Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра прикладної математики
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Застосування методів машинного навчання
для оцінювання кредитних ризиків
(тема)

Виконав:
студент _2_ курсу, групи _САУм-19-1
Коновалов В.С.
(прізвище, ініціали)
Спеціальність
124 Системний аналіз
(код і повна назва спеціальності)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма
Системний аналіз і управління
(повна назва освітньої програми)
Керівник доц. Гибкіна Н.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ПМ _____ Тевяшев А.Д.
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет _інформаційно-аналітичних технологій та менеджменту
Кафедра _прикладної математики
Рівень вищої освіти _другий (магістерський)
Спеціальність _124 Системний аналіз
(код і повна назва)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма _Системний аналіз і управління
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри ПМ _____
(підпис)
"____" _____ 2020 р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові _Коновалову Владиславу Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи _Застосування методів машинного навчання для оцінювання кредитних ризиків

затверджена наказом по університету від 23 жовтня 2020 р. № 1420 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 10 грудня 2020 р.

3. Вихідні дані до роботи _статистичні дані стосовно видачі та повернення кредитів, статистичні дані стосовно шахрайства під час оформлення заявки щодо отримання кредиту

4. Перелік питань, що потрібно опрацювати в роботі
1. Системний аналіз проблеми застосування методів бінарної класифікації у банківській діяльності
2. Вибір і обґрунтування методу розв'язання
3. Програмна реалізація
4. Результати обчислювального експерименту
5. Аналіз можливих застосувань

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій

1. Актуальність теми роботи
2. Постановка задачі
3. Системний аналіз проблеми
4. Метод чисельного аналізу
5. Результати обчислювального експерименту

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Підбір та вивчення технічної літератури за темою роботи	вересень 2020 р.	виконано
2	Вибір та обґрунтування методу	жовтень – листопад 2020 р.	виконано
3	Розробка алгоритму і програми	листопад – грудень 2020 р.	виконано
4	Проведення аналітичних досліджень та розрахунків	листопад – грудень 2020 р.	виконано
5	Робота над текстом пояснювальної записки	грудень 2020 р.	виконано
6	Представлення роботи на рецензію в ЕК	грудень 2020 р.	виконано

Дата видачі завдання _1_ вересня 2020 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Гибкіна Н.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 76 с., 8 табл., 23 рис., 4 дод., 18 джерел.

ЛОГІСТИЧНА РЕГРЕССІЯ, ЗМЕНШЕННЯ РОЗМІРНОСТІ, ROC-КРИВІ, ВАГОВІ КОЕФІЦІЄНТИ, БІНАРНА КЛАСИФІКАЦІЯ, МЕТОД ГОЛОВНИХ КОМПОНЕНТ, КРЕДИТНИЙ СКОРІНГ, FRAUD-СКОРІНГ.

Об’єкт дослідження – банківська система, яка складається з багатьох відділів, що працюють разом для надання послуг, пов’язаних з фінансовими операціями для існуючих чи нових клієнтів банку.

Мета дослідження – розв’язання задачі класифікації клієнтів банку.

Методи дослідження – методи зниження розмірності, методи бінарної класифікації.

У роботі проведено системний аналіз задачі бінарної класифікації клієнтів банку, які бажають отримати кредит, з метою максимізації відсотку клієнтів, що повернуть займані кошти банку, та виключення ймовірності того, що клієнт є шахраєм. Розв’язано задачу бінарної класифікації клієнтів за допомогою нейронної мережі. Із використанням тестової вибірки клієнтів були налаштовані вагові коефіцієнти для нейронної мережі. Були застосовані методи зниження розмірності для вибірок із значним обсягом даних. Також було розроблено програмний продукт, який дозволяє розрахувати ймовірність успішного прогнозування, заснованого на даних клієнта. За допомогою розробленого програмного продукту проведено ряд обчислювальних експериментів та виконано аналіз отриманих результатів.

ABSTRACT

Introductory note: 76 pages, 8 tables, 23 figures, 4 appendixes, 18 sources.

LOGISTIC REGRESSION, DIMENSIONAL REDUCTION, ROC-CURVE, WEIGHT COEFFICIENTS, BINARY CLASSIFICATION, MAIN COMPONENT METHOD, CREDIT SCORING, FRAUD-SCORING.

The research object is a banking system, which consists of many departments that work together to provide a variety of financial services to the bank's clients.

The purpose of the research is to solve the problem of client classification.

Methods of research – methods of reducing the dimension, methods of binary classification.

The system analysis of the problem of the binary classification of bank clients wishing to obtain a loan is carried out in the work, in order to maximize the likelihood that the client who received the loan will return it. The binary classification problem with the help of a neural network is solved. The weighting coefficients for the neural network were obtained using customer test data samples for clients. Dimension reduction methods were applied for samples with a large amount of data for binary classification methods. A software product has been developed, which allows obtaining the probability based on customer data. A number of computational experiments were carried out with the help of the developed software product and a numerical analysis of the results was performed.

ЗМІСТ

ВСТУП

Сучасні технології та можливості змінили підхід до організації діяльності підприємств і призводять до стрімкого збільшення об’ємів даних, які потрібно швидко опрацьовувати для надання послуг клієнтам. Розвиваються не тільки потужності пристроїв, але й технології, завдяки котрим ми працюємо з даними. Саме тому набувають популярності математичні напрями інформаційних технологій, що стрімко розвиваються останнім часом, а, в особливості, аналіз даних та методи машинного навчання. Завдяки аналізу даних з’явилась можливість обробляти не досяжні дотепер обсяги даних майже без втручання людини. Це збільшує продуктивність на підприємствах, та допомагає оптимізувати роботу за рахунок перекладання аналізу великого масиву даних на комп’ютерну техніку.

Схожі задачі обробки великих обсягів даних виникають у багатьох сферах діяльності, зокрема, у банківській діяльності під час аналізу потоків та операцій, балансу активів та пасивів тощо. Однією з найважливіших задач для банківської сфери є задача про надання кредитних послуг та інвестування коштів у бізнес-проекти.

Розв’язання цієї задачі останнім часом все більше і більше переноситься на сучасні технології та може бути здійснено методами машинного навчання, які з легкістю реалізують методи бінарної класифікації об’єктів. Розглянемо етапи видачі кредиту. Клієнт банку, який бажає отримати кредит, повинен надати до установи велику кількість різноманітних документів, посвідчень, довідок тощо. Кожен з документів містить унікальну статистичну інформацію, яка з різних боків характеризує даного клієнта (вік, стаж роботи на останньому місці роботи, заробітна плата, кількість дітей, умови проживання тощо). Після отримання банком всієї необхідної інформації ці документи потрібно проаналізувати і на основі цього видати висновки стосовно кожного клієнта. Очевидно, що ймовірність прийняття правильного рішення стосовно можливості видати кредит можна суттєво збільшити, якщо використовувати під час аналізу інформа-

цію щодо видачі і повернення аналогічних кредитів у минулому.

