590.0	0
...	...	...
934	5.0	0
1896	5.0	0
1600	4.0	0
1950	0.0	0
979	NaN	0

3672 rows × 2 columns

Рисунок 2.7 – Розподіл тривалості контакту з негативним результатом ($y = 0$)*Джерело: розроблено автором*

Усього 3672 випадки, коли людина відмовилася. Із дослідженого виходить, що контакт може тривати 54 хвилини та 13 секунд і все одно залишитись безуспішним. Дане спостереження можемо назвати викидом, адже подібних йому більше немає. Даний викид відрізняється від інших спостережень на 23 хвилини.

Дослідимо залежності між сімейним статусом клієнтів та питання, чи клієнти придбали депозитний продукт. Спершу дізнаємось, якої категорії сімейного статусу найбільше у датасеті (рис. 2.8).

```
marital
divorced    446
married     2509
single      1153
unknown      11
dtype: int64
```

Рисунок 2.8 – Кількість спостережень у кожній категорії змінної «marital» (сімейний статус клієнта)

Джерело: розроблено автором

У базі даних найбільше спостережень з одруженими клієнтами (married), адже їх кількість сягає 2509. Також досить багато неодружених клієнтів (single) – 1153 спостереження. Всього 11 клієнтів, сімейний стан яких співробітникам банківської установи невідомий (unknown – 11). Візуалізуємо даний розподіл за допомогою гістограми (рис. 2.9).

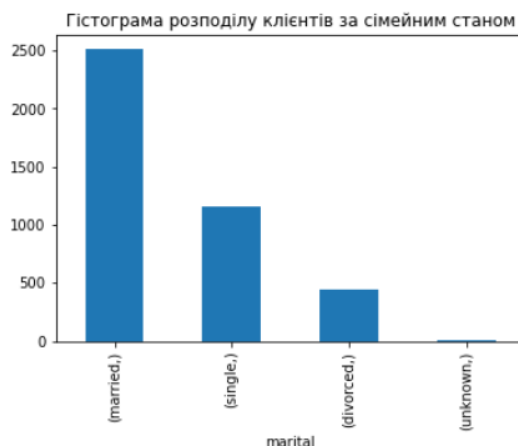


Рисунок 2.9 – Гістограма розподілу клієнтів за сімейний станом («marital»)

Джерело: розроблено автором

Як видно з рис. 2.9, 60% з усіх категорій змінної про сімейний стан займають одружені клієнти, 27,9% - це неодружені клієнти, 10,8% – розлучені. Продовжимо аналізувати детальніше саме одружених клієнтів, адже їх частка у датасеті є найбільшою.

Проведемо аналіз того, яке співвідношення між одруженими клієнтами, що мають особистий кредит та не мають кредиту на житло. Дослідимо можливість того, що вони придбають у такому разі депозитний продукт (рис. 2.10-2.11).

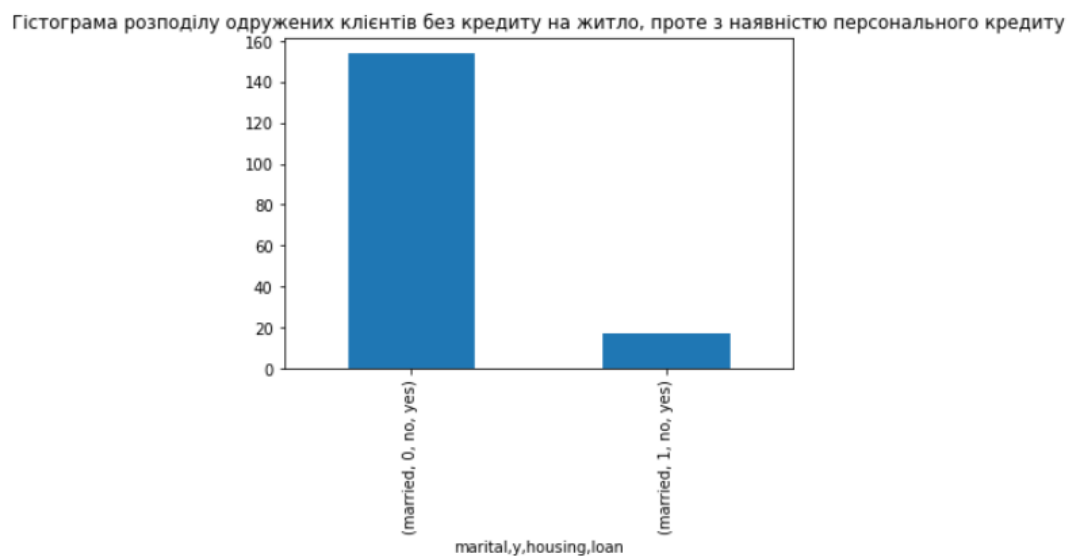


Рисунок 2.10 – Гістограма розподілу одружених клієнтів із персональним кредитом, проте без кредиту на житло

Джерело: розроблено автором

```
marital  y  housing  loan
married  0  no      yes    154
         1  no      yes    17
dtype: int64
```

Рисунок 2.11 – Гістограма розподілу одружених клієнтів із персональним кредитом, проте без кредиту на житло у цифрах

Джерело: розроблено автором

Даний аналіз показав, що існує набагато більше клієнтів з персональним кредитом (154), які в результаті не придбали депозитний продукт ($y = 0$), при цьому не маючи кредиту на житло. Усього 17 клієнтів при таких умовах погодились на умови даної маркетингової кампанії (рис. 2.11).

Перевіримо ситуацію, коли одружені клієнти не матимуть особистих та житлових кредитів. Можливо у такому випадку вони будуть більш схильними до придбання депозитного продукту (рис. 2.12).

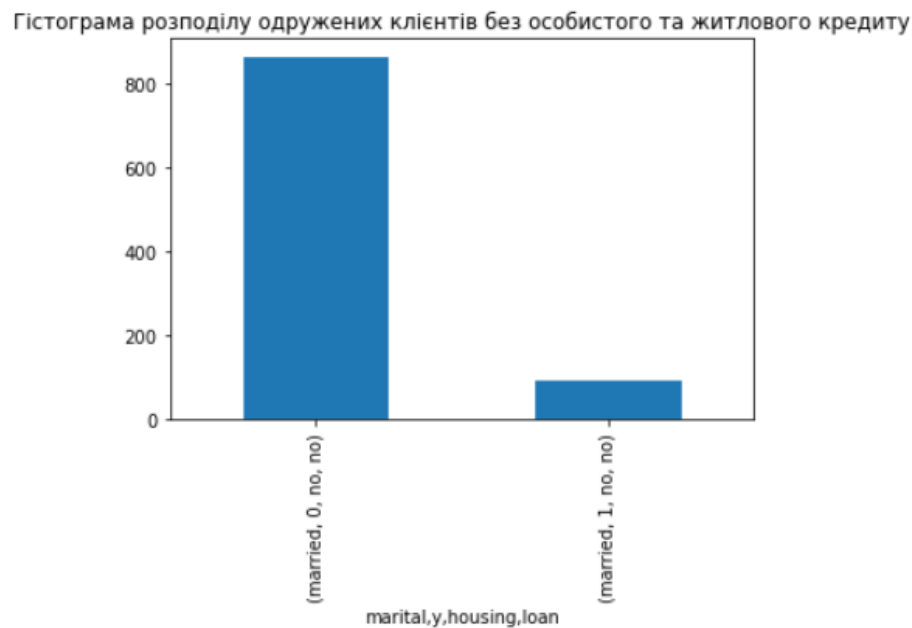


Рисунок 2.12 – Одружені клієнти без особистого та житлового кредиту

Джерело: розроблено автором

```
marital  y  housing  loan
married  0  no      no    865
         1  no      no    93
dtype: int64
```

Рисунок 2.13 – Одружені клієнти без особистого та житлового кредиту у цифрах

Джерело: розроблено автором

Отже, 865 одружених клієнтів без кредитів не погоджуються на депозитний продукт, що є вагомою різницею із тими, що погоджуються (93 спостереження) (рис. 2.13).

Перевіримо ситуацію з наявністю кредитного дефолту (рис. 2.14). Дослідимо питання, чи схильні клієнти з невикладеним кредитом до набуття ще однієї відповідальності?

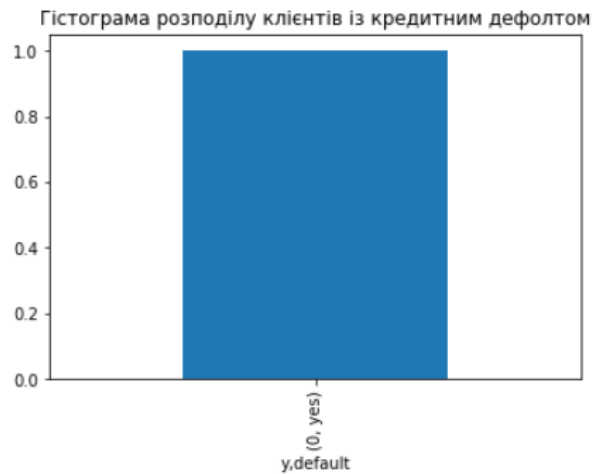


Рисунок 2.14 – Клієнти з наявністю кредитного дефолту

Джерело: розроблено автором

Жоден клієнт із невиконаними кредитними зобов'язаннями не придбає запропонований депозитний продукт. У висновку можемо сказати, що не слід витрачати ресурси компанії на клієнтів із кредитним дефолтом, адже спроба буде безуспішною.

Проаналізуємо, у які дні тижня краще всього зв'язуватись із клієнтами банку задля успішного результату (див. рис. 2.15).



Рисунок 2.15 – Гістограма розподілу купівлі депозиту за днями тижня

Джерело: розроблено автором

У цілому, різниця не є суттєвою, проте дана гістограма сигналізує, що краще зосередити ресурси банківської установи на понеділок (mon) та четвер (thu). Із досліджуваної вибірки 97 клієнтів придбали депозитний продукт у понеділок, а

96 – у четвер. Найменш продуктивним днем є середа (wed), адже тільки 82 клієнти погодились на пропозицію у цей день (рис. 2.15).

Перевіримо, наскільки кількість контактів, здійснених під час даної кампанії, взаємозалежить із таргетованою змінною «у» (чи придбав клієнт депозит). Для цього побудуємо теплову карту для змінних *campaign* (кількість контактів поточної кампанії) та *y* (чи купив клієнт строковий депозитний продукт) (див. рис. 2.16):



Рисунок 2.16 – Теплова карта кореляції між змінними «у» та «*campaign*»

Джерело: розроблено автором

Коефіцієнт автокореляції є від’ємним та незначним. Можемо вважати, що кількість контактів із клієнтом під час досліджуваної кампанії не впливають на бажаний результат.

Проаналізуємо теплову карту кореляцій між такими змінними: «*previous*» (загальна кількість контактів, здійснених до даної кампанії), «*campaign*» (кількість контактів, здійснених під час даної кампанії), «*duration*» (тривалість останнього контакту з клієнтом), «у» (результуюча змінна, що відповідає за продаж строкового депозиту) (рис. 2.17).



Рисунок 2.17 – Теплова карта кореляції між змінними «у», «campaign», «duration» та «previous»

Джерело: розроблено автором

Найбільший взаємозв'язок спостерігаємо між результуючою змінною «у» та «duration» - 0.41, що означає залежність між тривалістю контакту з клієнтом та вихідним результатом. Також присутня незначна кореляція між «у» та «previous» 0,26. Можна зробити припущення у взаємозалежності кількості попередніх контактів та результатом маркетингової кампанії. Якщо з клієнтом протягом різних кампаній відбувалась певна кількість контактів (чим більше, тим краще), то з великою вірогідністю у даного клієнта вибудовувалась довіра та лояльність до цієї установи, а, отже, можливість продати новий банківський продукт стане реальнішою.

Перевіримо змінну «campaign», що відповідає за кількість контактів з клієнтом, здійснених під час спостережуваної кампанії, на наявність викидів та нетипових значень за допомогою коробчастої діаграми (рис. 2.18).