Дана атестаційна робота присвячена застосуванню методів машинного навчання для розв’язання задачі бінарної класифікації клієнтів банку за їх кредитоспроможністю та можливістю шахрайства.

Метою розділу «Системний аналіз проблеми застосування методів бінарної класифікації у банківській діяльності та постановка задачі дослідження» є дослідження методів бінарної класифікації, визначення їх переваг та недоліків для різних типів даних і подальшого їх використання для певних задач бінарної класифікації. В результаті аналізу необхідно побудувати вербальну, морфологічну, функціональну та інформаційну моделі системи, а також виконати аналіз проблеми розв’язання задачі.

1 СИСТЕМНИЙ АНАЛІЗ ПРОБЛЕМИ ЗАСТОСУВАННЯ МЕТОДІВ БІНАРНОЇ КЛАСИФІКАЦІЇ У БАНКІВСЬКІЙ ДІЯЛЬНОСТІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Системний аналіз предметної області

1.1.1 Вербальна модель системи

Об’єкт дослідження – банківська система, яка складається з багатьох відділів, що працюють разом для надання різноманітних послуг, пов’язаних з фінансовими операціями для існуючих чи нових клієнтів банку.

Предмет дослідження – методи бінарної класифікації, за допомогою яких стане можливим класифікація клієнтів банку стосовно можливості повернення кредиту з максимальною ймовірністю та зменшення можливих випадків шахрайства під час проведення банківських операцій.

Система є соціальною, бо її основу складають, у першу чергу, люди. Також, в силу того, що вона взаємодіє з навколишнім середовищем, вона є відкритою.

Системна модель об'єкта створюється з метою дослідження етапів побудови математичної моделі, яка буде здатна класифікувати клієнтів на благонадійних і не благонадійних, спираючись на інформацію з анкет, заповнених при звертанні до банківської установи за кредитом. Розв’язання задачі класифікації включає в себе побудову і навчання нейронної мережі, оцінку точності прогнозу нейронної мережі будемо здійснювати, порівнюючи прогнозоване нейронною мережею значення та відповідну інформацію для людей з тестової вибірки, які повернули чи не повернули кредит.

Головний вихід системи – клієнт з міткою класу, яка відносить його до однієї з двох груп людей: тих, що повернуть кредит та тих, що не зроблять цього на першому етапі, та виділяти потенційних шахраїв серед клієнтів на другому етапі. На входи системи подається інформація стосовно клієнтів, які бажа-

ють отримати кредит. Головний вплив на систему надають вагові коефіцієнти, отримані у процесі навчання нейронної мережі за допомогою тренувального набору даних, які відображають впливовість кожного інформаційного поля клієнта цього банку.

1.1.2 Морфологічний опис системи

Призначенням моделі є дослідження нейронних мереж і їх можливостей для подальшого використання при прогнозуванні ймовірностей успішного повернення кредитних позичок, максимізації фінансових прибутків банківськими установами шляхом збільшення відсотку людей, які повернули кредит, та мінімізації випадків шахрайства під час звернення до банку. Для реалізації цілей системи використовують різні ресурси, до яких відносяться: інформаційні, технологічні тощо.

Морфологічний опис системи почнемо з опису зовнішнього середовища, що наведений у Додатку А на рисунку А.1.

Зовнішнє середовище – це сукупність всіх об’єктів, які не потрапляють у межі системи, зміна властивостей яких впливає на систему, а також тих об’єктів, чиї властивості змінюються у результаті функціонування системи.

Об’єкти зовнішнього середовища:

а) особисті якості клієнта – забезпечують необхідні для проведення досліджень умови;

б) економіка – впливає на кількість людей, які бажатимуть отримати кредит, і період, на який буде видано кредит, а також суми, на які будуть видані кредити, і відсоток, під який люди будуть його отримувати, а також кількість можливих випадків шахрайства під час звернення до банку;

в) фінансовий стан банку – впливає на кількість можливих клієнтів, які можуть отримати кредит.

Розглянемо модель типу «чорний ящик» системи «Задача бінарної класи-

фікації клієнтів банку», яка представлена у Додатку А на рисунку А.2. Така модель є вихідною при побудові моделі складної системи, вона звертає увагу розробника на взаємодію системи з зовнішнім середовищем. Виходом цієї системи є ймовірність того, що кредитна позика буде успішно повернена або того, що клієнт виявиться шахраєм (в залежності від поставленої задачі), а вхід – це інформація стосовно цього клієнта. Зміст «чорного ящика» не розкривається, тому що увага зосереджена тільки на межах системи. Межі, в свою чергу, виділяють цілісність системи, відособленість її від зовнішнього середовища і взаємодію системи і середовища.

1.1.3 Функціональна модель системи

Розглянемо функціональну модель системи з точки зору аналітика – людини, яка займається безпосереднім керуванням цим фрагментом системи. Метою функціонального моделювання є декомпозиція системи з метою виділення окремих функцій, важливих з точки зору розв’язання поставленої задачі.

Перша діаграма в ієрархії діаграм – контекстна діаграма IDEF0, рисунок А.3, Додаток А. Вона відображує функціонування системи в цілому.

В рамках методології IDEF0 процес зображують у вигляді набору елементів, в яких налагоджена взаємодія між собою, а також показані ресурси, що споживаються кожною роботою.

Після того, як опис контенту проведено, проводиться побудова наступних діаграм в ієрархії, тобто здійснюється декомпозиція складної системи для більш детального розгляду функціональної частини. Кожен наступний рівень декомпозиції діаграми є більш детальним описом однієї з робіт на попередній діаграмі. У нашому випадку розглядається функціонування системи «Задача бінарної класифікації клієнтів банку, що бажають отримати кредит». Друга діаграма демонструє основні функції системи і їх деталізації за рівнями, рисунки А.3, А.4 і А.5 Додатка А. Функціональна модель будується аналогічно для задачі Fraud-скорінгу.

1.1.4 Інформаційна модель системи

Дана модель використовується для опису документообігу та обробки інформації і називається діаграмою потоків даних (DFD). За допомогою неї можна легко побачити, де і як зберігаються дані, а також як саме відбувається документообіг у системі. Ці діаграми також можуть бути використані для виявлення проблем зв’язаних з документообігом і також можуть бути використані для надання рекомендацій стосовно запобіжних дій.

DFD-діаграма першого рівня декомпозиції представлена на рисунку А.6 Додатка А.