Дана діаграма пропонує розглядати кількість контактів, що перевищує число 6, як викиди. Медіана – це значення, яке ділить набір даних навпіл, тобто половина значень знаходиться вище, а половина - нижче цієї лінії. Медіана змінної «campaign» знаходиться на числі 2.5 (див. рис. 2.5). Прямокутник бокс-плоту вказує на степінь варіації розподілу у даній змінній. Він зсунутий до нижнього «вуса» (межі прямокутника), що вказує не то, що розподіл значень має схил вниз. Загалом, більшість спостережень набувають значень приблизно від 1 до 3.

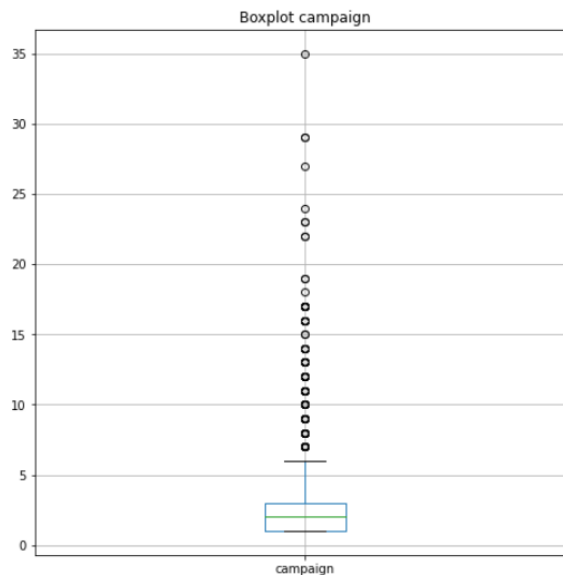


Рисунок 2.18 – Коробчаста діаграма змінної «campaign»

Джерело: розроблено автором

Перевіримо, починаючи з якої нетипової кількості контактів, здійснених під час даної кампанії, результуюча змінна «у» матиме негативний характер ($y = 0$). Для цього посортуємо дані за змінною «у» у низхідному порядку та методом підбору кількості контактів з'ясуємо дане питання (рис. 2.19).

	campaign	y
484	16	0
605	14	0
672	18	0
688	17	0
713	22	0
735	13	0
766	13	0
971	13	0
1488	17	0
1033	17	0
1053	17	0
1219	17	0
1265	14	0
1308	17	0
1328	13	0
1381	15	0
1386	13	0
1445	13	0
1457	17	0
3931	13	0

Рисунок 2.19 – Мінімальна кількість безуспішних контактів із клієнтом

Джерело: розроблено автором

Після аналізування вибірки даних було зрозуміло, що починаючи з 12 контактів, жоден клієнт не купив депозитний продукт. Робимо висновок, що

кількість контактів, яка перевищує 12, - не є раціональною для ефективного результату.

Після детального дослідницького аналізу даних, ми можемо висвітлити такі поради португальській банківській установі:

- *Щодо цільової аудиторії банку:* 35,5% усіх спостережень – це клієнти віком від 35 до 47 років, проте вони не є нашою цільовою аудиторією. За показником Event Rate (частота подій) було досліджено, що люди, віком від 57 років і вище (тобто пенсійного віку), більш схильні до придбання запропонованого продукту. Такі клієнти складають лише 5,4% вибірки, тому для максимізації цього числа слід розробляти маркетингові кампанії, які будуть розраховані на залучення клієнтів саме пенсійного віку.
- *Щодо сімейного стану та кредитів клієнтів:* 60% спостережень складають одружені люди. Аналіз показав, що одружені люди, які мають персональний кредит, скоріше за все, не погодяться на пропозицію. Вони будуть більш ретельно підходити до питання, адже, можливо, мають відповідальність у вигляді родини. На відміну від цього, неодружені клієнти, що займають 28% усіх спостережень, більш схильні до купівлі продукту. Також володіння житловим кредитом притаманне 52% вибірки клієнтів, на відміну від особистого кредиту – тільки 19% мають такий вид кредиту.
- *Щодо клієнтів із кредитним дефолтом:* абсолютно усі клієнти, що не виплатили кредитні зобов'язання, не погодяться на поточну кампанію. У результаті, слід попередньо закупити кредитну історію клієнтів, щоб мати уявлення про потенційні можливості продажів.
- *Щодо днів тижня:* за гістограмою розподілу успішних подій за днями тижня, було проаналізовано, що суттєва різниця відсутня, проте все одно краще зосереджувати основну кількість контактів на понеділок та четвер.
- *Щодо кількості контактів, які необхідно зробити із клієнтом:* дослідження вказує, що більше 12 контактів із клієнтом – це марнуванням ресурсів банку. У середньому відбувається 2,5 контактів, тому вважатимемо, що для

успішного результату знадобиться не більше, ніж 3 контакти із клієнтом банку.

- *Щодо тривалості контакту з клієнтами:* за аналізом зрозуміло, що розмова, яка триватиме менше однієї хвилини, не буде успішною для банку. У середньому телефонна розмова триває 4 хвилини, яких вистачає, щоб прослідкувати настрій та готовність клієнта до придбання депозиту. Довгий контакт, наприклад 54 хвилин, може не принести прибутку банку, а тільки змарнувати ресурси, тому слід раціонально розподіляти час.

2.3 Категоризація (біннінг) даних

Для подальшої роботи з даними та побудови якісних моделей нам потрібно зробити категоризацію (біннінг) даних. Доречність та ефективність біннінгу описано в першому розділі (див. п. 1.3 «Категоризація даних»).

Ця процедура проводиться з метою підвищення стійкості моделі до випадкових збурень та аномальних випадків в базі даних. Процес розбиття даних на категорії відносно бінарної таргетованої змінної зменшує ймовірність негативного впливу екстремальних викидів у кожній змінній на модель класифікації, перетворюючи їх в оцінку систематичного впливу категорії на залежну змінну.

Для проведення оптимального біннінгу кількісних та категоріальних змінних встановлюємо пакет клас «optbinning» мови Python. Продемонструємо біннінг числової змінної на прикладі «age» (вік клієнта) (рис. 2.20-2.24).

Записуємо її у змінну «x» та обираємо бінарну таргетовану змінну «y» з датасету (рис 2.20):

```
x = df['age'].values  
y = df.y
```

Рисунок 2.20 – Запис змінних «x» та «y»

Джерело: розроблено автором

Дискретизуємо у відповідності до «у»:

```
from optbinning import OptimalBinning
variable = 'age'
optb = OptimalBinning(name=variable, dtype="numerical", solver="cp")

optb.fit(x1, y)

OptimalBinning(name='age')
```

Рисунок 2.21 – Дискретизація змінної «age» відносно «у»

Джерело: розроблено автором

За допомогою функції «optb.status» у Python, перевіримо, чи знайдене оптимальне рішення дискретизації. Дана функція показала результат «OPTIMAL», що свідчить про її ефективність.

За допомогою функції «optb.binning_table» у Python можна переглянути таблицю метрик по змінній віку клієнта (рис. 2.22)

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 27.50)	304	0.073804	264	40	0.131579	-0.218864	0.003850	0.000480
1	[27.50, 33.50)	1001	0.243020	879	122	0.121878	-0.131169	0.004401	0.000550
2	[33.50, 35.50)	346	0.084001	328	18	0.052023	0.796709	0.039083	0.004760
3	[35.50, 47.50)	1465	0.355669	1342	123	0.083959	0.283799	0.025627	0.003193
4	[47.50, 57.50)	780	0.189366	694	86	0.110256	-0.017808	0.000060	0.000008
5	[57.50, inf)	223	0.054139	165	58	0.260090	-1.060431	0.089945	0.010744
6	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
7	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		4119	1.000000	3672	447	0.108521		0.162967	0.019735

Рисунок 2.22 – Метрики по змінній «age»

Джерело: розроблено автором

Утворилось 6 інтервалів, обмежених оптимальними точками розбиття (від 18 до 27; від 27 до 33; 33-35; 35-47; 47-57; 57-...). Детальніше про кожну метрику:

- Count: кількість записів для кожного біну;
- Count (%): відсоток записів для кожного біну;
- Non-event: кількість записів, при яких $y=0$ (негативна подія);
- Event: кількість записів, при яких $y=1$ (позитивна подія);
- Event rate: відсоток записів, при яких $y=1$ (позитивна подія);

- WoE: вагомість змінної (Weight of Evidence) показує узагальнену кількісну оцінку поведінки того чи іншого спостереження;
- IV: інформаційна значимість (Information value) характеризує, наскільки добре бінеризація розділяє класи змінної;
- JS: розбіжність Дженсена-Шеннона - метрика розбіжності між оригінальним розподілом змінної та розподілами, утвореними після бінеризації.

Аналізуючи показник Event Rate, можемо зазначити, що більшість клієнтів, які купили депозитний продукт, віком від 57 років, адже мають максимальний Event Rate – 0,26 (див. рис. 2.22). Зробимо візуалізацію показника Event Rate, із значення якого в подальшому будемо створювати датасет для роботи з даними (рис. 2.23).

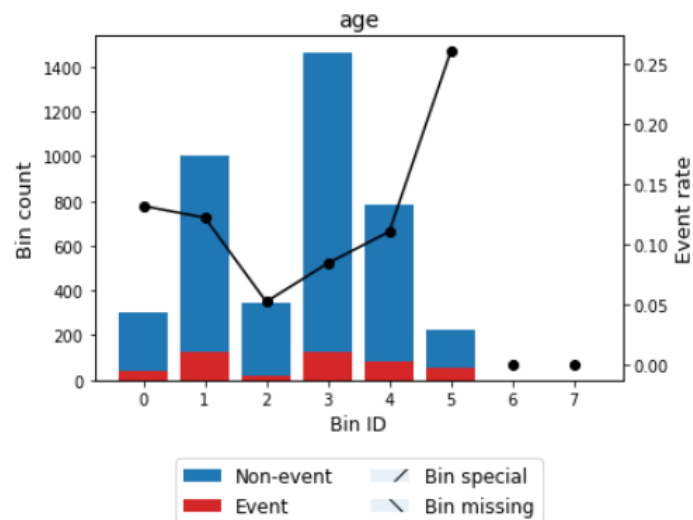


Рисунок 2.23 – Event Rate для категоризації змінної «age»

Джерело: розроблено автором

Із гістограми на рис. 2.23 бачимо, що значення частоти подій (Event Rate) є максимальним для 6 категорії (клієнти віком від 57 років і вище). Найменше значення – у другій категорії (27-33 роки). Дивлячись на різницю між синім кольором (без подій) та червоним (подія), розуміємо, що кількість негативних результатів (коли $y = 0$) значно переважає кількість позитивних результатів (коли $y = 1$).

Запишемо стовпчик дискретизованих значень величини «age» за метрикою «event_rate» у нову змінну «age_bins» та додамо до датасету, попередньо видаливши змінну «age» (рис. 2.24).

```
age_bins = optb.transform(x1, metric="event_rate")
print(age_bins)
```

[0.12187812 0.08395904 0.13157895 ... 0.13157895 0.26008969 0.05202312]

Рисунок 2.24 – Запис дискретизованої величини у змінну «age_bins»

Джерело: розроблено автором

Таким чином ми категоризували числову змінну, що відповідає за вік клієнтів «age».

Розглянемо візуалізацію біннінгу категоріальної змінної сімейного стану клієнтів «marital» та подивимось, яка категорія більш схильна до купівлі депозитного продукту (рис. 2.25-2.26).

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	[unknown, divorced]	457	0.110949	415	42	0.091904	0.184676	0.003519	0.000439
1	[married]	2509	0.609128	2257	252	0.100438	0.08643	0.004399	0.000550
2	[single]	1153	0.279922	1000	153	0.132697	-0.228616	0.015992	0.001995
3	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
4	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		4119	1.000000	3672	447	0.108521		0.023910	0.002984

Рисунок 2.25 – Біннінг змінної «marital»

Джерело: розроблено автором

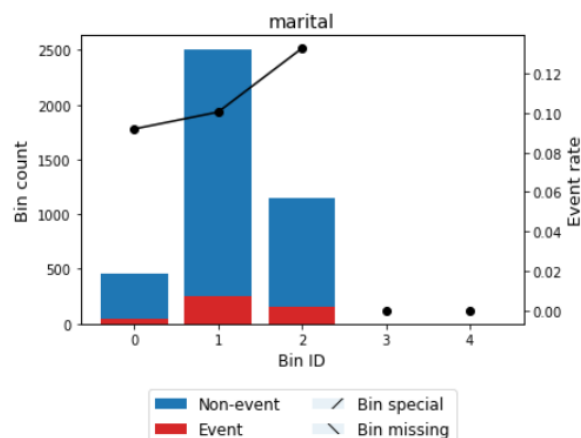


Рисунок 2.26 – Візуалізація біннінгу змінної «marital»

Джерело: розроблено автором

Отже, сімейний стан клієнтів розбився на три біни: невідомий сімейний стан та розлучені клієнти; одружені клієнти; неодружені. Найбільшу частоту подій (Event Rate) спостерігаємо у третьому біні (неодружені клієнти) – 0,132 (див. рис. 2.25). Найменша частота подій у нульовому біні – 0,091. Можемо зробити висновок, що хоча й одружені клієнти займають 60% з усієї вибірки, вони погоджуються на пропозицію від банку рідше (Event Rate – 0,1), ніж неодружені клієнти, що займають всього 28% з усіх спостережень (Event Rate – 0,132).

Таким чином дискретизуємо та записуємо до датасету кожні числові та категоріальні змінні. У результаті отримуємо датасет із дискретизованими величинами у відповідності до таргетованої бінарної змінної «у» (чи купив клієнт депозитний продукт: 0 – ні; 1 – так) (рис. 2.27). Більш детально утворений датасет можна передивитись на рис. Б.1 Додатку Б.

y	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome
0	0.121878	0.065891	0.100438	0.074913	0.119759	0.109425	0.110481	0.140271	0.064586	0.108073	0.240157	0.112608	0.108521	0.081749	0.081749
0	0.083959	0.086957	0.132697	0.104235	0.119759	0.107510	0.110481	0.051125	0.064586	0.108073	0.158392	0.109966	0.108521	0.081749	0.081749
0	0.131579	0.086957	0.100438	0.104235	0.119759	0.109425	0.110481	0.051125	0.126415	0.103145	0.080997	0.123583	0.108521	0.081749	0.081749
0	0.083959	0.086957	0.100438	0.074913	0.119759	0.107510	0.100000	0.051125	0.126415	0.108073	0.005333	0.111111	0.108521	0.081749	0.081749
0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.096413	0.113450	0.005333	0.123583	0.108521	0.081749	0.081749
0	0.121878	0.086957	0.132697	0.132075	0.119759	0.107510	0.110481	0.140271	0.301435	0.111628	0.035088	0.111111	0.108521	0.266779	0.266779
0	0.121878	0.130435	0.132697	0.132075	0.119759	0.109425	0.110481	0.140271	0.301435	0.113450	0.115607	0.109966	0.108521	0.081749	0.081749
0	0.083959	0.065891	0.100438	0.132075	0.062189	0.109425	0.110481	0.140271	0.096413	0.113450	0.005333	0.112608	0.108521	0.081749	0.081749
0	0.121878	0.086957	0.091904	0.119626	0.119759	0.107510	0.110481	0.140271	0.096413	0.105826	0.005333	0.123583	0.108521	0.266779	0.266779
0	0.052023	0.065891	0.100438	0.074913	0.062189	0.107510	0.110481	0.051125	0.064586	0.111628	0.066351	0.123583	0.108521	0.081749	0.081749
0	0.131579	0.086957	0.132697	0.074236	0.062189	0.109425	0.110481	0.140271	0.081575	0.111628	0.115607	0.123583	0.108521	0.081749	0.081749

Рисунок 2.27 –Датасет після біннінгу даних

Джерело: розроблено автором

РОЗДІЛ 3

КОМПЛЕКС МОДЕЛЕЙ ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ КОНТАКТУ З КЛІЄНТАМИ БАНКУ ПІД ЧАС МАРКЕТИНГОВОЇ КАМПАНІЇ З ПРОДАЖУ ДЕПОЗИТНОГО ПРОДУКТУ

3.1 Побудова логістичної регресії на «сирих» даних

Даний розділ починатимемо з побудови першої моделі логістичної регресії, побудованої на «сирих» даних. У цьому випадку дане визначення регресії означає модель, що буде будуватись на категоризованих даних без балансування вибірки та без відбору значущих факторів, тобто на «сирих» даних.

Визначаємо матрицю незалежних змінних: *age* (вік клієнта), *job* (професія клієнта), *marital* (сімейний стан), *education* (освіта), *default* (наявність кредитного дефолту), *housing* (наявність кредиту на житло), *loan* (наявність персонального кредиту), *contact* (спосіб контакту), *month* (останній контактний місяць року), *day_of_week* (останній контактний день року), *duration* (тривалість контакту), *campaign* (кількість контактів, здійснених під час даної кампанії), *pdays* (кількість днів, що минули з моменту останнього контакту), *previous* (загальна кількість контактів, до цієї кампанії), *poutcome* (результат попередньої маркетингової кампанії).

Та вектор цільової змінної: *y* (чи купив клієнт депозитний продукт).
Утворений результат можна переглянути на рис. 3.1.

```

      age      job  marital  education  default  housing      loan  \
0  0.121878  0.065891  0.100438  0.074913  0.119759  0.109425  0.110481
1  0.083959  0.086957  0.132697  0.104235  0.119759  0.107510  0.110481
2  0.131579  0.086957  0.100438  0.104235  0.119759  0.109425  0.110481
3  0.083959  0.086957  0.100438  0.074913  0.119759  0.107510  0.100000
4  0.083959  0.130435  0.100438  0.132075  0.119759  0.109425  0.110481

      contact      month  day_of_week  duration  campaign      pdays  previous  \
0  0.140271  0.064586      0.108073  0.240157  0.112608  0.108521  0.081749
1  0.051125  0.064586      0.108073  0.158392  0.109966  0.108521  0.081749
2  0.051125  0.126415      0.103145  0.080997  0.123583  0.108521  0.081749
3  0.051125  0.126415      0.108073  0.005333  0.111111  0.108521  0.081749
4  0.140271  0.096413      0.113450  0.005333  0.123583  0.108521  0.081749

      poutcome
0  0.081749
1  0.081749
2  0.081749
3  0.081749
4  0.081749
0      0
1      0
2      0
3      0
4      0
Name: y, dtype: int64

```

Рисунок 3.1 – Матриця незалежних змінних та вектор цільової змінної «у»

Джерело: розроблено автором

Розділимо дані на навчальну і тестову вибірки. Навчальну вибірку використовуємо для навчання моделі, а тестову – для перевірки якості побудованої моделі. Розбиття вибірки відбуватиметься автоматично, без балансування нулів та одиниць.

```

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    X, y, stratify = y, test_size=0.2, random_state=0)

```

Рисунок 3.2 – Розбиття вибірки на тестову та навчальну

Джерело: розроблено автором

- параметр «*stratify = y*» означає, що метод `train_test_split` повертає навчальні та тестові вибірки, які мають ті самі пропорції міток класів, що й набір вхідних даних. Наприклад, якщо цільова змінна «у» є бінарною змінною зі значеннями 0 або 1 і містить 89,15% нулів і 10,85% одиниць, то параметр «*stratify = y*» гарантуватиме, що при розбитті даних навчальна та тестова вибірка також міститимуть 89,15% нулів і 10,85 % одиниць.
- параметр «*test_size*» визначає розмір тестової вибірки. Як правило, *test_size* приймають рівним 0.2 або 0.3 (відповідно 20% та 30% від набору вхідних даних).

- параметр «*random_state*» гарантує, що створюватиметься те саме розбиття на навчальну та тестову вибірки кожного разу, коли запускатиметься код.

Створюємо і навчаємо класифікатор на навчальному наборі даних. Оцінимо якість класифікатора, порахувавши показник «*roc_auc_score*», саме за яким в подальшому будемо оцінювати якість моделі (рис. 3.3). Детальніше про цей показник можна почитати в першому розділі (див. п. 1.3 «Логістична регресія»).

Показник «*roc_auc_score*» - вказує на ймовірність того, що випадково вибраний позитивний випадок буде оцінений вище, ніж випадково вибраний негативний випадок.

	precision	recall	f1-score	support
0	0.90	0.99	0.94	735
1	0.60	0.13	0.22	89
accuracy			0.90	824
macro avg	0.75	0.56	0.58	824
weighted avg	0.87	0.90	0.87	824

AUC ROC: 0.5619735534663304

Рисунок 3.3 – Оцінка якості моделі на «сирих» даних після біннінгу

Джерело: розроблено автором

Відповідно всього 824 спостережень, з яких 735 належить до першого класу (ті клієнти, що не купили депозитний продукт), 89 - до другого (ті клієнти, що купили).

- *accuracy* відповідає за повну точність моделі. Повна точність являє собою кількість правильно класифікованих екземплярів даних над загальною кількістю екземплярів даних. Тобто точність даної моделі становить 90%, що дуже добре, адже тільки 10% були класифіковані неправильно. Але не завжди показник точності відповідає реальності, адже деколи дані можуть бути незбалансованими. Тому треба розглядати і інші показники.
- *precision* - це відношення правильно передбачених позитивних значень до загальних прогнозованих позитивних значень. В ідеалі *precision* повинен бути 1 (високим) для хорошого класифікатора. Високий *precision* свідчить про низький показник хибно позитивних результатів, на відміну від *accuracy*.

У даному випадку точність для класифікації стану кредиту становить 90% для тих, хто не купив депозит, та 60%, для тих, хто купив. Отже, результати точності для $y = 0$ є доволі непоганими. Для $y = 1$ класифікатор помилився на 40%, що є суттєвою помилкою.

- *recall* - це відношення правильно передбачених позитивних значень до фактичних позитивних значень. В ідеалі *recall* має бути 1 (високим) для хорошого класифікатора. Це також буде означати, що при класифікації немає фальшиво-негативних значень. У даному випадку *recall* становить 99% для тих, хто не придбав депозит, та 13%, для тих, хто придбав. Результати першої групи є дуже хорошими, адже вони близькі до 100%, а ось для другої результат є дуже низьким. Отже, дана модель погано класифікує клієнтів, які погодяться на пропозицію, адже кількість правильно передбачених позитивних значень становить 13% фактичних позитивних значень.
- *f1-score* - середньозважене значення *precision* та *recall*. Оцінка F1 стає 1 лише тоді, коли *precision* та *recall* обидва рівні 1. Дана оцінка враховує як помилкові спрацьовування, так і помилкові негативні. Отже, в даній класифікації оцінка F1 для стану продажі 0 ($y = 0$) становить 94%, а для стану продажі 1 ($y = 1$) - 22%.
- *Roc_auc_score* показує, що якість моделі дорівнює 56,19%, що є значно меншим за результат ассигасу. При виборі найкращої моделі ми будемо орієнтуватись саме на метрику *auc_roc*, адже вона оцінює якість моделі в цілому, не враховуючи рівень відсікання при тестуванні вибірки.

Отже, дана модель з високою точністю визначає клієнтів, які впродовж телефонної розмови не погодяться на купівлю депозитного продукту. Логістична регресія, що побудована на усіх змінних без балансування розбиття вибірки не зможе правильно класифікувати людей, що погодяться на купівлю.

Побудуємо та зробимо візуалізацію матриці помилок (рис. 3.4) та дізнаємось, як саме логістична регресія класифікувала спостереження. За допомогою даної візуалізації ми зможемо побачити, скільки і яких випадків було класифіковано невірно.