1.2 Аналіз сценаріїв вирішення проблеми вибору методу для розв’язання задачі бінарної класифікації

1.2.1 Модель аналізу проблеми

Об’єктом дослідження є система, що містить певну проблему (ПС-система), яка була сформульована на основі нашої моделі. Проблемою в даному випадку є складність вибору методу бінарної класифікації в залежності від інформації стосовно кожного клієнта, внаслідок чого виникає складність у виборі і реалізації методу розв’язання задачі бінарної класифікації, що є впливовою частиною моделювання та аналізу процесів моделювання.

Рішення проблеми – це перетворення ПС-системи з початкового, незадовільного стану в бажаний або необхідний стан.

На просторі розглянутих об'єктів необхідно виділити ряд критеріїв, які, на нашу думку, будуть впливати на досягнення необхідного результату. Ці компоненти розбиваються на чотири групи:

а) точність;

б) час роботи;

в) ресурси ЕОМ;

г) вартість розробки і підтримки.

Обирати будемо з множин альтернатив:

а) логістична регресія;

б) k найближчих сусідів;

в) дерева рішень.

Побудуємо ієрархічну стуктуру, використовуючи метод парних порівнянь. Ієрархічна структура приведена на рис. 1.1.

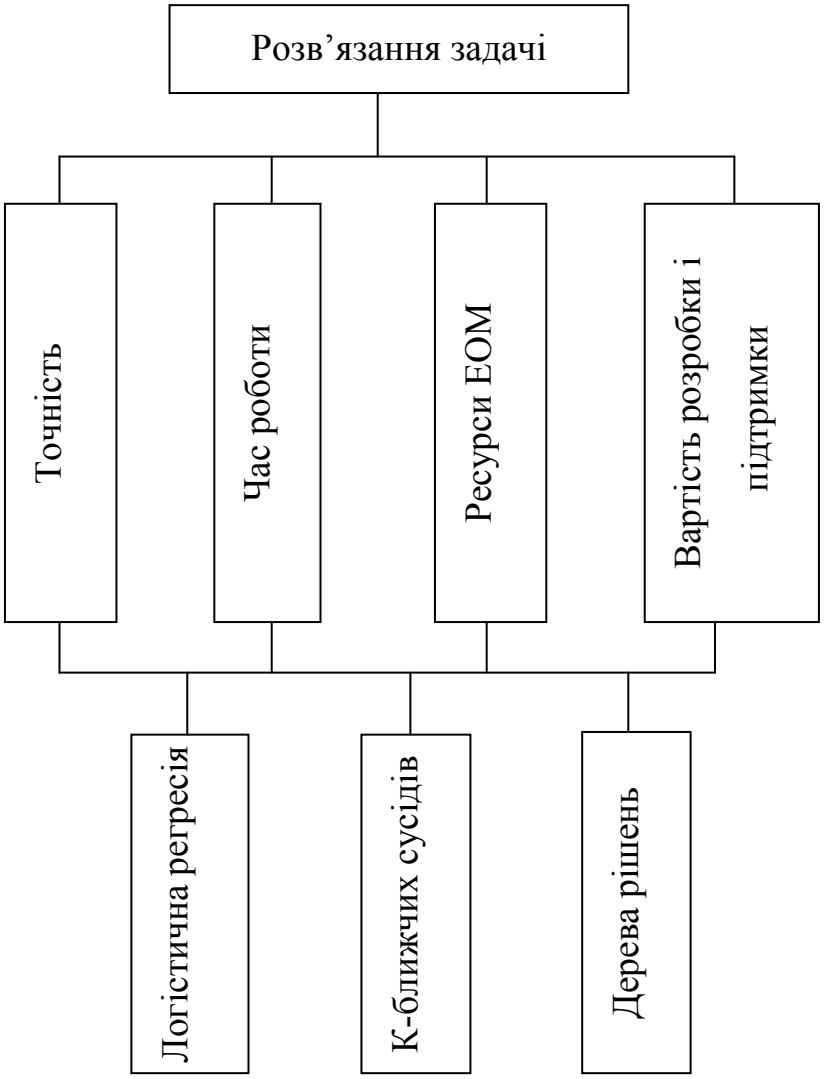

Рисунок 1.1 – Ієрархічна модель процесу аналізу розв'язання задачі

1.2.2 Оцінювання вектора пріоритетів незадоволеностей методом аналізу ієрархій

Побудуємо матрицю попарних порівнянь для елементів першого рівня ієрархії, тобто матрицю порівнянь важливості критеріїв за шкалою Т. Сааті, і розрахуємо вектор локальних пріоритетів критеріїв (таблиця 1.1).

Аналіз вектору локальних пріоритетів першого рівня показав, що найбільш впливовою компонентою виявилась точність $(W_1 = 0,467)$, наступна компонента – час роботи $(W_2 = 0,277)$, вартість $(W_3 = 0,160)$ та ресурси ЕОМ $(W_4 = 0,095)$. На рисунку 1.2 наведена діаграма пріоритетів.

Найбільше власне значення матриці суджень дорівнює $\lambda_{\max} = 4,0313$. Індекс узгодженості дорівнює $\text{ІУ} = 0,0104$ і відношення узгодженості $\text{ВУ} = 0,0116$, що дуже близько до 0,1, тому вважатимемо, що оцінки критеріїв були зроблені правильно. Зведемо отримані дані до однієї таблиці для більш детального і легкого аналізу отриманих результатів.

Таблиця 1.1 – Розрахунок вектору локальних пріоритетів критеріїв

	Точність	Час роботи	Ресурси ЕОМ	Вартість розробки	Середнє геометричне по рядках	Вектор пріоритетів
Точність	1	2	4	3	$x_1 = 2,21336$	$W_1 = 0,466849$
Час роботи	$\frac{1}{2}$	1	3	2	$x_2 = 1,31607$	$W_2 = 0,27759$
Ресурси ЕОМ	$\frac{1}{4}$	$\frac{1}{3}$	1	$\frac{1}{2}$	$x_3 = 0,45180$	$W_3 = 0,095295$
Вартість розробки	$\frac{1}{3}$	$\frac{1}{2}$	2	1	$x_3 = 0,759836$	$W_3 = 0,160267$
					$\sum_{i=1}^{3} x_i = 4,74107$	

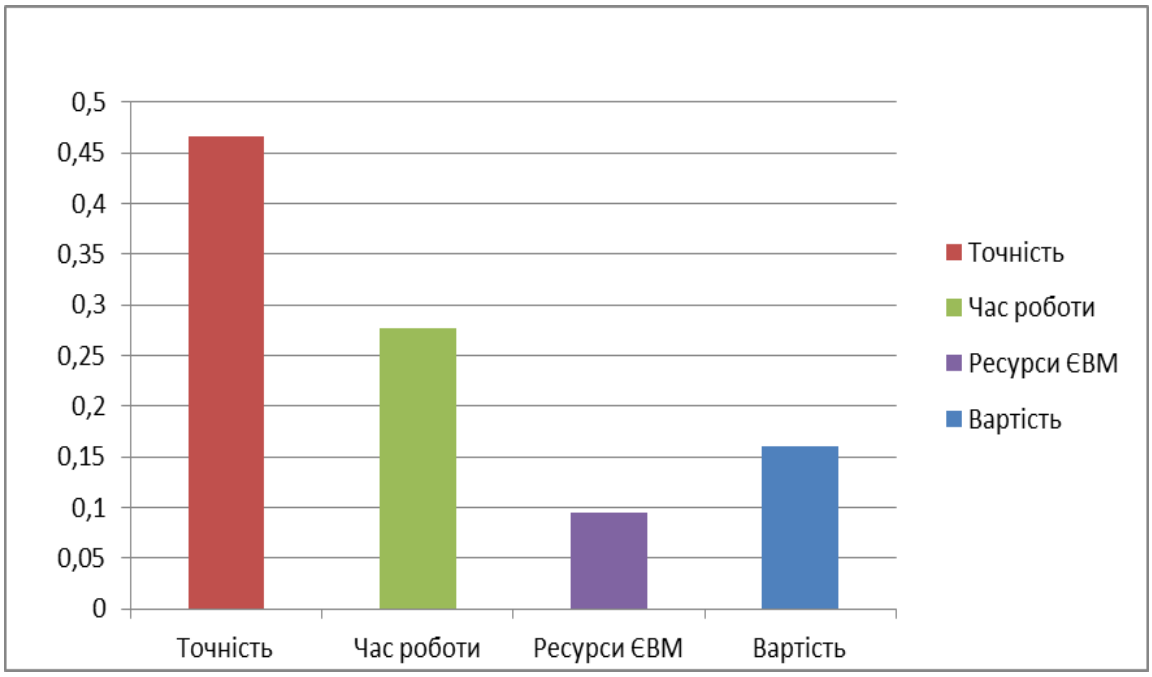

Рисунок 1.2 – Діаграма пріоритетів за критерієм порівняння

Далі переходимо до порівняння альтернатив за кожним з критеріїв. Побудуємо матрицю порівнянь для кожної з альтернатив за кожним критерієм і зведемо отримані дані у таблиці для подальшого аналізу. У таблицях 1.2–1.5 наведена інформація стосовно порівняння кожної альтернативи за кожним з критеріїв.

Таблиця 1.2 – Таблиця порівняння альтернатив за першим критерієм

Точність	Логістична регресія	k найближчих сусідів	Дерева рішень
Логістична Регресія	1	2	4
k найближчих сусідів	$\frac{1}{2}$	1	3
Дерева рішень	$\frac{1}{4}$	$\frac{1}{3}$	1

Таблиця 1.3 – Таблиця порівняння альтернатив за другим критерієм

Час роботи	Логістична регресія	k найближчих сусідів	Дерева рішень
Логістична Регресія	1	$\frac{1}{2}$	$\frac{1}{3}$
k найближчих сусідів	2	1	$\frac{1}{2}$
Дерева рішень	3	2	1

Таблиця 1.4 – Таблиця порівняння альтернатив за третім критерієм

Ресурси ЕОМ	Логістична регресія	k найближчих сусідів	Дерева рішень
Логістична Регресія	1	2	$\frac{1}{2}$
k найближчих сусідів	$\frac{1}{2}$	1	$\frac{1}{2}$
Дерева рішень	2	2	1

Таблиця 1.5 – Таблиця порівняння альтернатив за четвертим критерієм

Вартість	Логістична регресія	k найближчих сусідів	Дерева рішень
Логістична Регресія	1	1	1
k найближчих сусідів	1	1	1
Дерева рішень	1	1	1

Для кожної з альтернатив порахуємо вектори локальних пріоритетів:

$$\vec{p}_1^{\,A} = \begin{pmatrix} 0,558425 \\ 0,319618 \\ 0,121957 \end{pmatrix},\ \vec{p}_2^{\,A} = \begin{pmatrix} 0,163424 \\ 0,296961 \\ 0,539615 \end{pmatrix},\ \vec{p}_3^{\,A} = \begin{pmatrix} 0,310814 \\ 0,195800 \\ 0,493386 \end{pmatrix},\ \vec{p}_4^{\,A} = \begin{pmatrix} 0,333333 \\ 0,333333 \\ 0,333333 \end{pmatrix}.$$

Маємо $RI^A = 0,58$, оскільки матриця попарних порівнянь альтернатив – це матриця третього порядку.

Індекси узгодженості й відношень узгодженості для матриць попарних порівнянь по кожному критерію [1]:

$$CI_{K1}^A = 0,00914735,\ CI_{K2}^A = 0,00460136,\ CI_{K3}^A = 0,0268108,\ CI_{K4}^A = 0,$$
$$CR_{K1}^A = 0,0157713,\ CR_{K2}^A = 0,00793337,\ CR_{K3}^A = 0,0462255,\ CR_{K4}^A = 0.$$

Розрахуємо вектор глобальних пріоритетів:

$$\vec{p} = \begin{pmatrix} 0,558425 & 0,163424 & 0,310814 & 0,333333 \\ 0,319618 & 0,296961 & 0,195800 & 0,333333 \\ 0,121957 & 0,539615 & 0,493386 & 0,333333 \end{pmatrix} \cdot \begin{pmatrix} 0,466849 \\ 0,277590 \\ 0,095295 \\ 0,160267 \end{pmatrix} = \begin{pmatrix} 0,389106 \\ 0,303728 \\ 0,307166 \end{pmatrix}.$$

Розрахуємо індекс узгодженості й відношення узгодженості для всієї ієрархії: $CI = CI^K + \vec{p}^{\,K}$, $\overrightarrow{CI^A} = 0,019$, $RI = RI^K + RI^A = 1,48$, $CR = \dfrac{CI}{RI} = 0,013$.

1.2.3 Модель вирішення проблеми

Узагальнюючи отримані результати ми, як особа, яка приймає рішення, повинні зробити висновки на основі отриманої інформації.

Максимальною компонентою вектора глобальних пріоритетів є перша альтернатива, а, отже, для бінарної класифікації рекомендується обрати метод логістичної регресії, оскільки він дає найбільш точні результати, виходячи з цілей і показників якості, найбільш важливих з точки зору особи, яка приймає рішення [2].

Лістинг програми розрахунків за методом аналізу ієрархій приведено у додатку Б.