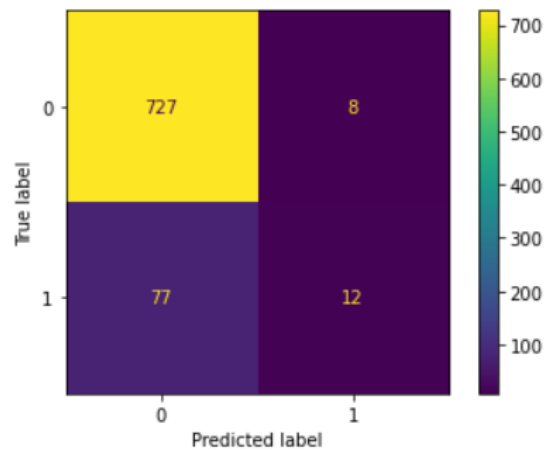


Рисунок 3.4 – Матриця помилок логістичної регресії на «сирих» даних

Джерело: розроблено автором

- *TP – істинно позитивний результат*, значення якої можна побачити в лівому верхньому куті. Це ті значення, що були класифіковані, як ті клієнти, що не купляють депозит і вони справді його не купляють. У даному випадку більшість значень (727) є істинно позитивними, що вказує на правильну класифікацію.
- *FP - помилково позитивний результат*, значення якої можна побачити в правому верхньому куті. Це ті значення, що були класифіковані як клієнти, що купили депозит, а насправді вони не купили. Значення цього показника є маленьким (8).
- *TN – істинно негативний результат*, значення якої можна побачити в правому нижньому куті. Це ті значення, що були класифіковані, як клієнти, що придбали депозитний продукт і вони справді його придбали. Їх кількість - 12, що є значно менше за кількість помилково негативного результату - 77. Підтверджує погано роботу класифікатора при $y = 1$.
- *FN - помилково негативний результат*, значення якої можна побачити в лівому нижньому куті. Це ті значення, що були класифіковані як клієнти, що не придбають депозит, а насправді вони його придбають. Для даного випадку кількість цих людей є доволі великою - 77.

Ми прагнемо до більшої кількості істинно позитивних та істинно негативних результатів, тому спробуємо покращити значення показника «*roc_auc_score*» та якість моделі класифікації.

3.2 Логістична регресія на збалансованій вибірці

Після проведення категоризації даних та побудови моделі регресії, здійснимо спробу покращення ефективності результатів моделювання за рахунок балансування вибірки. Доречність балансування доведено у статті [30] (див. п. 1.3 «Балансування вибірки»). Типовою проблемою при моделюванні складних процесів, зокрема продажі депозитних продуктів, є те, що подія, яка досліджується, є рідкісною. Тобто купівля відбувається набагато рідше, ніж відмова від пропозиції. Цей дисбаланс класів сильно впливає на продуктивність традиційних класифікаторів, тому вирішено збалансувати тренінгову вибірку таким чином, щоб кількість нулів та одиниць була однаковою.

Спершу запишемо усі спостереження, при яких $y = 0$ (тобто клієнт не купив депозитний продукт), у клас «*class_0*». Ситуації, при яких $y = 1$ (клієнт погодився придбати депозитний продукт), записуємо у клас «*class_1*».

Із класу «*class_0*» випадково вибираємо кількість спостережень, що дорівнюють кількості з класу «*class_1*». Відомо, що база даних португальської банківської установи складається з 447 спостережень, при яких клієнти погодилися на умови поточної кампанії, тобто придбали продукт. З цього виходить, що ми випадково обираємо 447 спостережень, при яких $y = 0$ та записуємо у новий датасет з назвою «*class_0_sample*» (рис. 3.5). Детальніше даний проміжний датасет можна передивитись на рис. В.1 Додатку В.

```
class_0_sample = class_0.sample(n=len(class_1), random_state=0)
class_0_sample
```

	y	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poi
1530	0	0.260090	0.206128	0.091904	0.074913	0.062189	0.109425	0.110481	0.051125	0.126415	0.108073	0.035088	0.109966	0.108521	0.081749	0.
131	0	0.083959	0.095137	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.081749	0.
4082	0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.051125	0.064586	0.111628	0.158392	0.112608	0.108521	0.081749	0.
3358	0	0.083959	0.095137	0.091904	0.132075	0.119759	0.107510	0.110481	0.051125	0.064586	0.105826	0.510903	0.053942	0.108521	0.081749	0.
3508	0	0.121878	0.065891	0.132697	0.104235	0.119759	0.109425	0.110481	0.051125	0.301435	0.111628	0.115607	0.123583	0.108521	0.081749	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1913	0	0.110256	0.086957	0.100438	0.132075	0.119759	0.109425	0.110481	0.051125	0.126415	0.105826	0.021277	0.109966	0.108521	0.081749	0.
2066	0	0.052023	0.115774	0.132697	0.119626	0.119759	0.109425	0.110481	0.140271	0.100629	0.105826	0.035088	0.123583	0.108521	0.081749	0.
283	0	0.083959	0.086957	0.100438	0.074913	0.119759	0.107510	0.110481	0.051125	0.064586	0.103145	0.066351	0.112608	0.108521	0.081749	0.
2361	0	0.052023	0.065891	0.100438	0.074913	0.119759	0.107510	0.110481	0.051125	0.064586	0.103145	0.240157	0.123583	0.108521	0.081749	0.
526	0	0.083959	0.065891	0.100438	0.132075	0.062189	0.107510	0.110481	0.051125	0.064586	0.105826	0.240157	0.025532	0.108521	0.081749	0.

447 rows × 16 columns

Рисунок 3.5 – Датасет «class_0_sample»

Джерело: розроблено автором

Третім етапом об'єднуємо клас «class_1», де зберігаються усі успішні спостереження, та датасет «class_0_sample» з однаковою кількістю спостережень. Отримуємо датасет «data1», що склеться з 894 спостережень (рис. 3.6).

Даний датасет будемо використовувати для побудови логістичної регресії. Детальніше його можна передивитись на рис. В.1 Додатку В.

```
data1 = pd.concat([class_0_sample, class_1], axis=0)
data1
```

	y	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poi
1530	0	0.260090	0.206128	0.091904	0.074913	0.062189	0.109425	0.110481	0.051125	0.126415	0.108073	0.035088	0.109966	0.108521	0.081749	0.
131	0	0.083959	0.095137	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.081749	0.
4082	0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.051125	0.064586	0.111628	0.158392	0.112608	0.108521	0.081749	0.
3358	0	0.083959	0.095137	0.091904	0.132075	0.119759	0.107510	0.110481	0.051125	0.064586	0.105826	0.510903	0.053942	0.108521	0.081749	0.
3508	0	0.121878	0.065891	0.132697	0.104235	0.119759	0.109425	0.110481	0.051125	0.301435	0.111628	0.115607	0.123583	0.108521	0.081749	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4034	1	0.260090	0.206128	0.100438	0.132075	0.119759	0.109425	0.100000	0.140271	0.100629	0.103145	0.080997	0.111111	0.108521	0.081749	0.
4066	1	0.260090	0.206128	0.100438	0.088578	0.119759	0.109425	0.110481	0.140271	0.301435	0.103145	0.510903	0.123583	0.108521	0.266779	0.
4069	1	0.260090	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.064586	0.113450	0.510903	0.112608	0.108521	0.081749	0.
4088	1	0.083959	0.095137	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.301435	0.113450	0.158392	0.123583	0.108521	0.266779	0.
4089	1	0.131579	0.130435	0.132697	0.132075	0.119759	0.109425	0.100000	0.140271	0.301435	0.108073	0.021277	0.123583	0.108521	0.266779	0.

894 rows × 16 columns

Рисунок 3.6 – Збалансована вибірка даних

Джерело: розроблено автором

Для якісної класифікації видаляємо з датасету дві змінні: «campaign» та «duration». Перша змінна відповідає за кількість контактів, здійснених під час даної кампанії та для даного клієнта. Друга змінна вказує на тривалість останнього контакту з клієнтом. Наявність цих змінних в датасеті сильно впливатиме на

таргетовану бінарну залежну змінну «у» (чи придбає клієнт продукт), адже ми завчасно не можемо знати, кількість та тривалість контактів з певним клієнтом. При прогнозуванні продажу депозитного продукту ці дві змінні треба виключити із аналізу.

Визначаємо матрицю залежних змінних, видаляючи змінні «*campaign*» та «*duration*». Визначаємо вектор залежної змінної «у» (рис. 3.7).

```

      age      job      marital      education      default      housing      loan \
1530  0.260090  0.206128  0.091904  0.074913  0.062189  0.109425  0.110481
131   0.083959  0.095137  0.100438  0.132075  0.119759  0.109425  0.110481
4082  0.083959  0.130435  0.100438  0.132075  0.119759  0.109425  0.110481
3358  0.083959  0.095137  0.091904  0.132075  0.119759  0.107510  0.110481
3508  0.121878  0.065891  0.132697  0.104235  0.119759  0.109425  0.110481

      contact      month      day_of_week      pdays      previous      poutcome
1530  0.051125  0.126415      0.108073  0.108521  0.081749  0.081749
131   0.140271  0.064586      0.111628  0.108521  0.081749  0.081749
4082  0.051125  0.064586      0.111628  0.108521  0.081749  0.081749
3358  0.051125  0.064586      0.105826  0.108521  0.081749  0.081749
3508  0.051125  0.301435      0.111628  0.108521  0.081749  0.081749
1530   0
131    0
4082   0
3358   0
3508   0
Name: y, dtype: int64

```

Рисунок 3.7 – Вибірка даних, не включаючи змінні «*campaign*» та «*duration*»

Джерело: розроблено автором

Розділяємо збалансовану вибірку на тестову та тренінгову. Тепер у кожній вибірці буде 50% нулів та 50% одиниць, що допоможе покращити якість класифікації.

Створюємо та навчаємо класифікатор на тренінговій вибірці. Робимо прогноз цільової змінної на тестовій вибірці та оцінюємо якість класифікації (рис. 3.8).

```

      precision      recall      f1-score      support
0           0.62         0.81         0.70          90
1           0.72         0.49         0.59          89

      accuracy
macro avg      0.67         0.65         0.64         179
weighted avg    0.67         0.65         0.64         179

AUC-ROC: 0.6527465667915107

```

Рисунок 3.8 – Результати моделі на збалансованій вибірці, не включаючи змінні «*campaign*» та «*duration*»

Джерело: розроблено автором

Отже, тепер точність моделі за показником «ассигасу» показує більш правдоподібний результат (65%), ніж оцінка моделі, побудованої на «сирих» даних (див. рис. 3.3), де повна точність моделі становила 90%. Дане явище ми можемо пояснити тим, що ассигасу оцінює точність моделі, залежачи від рівня відсікання. Саме тому було прийняте рішення спиратись на результати показника інтегральної точності моделі – «*roc_auc*». Порівняємо результати двох моделей більш ретельно (див. рис. 3.3; 3.8):

- За метрикою *roc_auc* якість моделі на збалансованій вибірці та не включаючи змінні «*campaign*» та «*duration*» зросла на 9% та становить 65,27%.
- Точність класифікації стану кредиту становить 62% для тих, хто не купив депозит, та 72%, для тих, хто купив. Отже, тепер класифікатор краще прогнозує клієнтів, що купляють депозитний продукт (при $y = 1$) – 72%, проте 37% клієнтів, що не купили депозитний продукт, було класифіковано неправильно.
- Показник *recall* вказує на відношення правильно передбачених позитивних значень до фактичних позитивних значень. За даним показником класифікатор краще працює на негативних спостереженнях (коли $y = 0$), адже *recall* є вищим на 32%, ніж при $y = 1$.

Побудуємо та зробимо візуалізацію матриці помилок для щойно побудованої логістичної регресії (рис. 3.9).

За матрицею помилок бачимо, що класифікатор не є дуже якісним, адже хибно класифікував 45 клієнтів з тих, хто купив продукт, та 17 клієнтів з тих, хто не купив депозитний продукт (рис. 3.9).

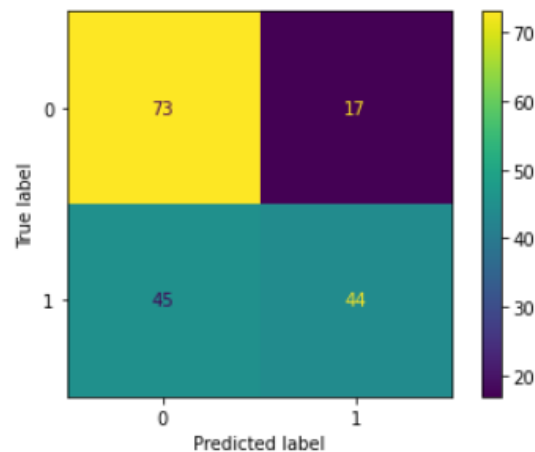


Рисунок 3.9 – Матриця помилок для моделі на збалансованій вибірці, не включаючи змінні «*campaign*» та «*duration*»

Джерело: розроблено автором

Таким чином, можемо сказати, що дана модель краще класифікує клієнтів, що не купляють депозитний продукт, проте все одно загальна якість класифікації залишається низькою. Незважаючи на це, якість цієї моделі зросла на 9% (рис. 3.8), в порівнянні з логістичною регресією, побудованою на «сирих» даних (рис. 3.3). У наступних пунктах спробуємо покращити класифікацію клієнтів за рахунок застосування різних процедур відбору найбільш значущих факторів з їх повної множини.

3.3 Побудова логістичних регресій із застосуванням процедур відбору значущих факторів

3.3.1 Backward Feature Selection із збалансованою вибіркою

Актуальність та доречність використання Stepwise Feature Selection (Backward та Forward Feature Selection) обґрунтована в розділі 1 (див. п. 1.3 «Stepwise Feature Selection для логістичної регресії»). Дані методи

використовуються для відбору тільки значущих змінних, які покращуватимуть якість класифікації та не спричинятимуть перенавчання моделі.

Для використання вибору змінних у зворотному напрямку виконуються такі дії:

1. Визначення матриці незалежних змінних та вектор цільової змінної (y). Виключаємо з аналізу також змінну «*campaign*» та «*duration*», які відповідають за кількість контактів з клієнтом під час даної кампанії та тривалістю контакту відповідно.
2. Розбиваємо вибірку та тестову та тренінгову. На цих вибірках залишатиметься однакова кількість нулів та одиниць, адже Backward Feature Selection будуватиметься на збалансованій вибірці.
3. Підключаємо клас «*SequentialFeatureSelector*», за допомогою якого будемо робити покроковий вибір кращих змінних, використовуючи логістичну регресію та включаємо такі параметри:
 - `k_features=1, len(X.columns)` – включаємо до покрокового відбору даних усі незалежні змінні (`X.columns`);
 - `forward=False` – ми будуємо зворотній покроковий відбір змінних (Backward Feature Selection), а параметр «`forward`» відповідає за прямий відбір даних (Forward Feature Selection);
 - `verbose=2` – даний параметр допомагає подивитись кожний етап та набір змінних, на яких навчається модель;
 - `scoring='roc_auc'` - показує, яка ймовірність того, що випадково вибраний позитивний випадок буде оцінений вище, ніж випадково вибраний негативний випадок. Використовується як метрика точності моделі.
 - `cv=5` – виконуємо п'ятикратну (за замовчуванням) крос-валідацію.
4. Отримуємо перелік рекомендованих до включення змінних та точність моделі: *age, job, default, housing, contact, month, pdays, previous, poutcome*. Код написання Backward Feature Selection із збалансованою вибіркою та процес навчання моделі можна переглянути на рис. Д.1 Додатку Д.

Backward Feature Selection рекомендує включати до моделі такі змінні: вік клієнта; професія; наявність кредитного дефолту; вид контакту; місяць; кількість днів, що минули з моменту останнього контакту та загальна кількість контактів, здійснених до цієї кампанії; загальну кількість контактів, здійснених до цієї кампанії та результат попередньої маркетингової кампанії.

Іншими словами, включити до моделі усі змінні, крім освіти клієнта (*education*); сімейного стану (*marital*); наявності персонального кредиту (*loan*) та останнього контактного дня тижня (*day_of_week*).

За *roc_auc* метрикою точність логістичної регресії з використанням даного набору змінних дорівнює 0.75025. Класифікація покращилась на 10% в порівнянні із логістичною регресією на збалансованій вибірці, включаючи усі змінні.

3.3.2 Forward Feature Selection із збалансованою вибіркою

Для використання прямого вибору змінних виконуються такі ж дії, як і для Backward Feature Selection (див. п. 1.3 «Stepwise Feature Selection для логістичної регресії»), проте при виконанні прямого вибору змінних, задаємо параметр *direction = forward*. Код написання Forward Feature Selection із збалансованою вибіркою та процес навчання моделі можна переглянути на рис. Д.2 Додатку Д.

Forward Feature Selection рекомендує включати до моделі такі змінні: *age*, *job*, *default*, *housing*, *contact*, *month*, *previous*, *poutcome*.

Тобто вік клієнта, професію, наявність кредитного дефолту, наявність кредиту на житло, спосіб контакту, останній контактний місяць, загальну кількість контактів, здійснених до цієї кампанії та результат попередньої маркетингової кампанії.

Іншими словами, Forward Feature Selection включає до аналізу ті ж змінні, що й Backward Feature Selection, проте без змінної «*pdays*», що відповідає за кількість днів, що минули з моменту останнього контакту з попередньої кампанії.

Точність моделі з використанням цих змінних дорівнюватиме 0.75025. Такою ж оцінкою регресії володіє модель із змінними, обраними методом Backward Feature Selection. Отже, виключення або включення змінної «*pdays*» не впливає на якість моделі.

3.4 Оптимізація біннінгу даних та вибір найкращої моделі

3.4.1 Покращення біннінгу даних за допомогою параметра «*cat_cutoff*»

За допомогою параметру «*cat_cutoff*» для кожної категоріальної змінної можна створити окрему категорію, для яких частка входжень буде нижче, чим значення *cat_cutoff*. Методом підбору оптимального значення встановили для кожної категоріальної змінної *cat_cutoff* = 0.05. Розглянемо на прикладі для змінної «*job*»:

- Додаємо параметр «*cat_cutoff* = 0.05» при дискретизації змінної «*job*» відповідно до бінарної таргетованої змінної «*y*» (рис. 3.10).

```
from optbinning import OptimalBinning
variable = 'job'

optb = OptimalBinning(name=variable, dtype="categorical", solver="mip", cat_cutoff=0.05)
optb.fit(x2, y)
```

```
▼ OptimalBinning
OptimalBinning(cat_cutoff=0.05, dtype='categorical', name='job', solver='mip')
```

Рисунок 3.10 – Параметр *cat_cutoff*=0.05» для змінної «*job*»

Джерело: розроблено автором

- За допомогою функції «*opt.splits*» переглядаємо, яким чином утворили біни (рис. 3.11):.

```
optb.splits
```

```
[array(['blue-collar'], dtype=object),
 array(['services'], dtype=object),
 array(['management'], dtype=object),
 array(['technician'], dtype=object),
 array(['admin.'], dtype=object),
 array(['retired', 'self-employed', 'entrepreneur', 'unemployed',
       'housemaid', 'student', 'unknown'], dtype=object)]
```

Рисунок 3.11 – Утворені категорії для змінної «job»

Джерело: розроблено автором

Останній бін містить 7 видів професій клієнтів, включаючи: пенсійних клієнтів, фрілансерів, перекладачів, безробітних, домогосподарів, студентів та клієнтів з невідомою професією. Даних професій по всьому датасету сумарно менше, ніж 5%, саме тому біннінг записав їх в одну категорію.

Подивимось як змінились якості моделей після додавання параметру «*cat_cutoff*», що становить 5% до біннінгу даних:

- Логістична регресія, побудована на всіх змінних та не збалансованій вибірці: метрика «*roc_auc_score*» зросла на 5% та тепер становить 0.613;
- Логістична регресія на збалансованій вибірці, не включаючи змінні «*campaign*» та «*duration*»: метрика «*roc_auc_score*» вказує, що якість моделі зменшилась на 1,7% і тепер становить 0.635. Дана модель залишається кращою за попередню логістичну регресію на 2,2%;
- Логістична регресія методом Backward Feature Selection на збалансованій вибірці та не включаючи змінні «*campaign*» та «*duration*»: даний метод після додавання параметру «*cat_cutoff*» виключив із моделі три змінні (професію клієнта (*job*), наявність кредитного дефолту (*default*) та загальну кількість контактів, здійснених до цієї кампанії (*previous*)). Оцінка якості даної моделі зросла 1,788% та тепер становить 0.7681 (за показником *roc_auc*);
- Логістична регресія методом Forward Feature Selection на збалансованій вибірці, не включаючи змінні «*campaign*» та «*duration*»: у даному випадку послідовний відбір змінних показав результат такий же, як Backward Feature Selection після додавання параметру «*cat_cutoff*» до біннінгу даних. Тобто

рекомендований набір даних співпадає, а, отже, логістична регресія має однаковий показник якості - 0.7681 (за показником *roc_auc*).

3.4.2 Узагальнення результатів та вибір найкращої моделі

Показники ефективності побудованих у бакалаврській роботі 8 логістичних моделей наведено на рис. 3.12.

Модель	Показник ефективності моделі (<i>auc_roc</i>)
<i>До балансування вибірки</i>	
1. Логістична регресія на "сирих даних"*	0,5619 (56,19%)
<i>Після балансування вибірки (50% на 50%) та виключенням змінних «campaign» та «duration»</i>	
2. Логістична регресія	0,6527 (65,27%)
3. Логістична регресія методом Backward Feature Selection із запропонованим набором даних**	0,7525 (75,25%)
4. Логістична регресія методом Forward Feature Selection із запропонованим набором даних***	0,7525 (75,25%)
Параметр <i>cat_cutoff=0,05</i>	
<i>До балансування вибірки</i>	
5. Логістична регресія на "сирих даних"	0,6132 (61,32%)
<i>Після балансування вибірки (50% на 50%) та виключенням змінних «campaign» та «duration»</i>	
6. Логістична регресія	0,6352 (63,52%)
7. Логістична регресія методом Backward Feature Selection із запропонованим набором даних****	0,7681 (76,81%)
8. Логістична регресія методом Forward Feature Selection із запропонованим набором даних****	0,7681 (76,81%)

Рисунок 3.12 – Результати моделей

Джерело: розроблено автором

* - дані, що були отримані після категоризації та до балансування вибірки, із включенням до моделі усіх запропонованих змінних; ** - метод Backward Feature Selection запропонував такий набір змінних: *age, job, default, housing, contact, month, pdays, previous, poutcome*; *** - метод Forward Feature Selection запропонував такий набір змінних: *age, job, default, housing, contact, month, previous, poutcome*; **** - після додавання параметру «*cat_cutoff*» Backward та Forward Feature Selections запропонували однаковий набір змінних: *age, loan, contact, month, pdays, poutcome*.

Отже, найефективнішими моделями виявились дві, що утворились за таких умов:

- проведення балансування вибірки так, щоб сума одиниць (позитивних подій) дорівнювала сумі нулів (негативних подій);

- додавання параметру «*cat_cutoff*» до біннінгу даних. Тобто віднесення спостережень, частка яких займала менше 5% по всій вибірці, до окремої категорії;
- застосування методів Stepwise Feature Selection (Backward та Forward Feature Selections) до побудови логістичної регресії.

Зробивши усі дії, що описані вище, нам вдалося підвищити якість класифікації з 56,19% до 76,81%, тобто на 20,62 п.п. (або на 36,7%), що є гарним результатом. Найкращі моделі (сьома та восьма, див. рис. 12) включають такі змінні як: вік клієнта; інформацію про особистий кредит (так, є чи невідомо); вид контакту (за допомогою мобільного телефону чи стільникового); останній контактний місяць року; кількість днів, що минули з моменту останнього контакту, та результат попередньої маркетингової кампанії банку (чи купив клієнт запропонований продукт). Загалом, із 15 запропонованих факторів, що описують клієнта, було досліджено, що для проведення ефективної класифікації, буде достатньо тільки 6 вище описаних факторів.

Найгіршою виявилась модель, що була побудована на категоризованих даних, проте без балансування вибірки – оскільки кількість негативних результатів (не продажі депозитного продукту) значно перевищувала кількість позитивних, тому точність класифікації була значно меншою.