1.3 Змістовна та формальна постановка задачі

1.3.1 Змістовна постановка задачі

Відбувається розв’язання задачі класифікації клієнтів банку відносно їх кредитоспроможності та можливості шахрайських дій під час звернення до уповноваженої установи з питання отримання кредитної позики певного розміру чи інших фінансових операцій. Для оцінювання спроможності кожного клієнта відносно повернення фінансових коштів, позичених у банка, та ймовірності противоправних дій будемо використовувати особисту інформацію, яку банк отримує під час попередньої обробки даних під час звернення клієнта.

Проведені розрахунки дозволять спрогнозувати ймовірність повернення коштів, що були позичені кожним окремим клієнтом. На основі отриманих даних керівництво банку зможе приймати рішення щодо можливості взаємодії з цією людиною та, зокрема, видачі їй кредитного займу.

Специфічність даної задачі полягає у тому, що на вході ми маємо велику кількість різноманітних даних, не пов’язаних на перший погляд, які описують клієнта з різних сторін. Тому, окрім безпосередньої задачі класифікації клієнтів, актуальною також стає задача зменшення розмірності простору ознак та виділення серед них найбільш значущих ознак.

Для розв’язання задачі класифікації клієнтів банку у атестаційній роботі

обрано метод логістичної регресії, а зменшення розмірності виконуватиметься за допомогою методу головних компонент. Виділення найбільш значущих критеріїв будемо здійснювати за допомогою методу покрокової логістичної регресії. Для оцінки якості класифікатору на основі логістичної регресії будемо порівнювати отримані ним прогнози, з прогнозами класифікаторів інших типів.

1.3.2 Формальна постановка задачі

Розглянемо множину пар $(\vec{x}_i, y_i)$, $i = \overline{1,m}$, де $\vec{x}_i \in \mathbb{R}^n$ – вектор параметрів i-го об'єкта, $y_i \in \{-1;+1\}$ – його цільова ознака. При розв'язанні задачі прийняття рішення стосовно видачі кредиту такими об'єктами будуть попередні клієнти, що вже отримали кредит, та термін їх кредитування завершився, і про яких відомо, що вони повернули (+1) або не повернули (–1) кредит. У задачі виявлення кредитних махінацій мітка класу показуватиме, чи був цей випадок шахрайством (+1) чи ні (–1) [3].

Алгоритм бінарної класифікації матиме вигляд:

$$a(\vec{x}, \vec{w}) = sign\left(\sum_{j=1}^{n} w_j x_j \right) = sign\langle \vec{x}, \vec{w} \rangle,$$

де w_j – ваги, кожної ознаки.

Задача навчання лінійного класифікатора полягає у встановленні значень ваг w_j, $j = \overline{1,n}$, за заданою вибіркою $(\vec{x}_i, y_i)$, $i = \overline{1,m}$. У задачі класифікації на основі логістичної регресії для цього необхідно мінімізувати емпіричну функцію похибок [4]

$$Q(\vec{w}) = \sum_{i=1}^{m} \ln(1 + e^{-\langle \vec{x}_i, \vec{w} \rangle y_i}) \to \min_{w_j, j=\overline{1,n}}.$$

Побудований лінійний класифікатор дозволить робити висновки стосовно можливості видачі кредиту будь-якому новому клієнту, що описується вектором параметрів $\vec{x}$, та оцінювання ймовірності повернення чи неповернення їм цього кредиту або здійснення кредитних махінацій [5]:

$$\mathrm{P}\{y \mid \vec{x}\} = \frac{1}{1+e^{-y\langle \vec{x}, \vec{w} \rangle}}.$$

Чим вище отримане значення, тим більша ймовірність того, що розглядуваний клієнт буде благонадійним [6]. В залежності від результату, отриманого за допомогою алгоритму, співробітник банку зможе використати цю інформацію за для рішення щодо можливої співпраці з цим клієнтом. Рівень довіри, починаючи з якого, банк дозволяє ухвалювати кредитні позики, можна буде корегувати в залежності від банку, відділення чи міста, у якому знаходиться банк [2].

1.4 Постановка задачі дослідження

Спираючись на системний аналіз системи, а також на аналіз проблеми вибору оптимального методу розв’язання розглядуваної проблеми, сформулюємо перелік задач дослідження даної атестаційної роботи:

– сформулювати задачу бінарної класифікації клієнтів банку щодо можливості отримання кредитної позики та здійснення кредитних махінацій;

– обрати метод розв’язання поставленої задачі класифікації, спираючись на висновки, отримані під час проведення системного аналізу даної проблеми;

– розв’язати задачу бінарної класифікації клієнтів банку за статистичними даними, зібраними з анкет, заповнених клієнтами під час звернення до банку;

– створити програмний продукт, який дозволить розв’язати задачу бінарної класифікації клієнтів банку.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

2.1 Методи бінарної класифікації

Сьогодні для розв'язання задач, які потребують обробки і аналізу великого обсягу різнорідних даних, вже створено велику кількість методів й алгоритмів класифікації і регресії, а також велику кількість їх модифікацій. В даному розділі розглянемо найвідоміші з цих методів, зокрема:

– лінійну регресію;

– логістичну регресію;

– метод k найближчих сусідів;

– простий класифікатор Байєса;

– метод дерева рішень.

2.1.1 Лінійна регресія

Лінійна регресія використовується для моделювання лінійної залежності між неперервною вхідною змінною і набором вхідних змінних. При певних умовах рівняння лінійної регресії є незамінним і дуже якісним інструментом аналізу і прогнозування [7]. Модель лінійної регресії є найбільш розповсюдженим і простим рівнянням залежності між вхідними й вихідними даними. Крім того, побудоване рівняння лінійної регресії можна буде використати як початкову точку для аналізу даних [8].

Теоретична лінійна регресійна модель має вигляд:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon,$$

де y – вихідна змінна моделі;

$x_1, x_2, ..., x_n$ – вхідні змінні;

β_i, $i=\overline{1,n}$, – коефіцієнти лінійної регресії;

β_0 – вільний член;

ε – оцінка теоретичного випадкового відхилення.

Для визначення значень теоретичних коефіцієнтів для лінійної регресії необхідно знати і використовувати розподіл генеральної сукупності, що практично неможливо [7].

За вибіркою обмеженого об'єму можливо побудувати емпіричне рівняння регресії:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n,$$

де y – вихідна змінна моделі;

$x_1, x_2, ..., x_n$ – вхідні змінні;

b_i, $i=\overline{1,n}$, – коефіцієнти лінійної регресії;

b_0 – вільний член.