ВИСНОВКИ

У даній кваліфікаційній роботі вирішено задачу моделювання ефективності контакту з клієнтом банку під час маркетингової кампанії з продажів депозитного продукту. У ході виконання дипломної роботи виконані такі завдання:

- Досліджено сучасний стан ринку банківських послуг, а також проведений критичний аналіз маркетингових кампаній в банківській сфері;
- Досліджений математичний інструментарій побудови моделей аналізу поведінки споживачів та виявлені основні закономірності та тенденції серед клієнтів банківської установи;
- У програмному середовищі Python було побудовано вісім логістичних регресій, за допомогою яких проводились експериментальні дослідження адекватності моделей на основі реальних даних португальської банківської установи.
- Сформульовані висновки та пропозиції щодо оптимізації роботи банку наведені нижче.

Після детального дослідницького аналізу даних, ми можемо висвітлити такі поради португальській банківській установі:

- *Щодо цільової аудиторії банку:* 35,5% усіх спостережень – це клієнти віком від 35 до 47 років, проте вони не є нашою цільовою аудиторією. За показником Event Rate (частота подій) було досліджено, що люди, віком від 57 років і вище (тобто пенсійного віку), більш схильні до придбання запропонованого продукту. Такі клієнти складають лише 5,4% вибірки, тому для максимізації цього числа слід розробляти маркетингові кампанії, які будуть розраховані на залучення клієнтів саме пенсійного віку.
- *Щодо сімейного стану та кредитів клієнтів:* 60% спостережень складають одружені люди. Аналіз показав, що одружені люди, які мають персональний кредит, скоріше за все, не погодяться на пропозицію. Вони будуть більш ретельно підходити до питання, адже, можливо, мають відповідальність у

вигляді родини. На відміну від цього, неодружені клієнти, що займають 28% усіх спостережень, більш схильні до купівлі продукту. Також володіння житловим кредитом притаманне 52% вибірки клієнтів, на відміну від особистого кредиту – тільки 19% мають такий вид кредиту.

- *Щодо клієнтів із кредитним дефолтом:* абсолютно усі клієнти, що не виплатили кредитні зобов'язання, не погодяться на поточну кампанію. У результаті, слід попередньо закупити кредитну історію клієнтів, щоб мати уявлення про потенційні можливості продажів.
- *Щодо днів тижня:* за гістограмою розподілу успішних подій за днями тижня, було проаналізовано, що суттєва різниця відсутня, проте все одно краще зосереджувати основну кількість контактів на понеділок та четвер.
- *Щодо кількості контактів, які необхідно зробити із клієнтом:* дослідження вказує, що більше 12 контактів із клієнтом – це марнуванням ресурсів банку. У середньому відбувається 2,5 контактів, тому вважатимемо, що для успішного результату знадобиться не більше, ніж 3 контакти із клієнтом банку.
- *Щодо тривалості контакту з клієнтами:* за аналізом зрозуміло, що розмова, яка триватиме менше однієї хвилини, не буде успішною для банку. У середньому телефонна розмова триває 4 хвилини, яких вистачає, щоб прослідкувати настрій та готовність клієнта до придбання депозиту. Довгий контакт, наприклад 54 хвилин, може не принести прибутку банку, а тільки змарнувати ресурси, тому слід раціонально розподіляти час.

В процесі виконання дослідження був запропонований методологічний підхід до моделювання ефективності контакту з клієнтом банку, який складається з етапів категоризації, балансування вибірки, відбору значущих факторів, що дали можливість побудувати комплекс математичних моделей, провести їх експериментальне дослідження та визначити доцільність застосування різноманітних процедур задля отримання найбільш адекватного прогнозу. Найбільш ефективні модель включала такі змінні, як: вік клієнта; інформацію про особистий кредит (так, є чи невідомо); вид контакту (за допомогою мобільного

телефону чи стільникового); останній контактний місяць року; кількість днів, що минули з моменту останнього контакту, та результат попередньої маркетингової кампанії банку (чи купив клієнт запропонований продукт). Загалом, із 15 запропонованих факторів, що описують клієнта, для проведення ефективної класифікації виявилось достатньо тільки 6 вище описаних факторів.

Найгіршою виявилась модель, що була побудована на категоризованих даних, проте без балансування вибірки, адже кількість негативних результатів (не продажі депозитного продукту) значно перевищувала кількість позитивних, через що точність класифікації була значно меншою.

Тож можна зробити висновок, що усі виконані етапи дозволили нам побудувати ефективну модель класифікації, яка в майбутньому допоможе банківській установі знизити рівень витрат та ресурсів, а також оптимізувати роботу та підвищити прибуток банку.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Sheth, J. N., & Sisodia, R. S. (2002). Marketing productivity: Issues and analysis. *Journal of Business Research*, 55(5), 349-362. [https://doi.org/10.1016/S0148-2963\(00\)00164-8](https://doi.org/10.1016/S0148-2963(00)00164-8)
2. Hooley, G. J., Greenley, G. E., Cadogan, J. W., & Fahy, J. (2005). The performance impact of marketing resources. *Journal of Business Research*, 58(1), 18-27. [https://doi.org/10.1016/S0148-2963\(03\)00109-7](https://doi.org/10.1016/S0148-2963(03)00109-7)
3. Csikósová, A., Čulková, K., & Janošková, M. (2016). Evaluation of quantitative indicators of marketing activities in the banking sector. *Journal of Business Research*, 69(11), 5028-5033. <https://doi.org/10.1016/j.jbusres.2016.04.075>
4. Krizanová, A., Lăzăroiu, G., Gajanová, L., Klietková, J., Nadanyiova, M., & Moravcikova, D. (2019). The Effectiveness of Marketing Communication and Importance of Its Evaluation in an Online Environment. *Sustainability*, 11(24), 7016. <https://doi.org/10.3390/su11247016>
5. Kumar, V., & Mittal, S. (2020). Mobile marketing campaigns: practices, challenges and opportunities. *International Journal of Business Innovation and Research*, 21(4), 523-539. <https://doi.org/10.1504/IJBIR.2020.105996>
6. Priyanka, D., (2015). Banking & Financial Institutions. Banking System. RDS College, Muzzarpur. <http://www.rdscollege.ac.in/studymaterial/1596117233.pdf>
7. Банківська система. (2016). ВУЕ. https://vue.gov.ua/Банківська_система
8. Hemming, G., (2020). 7 Types of Banks and Their Features. *ABC Finance Ltd.* ABC Finance. <https://abcfinance.co.uk/bank/types/>
9. Ebeling, A. (2022). European investment bank loan appraisal, the EU climate bank. *International Economics*, 172, 203-216. <https://doi.org/10.1016/j.inteco.2022.09.003>
10. Sakawa, H., Watanabel, N., Yamauchi, S., & Liu, R. (2023). The effect of Tobin's q on investment in a bank-based financial system: Evidence from Japan.

- Pacific-Basin Finance Journal*, 77, 101880.
<https://doi.org/10.1016/j.pacfin.2022.101880>
11. Bossone, B. (2021). Commercial bank seigniorage and the macroeconomy. *International Review of Financial Analysis*, 76, 101775.
<https://doi.org/10.1016/j.irfa.2021.101775>
 12. Ksoll, C., Lilleør, H. B., Lønborg, J. H., & Rasmussen, O. D. (2016). Impact of Village Savings and Loan Associations: Evidence from a cluster randomized trial. *Journal of Development Economics*, 120, 70-85.
<https://doi.org/10.1016/j.jdeveco.2015.12.003>
 13. de França Filho, G. C., Júnior, J. T. S., & Rigo, A. S. (2012). Solidarity finance through community development banks as a strategy for reshaping local economies: Lessons from Banco Palmas. *Revista de Administração*, 47(3), 500-515. <https://doi.org/10.5700/rausp1054>
 14. Chaudhary, V., Mandaviya, M., Sameen, H. S., Maroor, J. P., & Bharti, A. (2022). A study on effect of various demographic factors on preference of consumers towards online banking usage. *Materials Today: Proceedings*, 51, 84-87. <https://doi.org/10.1016/j.matpr.2021.04.560>
 15. Інтернет-джерело "Національний банк України" <https://bank.gov.ua/>
 16. Banks: Meaning, Examples, Functions & Characteristics. StudySmarter UK. <https://www.studysmarter.co.uk/explanations/macroeconomics/financial-sector/banks/>
 17. Mirković, M. & Kulina, D. (2016). Marketing Communications as Important Segment of the Marketing Concept. *ECONOMICS*, 4(1) 165-180. <https://doi.org/10.1515/eoik-2015-0031>
 18. Abdul, R. B. (2018). Optimization of Bank Expenses on Marketing Communications. *Financial Markets, Institution and Risks*, 62(3), 5-14. [https://doi.org/10.21272/fmir.2\(1\).5-14.2018](https://doi.org/10.21272/fmir.2(1).5-14.2018)
 19. Cătălin-Valeriu Curmei. (2021). Some Considerations on Integrating Corporate Social Responsibility in the Marketing Strategy of Banking Institutions. In

- Resilience and Economic Intelligence Through Digitalization and Big Data Analytics*, 350-358. Sciendo. <https://doi.org/10.2478/9788366675704-035>
20. Вальд Б.О. (2022). Більше мільярда на рекламу: хто з українських банків торік найбільше витрачав на своє просування. *UCAP.IO - український капітал, юкап, головне про економіку, бізнес, державу*. <https://ucap.io/bilshe-milyarda-na-reklamu-hto-z-ukrayinskyh-bankiv-torik-najbilshe-vytrachav-na-svoye-prosuvannya/>
 21. What is Exploratory Data Analysis? *IBM*. IBM - Deutschland IBM. <https://www.ibm.com/topics/exploratory-data-analysis>
 22. Chakri, P., Pratap, S., & Gouda, S. K. (2023). An exploratory data analysis approach for analyzing financial accounting data using machine learning. *Decision Analytics Journal*, 7, 100212. <https://doi.org/10.1016/j.dajour.2023.100212>
 23. Md. Abdullah Al Mamun. (2023). Exploring the factors that affecting adoption of mobile banking in Bangladesh. *Exploring the factors that affecting adoption of mobile banking in Bangladesh*, 8, 66-69. <https://doi.org/10.5038/2640-6489.8.1.1235>.
 24. What is Exploratory Data Analysis? *GeeksforGeeks*. <https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>
 25. Kleban, Y. (2019). Studying the methods of data transformation in the context of increasing the effectiveness of credit scoring models. *Neuro-Fuzzy Modeling Techniques in Economics*, 8, 94-123. <http://doi.org/10.33111/nfmte.2019.094>
 26. Jopia, H. (2015, March 24). *R Package 'smbinning': Optimal Binning for Scoring Modeling*. Retrieved from <https://blog.revolutionanalytics.com/2015/03/r-packagesmbinning-optimal-binning-for-scoring-modeling.html>.
 27. Kolyada, Y. V., & Bondar, V. A. (2016). Binninh u neyromerezhevykh skorynhovykh modelyakh. *Neyro-nechitki tekhnolohiyi modelyuvannya v ekonomitsi (Neuro-Fuzzy Modeling Techniques in Economics)*, 5, 60–80 [in Ukrainian] <https://doi.org/10.33111/nfmte.2016.060>

28. Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. New Jersey, NJ: John Wiley and Sons.
29. Pyrohov, V., & Turchenko, S. (2020). Research of the dependence of the efficiency of modeling the creditworthiness of borrowers on the method of forming a control subset. *Neuro-Fuzzy Modeling Techniques in Economics*, 9, 156-174. <http://doi.org/10.33111/nfmte.2020.156>
30. Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1), 224-238. <https://doi.org/10.1016/j.ijforecast.2011.07.006>
31. Fair Isaac Corporation. (2014). *Building Powerful, Predictive Scorecards: An overview of Scorecard module for FICO Model Builder*. Retrieved from http://www.fico.com/en/wp-content/secure_upload/Building_Powerful_Predictive_Scorecards_1991WP.pdf.
32. Swaminathan, S. (2018). Logistic Regression — Detailed Overview. *Medium*. <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>
33. Stepwise regression. (2023, January 25). In *Wikipedia*. https://en.wikipedia.org/wiki/Stepwise_regression
34. UCI Machine Learning Repository: Bank Marketing Data Set. (database). <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

ДОДАТКИ

Додаток А

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	y
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	no
57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	no
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	no
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	no
56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	no
45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0	nonexistent	no
59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent	no
41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217	1	999	0	nonexistent	no
24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent	no
25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent	no
41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55	1	999	0	nonexistent	no
25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent	no
29	blue-collar	single	high.school	no	no	yes	telephone	may	mon	137	1	999	0	nonexistent	no
57	housemaid	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1	999	0	nonexistent	no
35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	146	1	999	0	nonexistent	no
54	retired	married	basic.9y	unknown	yes	yes	telephone	may	mon	174	1	999	0	nonexistent	no
35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	312	1	999	0	nonexistent	no
46	blue-collar	married	basic.6y	unknown	yes	yes	telephone	may	mon	440	1	999	0	nonexistent	no
50	blue-collar	married	basic.9y	no	yes	yes	telephone	may	mon	353	1	999	0	nonexistent	no
39	management	single	basic.9y	unknown	no	no	telephone	may	mon	195	1	999	0	nonexistent	no
30	unemployed	married	high.school	no	no	no	telephone	may	mon	38	1	999	0	nonexistent	no
55	blue-collar	married	basic.4y	unknown	yes	no	telephone	may	mon	262	1	999	0	nonexistent	no
55	retired	single	high.school	no	yes	no	telephone	may	mon	342	1	999	0	nonexistent	no
41	technician	single	high.school	no	yes	no	telephone	may	mon	181	1	999	0	nonexistent	no
37	admin.	married	high.school	no	yes	no	telephone	may	mon	172	1	999	0	nonexistent	no
35	technician	married	university.degree	no	no	yes	telephone	may	mon	99	1	999	0	nonexistent	no

Рисунок А.1 – База даних, взята із сховища машинного навчання

Джерело: Розроблено автором за даними [31]

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	y
30	blue-collar	married	basic.9y	no	yes	no	cellular	may	fri	487	2	no	0	nonexistent	0
39	services	single	high.school	no	no	no	telephone	may	fri	346	4	no	0	nonexistent	0
25	services	married	high.school	no	yes	no	telephone	jun	wed	227	1	no	0	nonexistent	0
38	services	married	basic.9y	no	unknown	unknown	telephone	jun	fri	17	3	no	0	nonexistent	0
47	admin.	married	university.degree	no	yes	no	cellular	nov	mon	58	1	no	0	nonexistent	0
32	services	single	university.degree	no	no	no	cellular	sep	thu	128	3	no	2	failure	0
32	admin.	single	university.degree	no	yes	no	cellular	sep	mon	290	4	no	0	nonexistent	0
41	entrepreneur	married	university.degree	unknown	yes	no	cellular	nov	mon	44	2	no	0	nonexistent	0
31	services	divorced	professional.course	no	no	no	cellular	nov	tue	68	1	no	1	failure	0
35	blue-collar	married	basic.9y	unknown	no	no	telephone	may	thu	170	1	no	0	nonexistent	0
25	services	single	basic.6y	unknown	yes	no	cellular	jul	thu	301	1	no	0	nonexistent	0
36	self-employed	single	basic.4y	no	no	no	cellular	jul	thu	148	1	no	0	nonexistent	0
36	admin.	married	high.school	no	no	no	telephone	may	wed	97	2	no	0	nonexistent	0
47	blue-collar	married	basic.4y	no	yes	no	telephone	jun	thu	211	2	no	0	nonexistent	0
29	admin.	single	high.school	no	no	no	cellular	may	fri	553	2	no	0	nonexistent	0
27	services	single	university.degree	no	no	no	cellular	jul	wed	698	2	no	0	nonexistent	0
44	admin.	divorced	university.degree	no	no	no	cellular	jul	wed	191	6	no	0	nonexistent	0
46	admin.	divorced	university.degree	no	yes	no	telephone	jul	mon	59	4	no	0	nonexistent	0
45	entrepreneur	married	university.degree	unknown	yes	yes	cellular	aug	mon	38	2	no	0	nonexistent	0
50	blue-collar	married	basic.4y	no	no	yes	cellular	jul	tue	849	1	no	0	nonexistent	1
55	services	married	basic.6y	unknown	yes	no	cellular	jul	tue	326	6	no	0	nonexistent	0
39	technician	divorced	high.school	no	no	no	cellular	mar	mon	222	1	12	2	success	1
29	technician	single	university.degree	no	yes	yes	cellular	aug	wed	626	3	no	0	nonexistent	0
40	management	married	high.school	no	no	yes	cellular	aug	wed	119	1	no	0	nonexistent	0
44	technician	married	professional.course	unknown	yes	no	telephone	may	fri	388	7	no	0	nonexistent	0
38	technician	married	professional.course	no	yes	no	cellular	aug	mon	479	1	no	0	nonexistent	1

Рисунок А.2 – Відредагована база даних

Джерело: Розроблено автором

Додаток Б

	y	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutc
0	0	0.121878	0.067873	0.100438	0.074913	0.119759	0.109425	0.110481	0.140271	0.064586	0.108073	0.240157	0.112608	0.087648	0.081749	0.08
1	0	0.083959	0.089059	0.132697	0.104235	0.119759	0.108755	0.110481	0.051125	0.064586	0.108073	0.158392	0.109966	0.087648	0.081749	0.08
2	0	0.131579	0.089059	0.100438	0.104235	0.119759	0.109425	0.110481	0.051125	0.126415	0.103145	0.080997	0.123583	0.087648	0.081749	0.08
3	0	0.083959	0.089059	0.100438	0.074913	0.119759	0.085714	0.085714	0.051125	0.126415	0.108073	0.005333	0.111111	0.087648	0.081749	0.08
4	0	0.083959	0.130435	0.100438	0.128956	0.119759	0.109425	0.110481	0.140271	0.096413	0.113450	0.005333	0.123583	0.087648	0.081749	0.08
5	0	0.121878	0.089059	0.132697	0.128956	0.119759	0.108755	0.110481	0.140271	0.443350	0.111628	0.035088	0.111111	0.087648	0.266779	0.14
6	0	0.121878	0.130435	0.132697	0.128956	0.119759	0.109425	0.110481	0.140271	0.443350	0.113450	0.115607	0.109966	0.087648	0.081749	0.08
7	0	0.083959	0.134969	0.100438	0.128956	0.062267	0.109425	0.110481	0.140271	0.096413	0.113450	0.005333	0.112608	0.087648	0.081749	0.08
8	0	0.121878	0.089059	0.091928	0.119626	0.119759	0.108755	0.110481	0.140271	0.096413	0.105826	0.005333	0.123583	0.087648	0.266779	0.14
9	0	0.052023	0.067873	0.100438	0.074913	0.062267	0.108755	0.110481	0.051125	0.064586	0.111628	0.066351	0.123583	0.087648	0.081749	0.08
10	0	0.131579	0.089059	0.132697	0.074561	0.062267	0.109425	0.110481	0.140271	0.081575	0.111628	0.115607	0.123583	0.087648	0.081749	0.08
11	0	0.083959	0.134969	0.132697	0.088578	0.119759	0.108755	0.110481	0.140271	0.081575	0.111628	0.066038	0.123583	0.087648	0.081749	0.08
12	0	0.083959	0.130435	0.100438	0.104235	0.119759	0.108755	0.110481	0.051125	0.064586	0.103145	0.021277	0.112608	0.087648	0.081749	0.08
13	0	0.083959	0.067873	0.100438	0.088578	0.119759	0.109425	0.110481	0.051125	0.126415	0.111628	0.080997	0.112608	0.087648	0.081749	0.08
14	0	0.121878	0.130435	0.132697	0.104235	0.119759	0.108755	0.110481	0.140271	0.064586	0.108073	0.240157	0.112608	0.087648	0.081749	0.08
15	0	0.131579	0.089059	0.132697	0.128956	0.119759	0.108755	0.110481	0.140271	0.081575	0.103145	0.510903	0.112608	0.087648	0.081749	0.08
16	0	0.083959	0.130435	0.091928	0.128956	0.119759	0.108755	0.110481	0.140271	0.081575	0.103145	0.066351	0.053942	0.087648	0.081749	0.08
17	0	0.083959	0.130435	0.091928	0.128956	0.119759	0.109425	0.110481	0.051125	0.081575	0.113450	0.005333	0.109966	0.087648	0.081749	0.08
18	0	0.083959	0.134969	0.100438	0.128956	0.062267	0.109425	0.102256	0.140271	0.100629	0.113450	0.005333	0.112608	0.087648	0.081749	0.08
19	1	0.110256	0.067873	0.100438	0.088578	0.119759	0.108755	0.102256	0.140271	0.081575	0.105826	0.510903	0.123583	0.087648	0.081749	0.08
20	0	0.110256	0.089059	0.100438	0.074561	0.062267	0.109425	0.110481	0.140271	0.081575	0.105826	0.158392	0.053942	0.087648	0.081749	0.08
21	1	0.083959	0.115774	0.091928	0.104235	0.119759	0.108755	0.110481	0.140271	0.443350	0.113450	0.080997	0.123583	0.625000	0.266779	0.64
22	0	0.121878	0.115774	0.132697	0.128956	0.119759	0.109425	0.102256	0.140271	0.100629	0.103145	0.510903	0.111111	0.087648	0.081749	0.08
23	0	0.083959	0.092593	0.100438	0.104235	0.119759	0.108755	0.102256	0.140271	0.100629	0.103145	0.021277	0.123583	0.087648	0.081749	0.08
24	0	0.083959	0.115774	0.100438	0.119626	0.062267	0.109425	0.110481	0.051125	0.064586	0.108073	0.158392	0.025532	0.087648	0.081749	0.08
25	1	0.083959	0.115774	0.100438	0.119626	0.119759	0.109425	0.110481	0.140271	0.100629	0.113450	0.240157	0.123583	0.087648	0.081749	0.08

Рисунок Б.1 –Біннінг даних

Джерело: Розроблено автором

Додаток В

	y	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	pou	
	1530	0	0.260090	0.206128	0.091904	0.074913	0.062189	0.109425	0.110481	0.051125	0.126415	0.108073	0.035088	0.109966	0.108521	0.081749	0.0
	131	0	0.083959	0.095137	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.081749	0.0
	4082	0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.051125	0.064586	0.111628	0.158392	0.112608	0.108521	0.081749	0.0
	3358	0	0.083959	0.095137	0.091904	0.132075	0.119759	0.107510	0.110481	0.051125	0.064586	0.105826	0.510903	0.053942	0.108521	0.081749	0.0
	3508	0	0.121878	0.065891	0.132697	0.104235	0.119759	0.109425	0.110481	0.051125	0.301435	0.111628	0.115607	0.123583	0.108521	0.081749	0.0
	2544	0	0.121878	0.115774	0.132697	0.119626	0.119759	0.109425	0.110481	0.140271	0.100629	0.103145	0.035088	0.111111	0.108521	0.081749	0.0
	672	0	0.260090	0.206128	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.100629	0.111628	0.005333	0.025532	0.108521	0.081749	0.0
	1130	0	0.110256	0.065891	0.132697	0.088578	0.119759	0.107510	0.110481	0.051125	0.064586	0.111628	0.115607	0.111111	0.108521	0.081749	0.0
	1021	0	0.052023	0.130435	0.132697	0.104235	0.119759	0.109425	0.100000	0.140271	0.064586	0.105826	0.066351	0.109966	0.108521	0.266779	0.2
	3838	0	0.110256	0.115774	0.100438	0.132075	0.062189	0.107510	0.110481	0.051125	0.126415	0.108073	0.005333	0.111111	0.108521	0.081749	0.0
	3734	0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.096413	0.113450	0.158392	0.112608	0.108521	0.081749	0.0
	2587	0	0.083959	0.065891	0.100438	0.074913	0.119759	0.109425	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.081749	0.0
	2432	0	0.083959	0.115774	0.100438	0.132075	0.062189	0.109425	0.100000	0.140271	0.301435	0.111628	0.035088	0.123583	0.108521	0.081749	0.0
	1669	0	0.083959	0.065891	0.100438	0.119626	0.119759	0.109425	0.110481	0.140271	0.301435	0.113450	0.021277	0.109966	0.108521	0.081749	0.0
	2988	0	0.083959	0.086957	0.100438	0.119626	0.119759	0.109425	0.110481	0.140271	0.081575	0.113450	0.005333	0.025532	0.108521	0.081749	0.0
	2030	0	0.131579	0.206128	0.132697	0.104235	0.119759	0.107510	0.110481	0.140271	0.064586	0.113450	0.240157	0.123583	0.108521	0.081749	0.0
	2209	0	0.121878	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.100629	0.108073	0.035088	0.112608	0.108521	0.081749	0.0
	2848	0	0.052023	0.065891	0.100438	0.074913	0.062189	0.107510	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.266779	0.2
	3566	0	0.121878	0.086957	0.132697	0.074913	0.119759	0.107510	0.110481	0.051125	0.064586	0.105826	0.021277	0.111111	0.108521	0.081749	0.0
	3479	0	0.121878	0.130435	0.100438	0.132075	0.119759	0.107510	0.100000	0.140271	0.100629	0.108073	0.005333	0.053942	0.108521	0.081749	0.0
	1952	0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.096413	0.113450	0.158392	0.123583	0.108521	0.081749	0.0
	2800	0	0.083959	0.095137	0.100438	0.132075	0.119759	0.107510	0.110481	0.140271	0.064586	0.113450	0.240157	0.123583	0.108521	0.081749	0.0
	2037	0	0.110256	0.065891	0.100438	0.074236	0.119759	0.107510	0.110481	0.140271	0.064586	0.103145	0.021277	0.123583	0.108521	0.081749	0.0
	3486	0	0.121878	0.065891	0.100438	0.088578	0.119759	0.107510	0.110481	0.051125	0.064586	0.113450	0.005333	0.112608	0.108521	0.081749	0.0
	3567	0	0.110256	0.065891	0.100438	0.088578	0.062189	0.109425	0.110481	0.140271	0.100629	0.111628	0.158392	0.123583	0.108521	0.081749	0.0