Коефіцієнти b_i обираються таким чином, щоб на заданий вектор $\vec{x} = (x_1, x_2, ..., x_n)$ регресійна модель формувала бажане вихідне значення y. Емпіричні коефіцієнти регресії b_0 і b_i, $i=\overline{1,n}$, є оцінками теоретичних коефіцієнтів β_0 і β_i, $i=\overline{1,n}$, а саме рівняння відображає тенденцію у поведінці змінних. Індивідуальні значення змінних в силу різних причин можуть відхилятися від модельних значень [7].

Одним з найбільш привабливих властивостей лінійної регресії є те, що її коефіцієнти можуть бути отримані за допомогою методу найменших квадратів [8].

Але модель лінійної регресії не є універсальною. Зазвичай, якщо для розв'язання задачі використовують модель лінійної регресії, на значення залежної змінної не накладають жодних обмежень. Але на практиці такі обмеження зустрічаються досить часто і можуть досить суттєво впливати на кінцевий результат. Наприклад, вихідна змінна може бути бінарною чи категоріальною.

Для таких випадків зазвичай використовують логістичну регресію, яка краще працює з такими типами даних [8].

2.1.2 Логістична регресія

Під час аналізу даних досить часто зустрічаються задачі, в яких вхідна змінна є категоріальною. У цьому випадку використання лінійної регресії ускладнюється. Тому при розв'язанні задач пошуку зв'язку між набором вхідних змінних і категоріальною вихідною зміною набула розповсюдження логістична регресія. Вона також є методом бінарної класифікації та дозволяє робити оцінку ймовірності реалізації (чи не реалізації) події в залежності від значень деяких залежних змінних. Лінією логістичної регресії, на відміну від лінійної регресії, не є пряма.

Умовне середнє для логістичної регресії має вигляд:

$$p(\vec{x}) = \frac{e^{w_0 + w_1 x_1 + \ldots + w_n x_n}}{1 + e^{w_0 + w_1 x_1 + \ldots + w_n x_n}},$$

де e – основа натурального логарифму;

p – ймовірність того, що станеться розглядувана подія;

$\beta_0, \beta_i, i = \overline{1,n}$, – коефіцієнти логістичної регресії;

$\vec{x}$ – вектор незалежних змінних.

Вищенаведену функцію називають логістичною. Значення $p(\vec{x})$ змінюється у діапазоні від 0 до 1. Якщо зробити припущення, що значення вихідної змінної y, рівне 1, розглядається як успіх, а значення 0 – як невдача, то $p(\vec{x})$ можна інтерпретувати як ймовірність успіху, а $1 - p(\vec{x})$ – невдачі.

Для оцінки коефіцієнтів логістичної регресії використовують не метод найменших квадратів, а метод максимальної правдоподібності.

Логарифмічна функція правдоподібності, що підлягає максимізації за змінними β_0, β_i, $i = \overline{1, n}$, має вигляд:

$$L(\vec{\beta} \mid \vec{x}) = \sum_{i=1}^{m} \{ y_i \ln(p(\vec{x}_i)) + (1 - y_i) \ln(1 - p(\vec{x}_i)) \}.$$

2.1.3 Простий класифікатор Байєса

Байєсовський підхід об'єднує групу алгоритмів класифікації, заснованих на принципі умовної ймовірності: для об'єкту за допомогою формули Байєса визначається апостеріорна ймовірність належності кожному класу і обирається той клас, для якого вона максимальна. Байєсовська класифікація спочатку використовувалась для нормалізування знання експертів в експертних системах, зараз вона також використовується у якості одного з методів Data Mining [9].

Особливе місце у даній області відведене простому класифікатору Байєса (Naive Bayes), в основі якого лежить припущення про незалежність ознак, що описують системи, які підлягають класифікації. Це припущення значно спрощує задачу, оскільки замість складної процедури оцінки багатовимірної щільності ймовірності потрібна оцінка декількох одномірних щільностей. Але, як показує практика, припущення щодо незалежності ознак рідко виконується, що і дало ім'я методу.

До основних переваг класифікатору можна віднести легкість програмної реалізації й низькі затрати на розрахунки. У тих випадках, коли ознаки дійсно незалежні, метод оптимальний. Головний його недолік – відносно низька якість класифікації у більшості реальних задач. Тому найчастіше його використовують як примітивний еталон для порівняння різних моделей, чи як блок для побудови більш складних алгоритмів [1].

Хоча припущення стосовно статистичної незалежності ознак досить рідко виконується на практиці, в машинному навчанні існує достатньо різних мето-

дів, які дозволяють відбирати найменш корельовані ознаки. Використання таких методів дозволяє суттєво підвищити ефективність байєсовських класифікаторів для розв'язання задач класифікації.

2.1.4 Метод k найближчих сусідів

Метод k найближчих сусідів (k-nearest neighbor algorithm) є примітивним метричним класифікатором для визначення класу об'єкту. Алгоритм був запропонований у роботі Фікса і Ходжеса [6] у 1951 році. Під метричним класифікатором мають на увазі алгоритм, побудований на оцінці схожості між об'єктами вибірки. У рамках даного методу повинна бути введена метрика відстані між об'єктами вибірки. Слід зазначити, що підбір метрики є одним з найважливіших аспектів застосування цього алгоритму на практиці [8].

Метод найближчих сусідів виділяється серед метричних класифікаторів тим, що процес класифікації об'єкту полягає у виборі класу, до якого належать найближчі до об'єкта спостереження. Більш точно, обирається та категорія, якій належить більша частина сусідів. На практиці кількість досліджуваних сусідніх спостережень встановлюють непарним для недопущення ситуації, коли однакове число сусідів належить різним класам [10].

Серед переваг цього методу можна виділити стійкість до впливання викидів у вибірці, що зумовлено малою ймовірністю для такого спостереження опинитися серед k найближчих сусідів, нескладна програмна реалізація і широкий простір для модифікації алгоритму за допомогою підбору найбільш підходящих метрик для досліджуваної задачі [9]. В свою чергу, головним недоліком є необхідність використовувати усі спостереження для класифікації одного об'єкту, що значно ускладнює практичне використання алгоритму [11].

2.1.5 Метод дерева рішень

Дерево рішень – це модель, яка являє собою сукупність правил для прийняття рішень. Графічно її можна уявити у вигляді деревовидної структури, де момент прийняття рішення відповідає так званим вузлам (decision nodes). У вузлах відбувається гілкування (branching), тобто розділення на так звані гілки (branches) в залежності від зробленого вибору. Кінцеві вузли називаються листками (leaf nodes), де кожен лист є кінцевим результатом послідовного прийняття рішень [12].

Дані, які підлягають класифікації, знаходяться у корні дерева. В залежності від рішення, яке було прийнято у вузлі, процес в решті решт зупиниться у одному з листків, де змінній відгуку присвоюється певне значення [12].

Метод дерева рішень реалізує принцип так званого рекурсивного ділення. У вузлах, починаючи з корінного, обирається ознака, значення якої використовується для розбиття усіх даних на два класи. Процес продовжується до тих пір, доки не виконається критерій зупинки [12]. Приклад реалізації алгоритму дерева рішень можна побачити на рисунку 2.1.

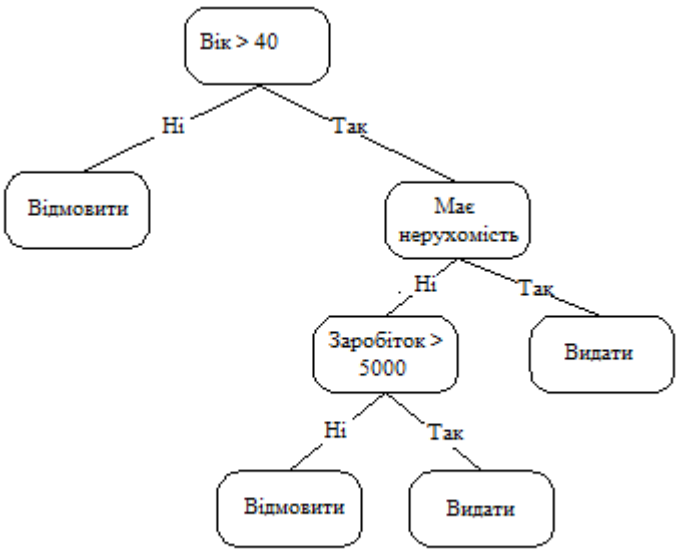

Рисунок 2.1 – Ілюстрація дерева рішень

2.2 ROC-аналіз

ROC-аналіз представляє собою методику графічного оцінювання ефективності моделей за допомогою двох показників – чутливості (*Se*) і специфічності (*Sp*).

Для задач бінарної класифікації у випадках, коли модель прогнозує ймовірність того, що спостереження відноситься до одного з двох класів, дуже вагомим є вибір точки відсікання, тобто значення ймовірності, яке розділяє два класи. За допомогою такої точки можна показати, після якого значення ймовірності на виході моделі один клас замінюється іншим. Корегуючи цю точку, ми можемо отримати можливість керування ймовірністю правильного впізнавання позитивних і негативних випадків. При зменшенні порогу відсікання збільшується ймовірність хибного розпізнавання позитивних спостережень (хибно позитивних результатів), а при збільшенні зростає ймовірність неправильного розпізнавання негативних спостережень (хибно негативних результатів) [13].

Метою ROC-аналізу є підбір такого значення точки відсікання, яке дозволить моделі розпізнавати позитивні чи негативні результати з найбільшою точністю і мінімізувати кількість хибно позитивних чи хибно негативних помилок, відповідно [14].

ROC-аналіз широко використовується у багатьох галузях, зокрема: медицині, біології, маркетингу, банківській справі та інших областях, де застосовується бінарна ймовірнісна класифікація.

В основі ROC-аналізу лежать ROC-криві. Графік ROC-кривої зображено на рисунку 2.2.

Для побудови ROC-кривої поріг відсікання повинен змінюватися у діапазоні від 0 до 1 с заданим кроком, наприклад 0,01, який буде регулювати кількість точок на графіку. На кожному значенні порогу заново розраховується значення специфічності (*Sp*) і чутливості (*Se*), тобто змінюється кількість розпізнаних похибок першого і другого роду. Чутливість (*Se*) відкладається на вісі ординат, а на вісі абсцис відкладається значення 100 – *Sp* [15].

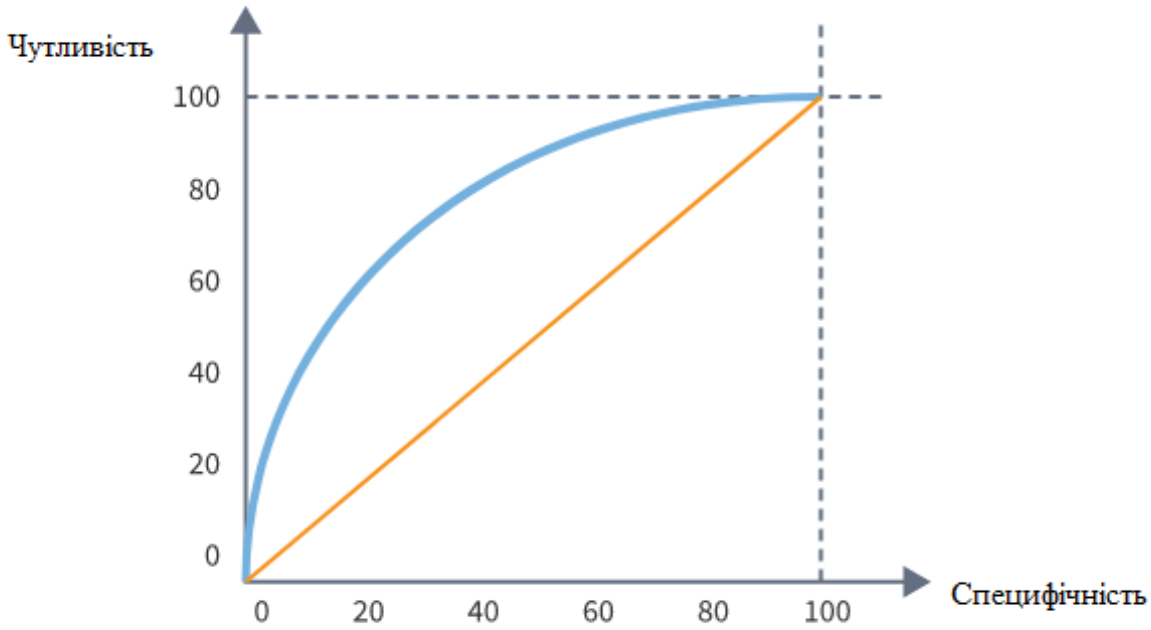

Рисунок 2.2 – Графік ROC-кривої

Ще одна корисна властивість ROC-кривої полягає у тому, що вона може дозволити добре оцінити якість побудованої моделі для класифікації. Визначити якість моделі можна за формою кривої: чим ближче вона до ідеального класифікатора, тим якісніша модель. Якщо ж крива близька до діагоналі, то модель не має сенсу.

Нажаль, порівняти ROC-криві і виявити більш ефективну модель не завжди є можливим. Тому їх можна порівнювати за допомогою площі під кривою (AUC). Площа під кривою характеризує якість прогнозуючої сили моделі, при цьому $AUC = 1$ відповідає ідеальному класифікатору, який є недосяжним на практиці, а $AUC = 0{,}5$ відповідає класифікатору дуже низької якості [16].