Рисунок В.1 – Проміжний датасет при балансуванні вибірки

Джерело: Розроблено автором

	y	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	pout	
	1530	0	0.260090	0.206128	0.091904	0.074913	0.062189	0.109425	0.110481	0.051125	0.126415	0.108073	0.035088	0.109966	0.108521	0.081749	0.08
	131	0	0.083959	0.095137	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.081749	0.08
	4082	0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.051125	0.064586	0.111628	0.158392	0.112608	0.108521	0.081749	0.08
	3358	0	0.083959	0.095137	0.091904	0.132075	0.119759	0.107510	0.110481	0.051125	0.064586	0.105826	0.510903	0.053942	0.108521	0.081749	0.08
	3508	0	0.121878	0.065891	0.132697	0.104235	0.119759	0.109425	0.110481	0.051125	0.301435	0.111628	0.115607	0.123583	0.108521	0.081749	0.08
	2544	0	0.121878	0.115774	0.132697	0.119626	0.119759	0.109425	0.110481	0.140271	0.100629	0.103145	0.035088	0.111111	0.108521	0.081749	0.08
	672	0	0.260090	0.206128	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.100629	0.111628	0.005333	0.025532	0.108521	0.081749	0.08
	1130	0	0.110256	0.065891	0.132697	0.088578	0.119759	0.107510	0.110481	0.051125	0.064586	0.111628	0.115607	0.111111	0.108521	0.081749	0.08
	1021	0	0.052023	0.130435	0.132697	0.104235	0.119759	0.109425	0.100000	0.140271	0.064586	0.105826	0.066351	0.109966	0.108521	0.266779	0.26
	3838	0	0.110256	0.115774	0.100438	0.132075	0.062189	0.107510	0.110481	0.051125	0.126415	0.108073	0.005333	0.111111	0.108521	0.081749	0.08
	3734	0	0.083959	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.096413	0.113450	0.158392	0.112608	0.108521	0.081749	0.08
	2587	0	0.083959	0.065891	0.100438	0.074913	0.119759	0.109425	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.081749	0.08
	2432	0	0.083959	0.115774	0.100438	0.132075	0.062189	0.109425	0.100000	0.140271	0.301435	0.111628	0.035088	0.123583	0.108521	0.081749	0.08
	1669	0	0.083959	0.065891	0.100438	0.119626	0.119759	0.109425	0.110481	0.140271	0.301435	0.113450	0.021277	0.109966	0.108521	0.081749	0.08
	2988	0	0.083959	0.086957	0.100438	0.119626	0.119759	0.109425	0.110481	0.140271	0.081575	0.113450	0.005333	0.025532	0.108521	0.081749	0.08
	2030	0	0.131579	0.206128	0.132697	0.104235	0.119759	0.107510	0.110481	0.140271	0.064586	0.113450	0.240157	0.123583	0.108521	0.081749	0.08
	2209	0	0.121878	0.130435	0.100438	0.132075	0.119759	0.109425	0.110481	0.140271	0.100629	0.108073	0.035088	0.112608	0.108521	0.081749	0.08
	2848	0	0.052023	0.065891	0.100438	0.074913	0.062189	0.107510	0.110481	0.140271	0.064586	0.111628	0.115607	0.123583	0.108521	0.266779	0.26
	3863	1	0.121878	0.086957	0.132697	0.104235	0.119759	0.109425	0.100000	0.140271	0.100629	0.103145	0.115607	0.123583	0.108521	0.266779	0.26
	3865	1	0.083959	0.130435	0.100438	0.132075	0.119759	0.107510	0.110481	0.140271	0.301435	0.105826	0.240157	0.123583	0.108521	0.081749	0.08
	3879	1	0.110256	0.086957	0.100438	0.074913	0.119759	0.109425	0.100000	0.140271	0.064586	0.113450	0.510903	0.053942	0.108521	0.266779	0.26
	3889	1	0.121878	0.130435	0.100438	0.132075	0.119759	0.109425	0.100000	0.140271	0.126415	0.105826	0.158392	0.123583	0.108521	0.081749	0.08
	3897	1	0.121878	0.086957	0.100438	0.119626	0.119759	0.109425	0.100000	0.140271	0.081575	0.111628	0.510903	0.109966	0.108521	0.081749	0.08
	3937	1	0.131579	0.206128	0.132697	0.132075	0.119759	0.107510	0.110481	0.140271	0.100629	0.111628	0.066351	0.123583	0.108521	0.081749	0.08
	3939	1	0.260090	0.095137	0.100438	0.088578	0.119759	0.109425	0.110481	0.140271	0.096413	0.105826	0.066351	0.112608	0.108521	0.081749	0.08
	3943	1	0.131579	0.130435	0.132697	0.132075	0.119759	0.109425	0.110481	0.140271	0.301435	0.108073	0.066038	0.112608	0.108521	0.081749	0.08

Рисунок В.2 – Збалансована вибірка

Джерело: Розроблено автором

```

lr = LogisticRegression(n_jobs=-1)
selector = SequentialFeatureSelector(lr,
                                   k_features=(1, len(X.columns)),
                                   forward=False,
                                   verbose=2,
                                   scoring='roc_auc',
                                   cv=5)

selector.fit(X_train, y_train)

[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 13 out of 13 | elapsed: 0.4s finished

[2023-05-26 11:40:24] Features: 12/1 -- score: 0.7499217527386541[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 12 out of 12 | elapsed: 0.3s finished

[2023-05-26 11:40:25] Features: 11/1 -- score: 0.7500978090766821[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 11 out of 11 | elapsed: 0.3s finished

[2023-05-26 11:40:25] Features: 10/1 -- score: 0.7501564945226918[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 10 out of 10 | elapsed: 0.2s finished

[2023-05-26 11:40:26] Features: 9/1 -- score: 0.7502543035993738[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 9 out of 9 | elapsed: 0.2s finished

[2023-05-26 11:40:26] Features: 8/1 -- score: 0.7502543035993738[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 8 out of 8 | elapsed: 0.2s finished

[2023-05-26 11:40:26] Features: 7/1 -- score: 0.7502151799687011[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 7 out of 7 | elapsed: 0.1s finished

[2023-05-26 11:40:26] Features: 6/1 -- score: 0.7499217527386541[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 6 out of 6 | elapsed: 0.1s finished

[2023-05-26 11:40:27] Features: 5/1 -- score: 0.7493544600938966[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 5 out of 5 | elapsed: 0.1s finished

[2023-05-26 11:40:27] Features: 4/1 -- score: 0.7463028169014085[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 4 out of 4 | elapsed: 0.0s finished

[2023-05-26 11:40:27] Features: 3/1 -- score: 0.7428990610328638[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 3 out of 3 | elapsed: 0.0s finished

[2023-05-26 11:40:27] Features: 2/1 -- score: 0.7183489827856026[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 co
ncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 2 out of 2 | elapsed: 0.0s finished

[2023-05-26 11:40:27] Features: 1/1 -- score: 0.6654929577464789

> SequentialFeatureSelector
> estimator: LogisticRegression
  > LogisticRegression

```

```

selected_features = selector.k_feature_names_
print(selected_features)

('age', 'job', 'default', 'housing', 'contact', 'month', 'pdays', 'previous', 'poutcome')

selected_score = selector.k_score_
print(selected_score)

0.7502543035993738

```

Рисунок Д.1 – Backward Feature Selection

Джерело: Розроблено автором

```

lr = LogisticRegression(n_jobs=-1)
selector = SequentialFeatureSelector(lr,
                                   k_features=(1, len(XFFS.columns)),
                                   forward=True,
                                   verbose=2,
                                   scoring='roc_auc',
                                   cv=5)

selector.fit(XFFS_train, yFFS_train)

[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 13 out of 13 | elapsed: 0.3s finished

[2023-05-26 11:40:28] Features: 1/13 -- score: 0.6654929577464789[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 12 out of 12 | elapsed: 0.3s finished

[2023-05-26 11:40:28] Features: 2/13 -- score: 0.7183489827856026[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 11 out of 11 | elapsed: 0.3s finished

[2023-05-26 11:40:28] Features: 3/13 -- score: 0.7428990610328638[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 10 out of 10 | elapsed: 0.3s finished

[2023-05-26 11:40:29] Features: 4/13 -- score: 0.7463028169014085[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 9 out of 9 | elapsed: 0.2s finished

[2023-05-26 11:40:29] Features: 5/13 -- score: 0.7493544600938966[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 8 out of 8 | elapsed: 0.2s finished

[2023-05-26 11:40:29] Features: 6/13 -- score: 0.7499217527386541[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 7 out of 7 | elapsed: 0.2s finished

[2023-05-26 11:40:30] Features: 7/13 -- score: 0.7502151799687011[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 6 out of 6 | elapsed: 0.2s finished

[2023-05-26 11:40:30] Features: 8/13 -- score: 0.7502543035993738[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 5 out of 5 | elapsed: 0.1s finished

[2023-05-26 11:40:30] Features: 9/13 -- score: 0.7502543035993738[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
oncurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 4 out of 4 | elapsed: 0.0s finished

[2023-05-26 11:40:30] Features: 10/13 -- score: 0.7501564945226918[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1
concurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 3 out of 3 | elapsed: 0.0s finished

[2023-05-26 11:40:30] Features: 11/13 -- score: 0.7500978090766821[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1
concurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 2 out of 2 | elapsed: 0.0s finished

[2023-05-26 11:40:30] Features: 12/13 -- score: 0.7499217527386541[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1
concurrent workers.
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s remaining: 0.0s
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 0.0s finished

[2023-05-26 11:40:30] Features: 13/13 -- score: 0.7491784037558685

> SequentialFeatureSelector
> estimator: LogisticRegression
  > LogisticRegression

```

```

selected_features_forward = selector.k_feature_names_
print(selected_features_forward)

('age', 'job', 'default', 'housing', 'contact', 'month', 'previous', 'poutcome')

selected_score_FFS = selector.k_score_
print(selected_score_FFS)

0.7502543035993738

```

Рисунок Д.2 – Forward Feature Selection

Джерело: Розроблено автором