Площа під кривою може бути розрахована, наприклад, методом трапецій:

$$AUC = \sum_{i=1}^{n} \left[\left(\frac{x_{i+1} + x_i}{2} \right) \left(y_{i+1} - y_i \right) \right],$$

де n – кількість точок;

x – координата точки на вісі абсцис;

y – координата точки на вісі ординат.

Наскільки добру прогнозуючу силу має модель, можна судити з таблиці 2.1. В ній представлена приблизна експертна шкала оцінки якості моделі в залежності від площі під ROC-кривою [17].

Таблиця 2.1 – Якість моделі в залежності від площі під кривою

Інтервал значень AUC	Якість моделі
0,9-1	Бездоганна
0,8-0,9	Дуже добра
0,7-0,8	Добра
0,6-0,7	Середня
0,5-0,6	Незадовільна

2.3 Метод головних компонент

Метод головних компонент є одним з найпопулярніших методів зменшення розмірності. Головна ідея цього методу полягає в послідовному виявленні напрямків, у яких дані мають найбільшу дисперсію.

Головні компоненти є множиною досліджуваних ознак $y^{(1)}, y^{(2)}, \ldots, y^{(p)}$, кожна з яких представляє собою деяку лінійну комбінацію вихідних ознак $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, що були отримані в результаті вимірів та опису об'єктів. Отримані в результаті такого перетворення нові ознаки $y^{(1)}, y^{(2)}, \ldots, y^{(p)}$ мають ряд зручних властивостей. Зазвичай вони впорядковані за ступенем розсіювання в досліджуваній сукупності об'єктів; перша ознака має найбільшу ступінь розсіювання, тобто найбільшу дисперсією [18].

Розглянемо деякі важливі властивості головних компонент:

а) сума квадратів відстаней від вихідних точок-спостережень X_1, X_2, …,

X_n до простору, натягнутого на перші p' головних компонент, найменша відносно всіх інших підпросторів вимірності p', отриманих за допомогою довільного лінійного перетворення вихідних координат;

б) серед всіх підпросторів заданої вимірності p' ($p' < p$), отриманих для досліджуваного факторного простору за допомогою довільного лінійного перетворення вихідних координат $x^{(1)}$, $x^{(2)}$, …, $x^{(p)}$, в підпросторі, натягнутому на перші p' головних компонент, найменш спотворюється сума квадратів відстаней між усіма можливими парами розглядуваних точок-спостережень;

в) серед всіх підпросторів заданої вимірності p' ($p' < p$), отриманих для досліджуваного факторного простору за допомогою довільного лінійного перетворення вихідних координат $x^{(1)}$, $x^{(2)}$, …, $x^{(p)}$, в підпросторі, натягнутому на перші p' головних компонент, найменш спотворюється відстань від розглядуваних точок-спостережень до їх спільного «центра тяжіння», а також кути між прямими, що з'єднують усі можливі пари точок-спостережень з їх спільним «центром тяжіння» [8].

2.4 Алгоритм розв'язання задачі бінарної класифікації клієнтів банку

Під час виконання атестаційної роботи необхідно розв'язати задачу класифікації. Процедура розв'язання полягає у визначенні такої функції $a: X \to Y$, яка кожному об'єкту $\vec{x} \in X$ ставить у відповідність значення $y \in Y$, що є міткою класу цього об'єкта. Значення змінної y є відомим на об'єктах тренувальної вибірки: $(\vec{x}_i, y_i)$, $i = \overline{1,m}$, де $\vec{x}_i \in \mathbb{R}^n$ – вектор параметрів i-го об'єкта, $y_i \in \{-1;+1\}$ – його цільова ознака.

При розв'язанні задачі прийняття рішення стосовно видачі кредиту такими об'єктами будуть попередні клієнти, що вже отримали кредит, та термін їх кредитування завершився, і про яких відомо, що вони повернули (+1) або не пове-

рнули (–1) кредит. У задачі виявлення кредитних махінацій мітка класу показуватиме, чи був цей випадок шахрайством (+1) чи ні (–1) [3].

Якщо у якості класифікатора використовувати логістичну регресію, то алгоритм бінарної класифікації матиме вигляд:

$$a(\vec{x},\vec{w}) = sign\left(\sum_{j=1}^{n} w_j x_j\right) = sign\langle \vec{x},\vec{w}\rangle,$$

де w_j – ваги, кожної ознаки, які необхідно визначити, мінімізуючи емпіричну функцію похибок:

$$Q(\vec{w}) = \sum_{i=1}^{m} \ln(1+e^{-\langle \vec{x}_i,\vec{w}\rangle y_i}) \to \min_{w_j,\, j=1,n}.$$

Виходячи з результатів системного та аналітичного аналізу за темою атестаційної роботи для розв’язання поставлених задач потрібно здійснити наступне:

– підготувати статистичні дані, необхідні для навчання класифікатора, що відповідає формальній постановці задачі, зокрема, провести перевірку рядків на наявність пустих значень, привести усі дані до бінарного вигляду та виконати нормування;

– навчити класифікатор, у основі якого лежить логістична регресія;

– виділити найбільш значущі критерії та використати їх для побудови класифікатора;

– навчити класифікатори інших типів для порівняння якості класифікації;

– спробувати покращити якість класифікатора логістичної регресії за рахунок підсилення його іншими класифікаторами;

– порівняти отримані результати і зробити висновки відносно якості системного аналізу проблеми вибору метода бінарної класифікації та якості класифікації.

За допомогою побудованого класифікатора можна прогнозувати ймовірності належності об’єктів кожному з класів. Зокрема, у разі використання логістичної регресії ця ймовірність для об’єкта, що описується вектором ознак $\vec{x} \in X$, обчислюється за формулою:

$$\mathrm{P}\{y \mid \vec{x}\} = \frac{1}{1 + e^{-y\langle \vec{x}, \vec{w} \rangle}}.$$

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Wolfram Mathematica 11.0 як система символьної математики

Для отримання можливості виконання складних розрахунків і маніпуляції великими обсягами даних були створені системи комп'ютерної математики, такі як Mathcad, MATLAB, Mathematica та інші. Вони мають можливість за короткий час виконувати великі обсяги роботи, пов'язаної з виконанням арифметичних і логічних операцій над числами або масивами чисел. Результатом виконання такої програми завжди є число чи масив чисел, через його блокову структуру. Сучасні технології дозволяють таким чином виконувати безліч обчислень за досить короткий відносно об'єму даних проміжок часу.

Mathematica 11.0 є однією з простих і потужних систем символьної математики. Пакет має великі графічні можливості. Завдяки постійному розвитку протягом трьох десятиліть система Mathematica не має собі рівних і з кожним оновленням отримує все більше можливостей для роботи у великому діапазоні задач і дає унікальні можливості щодо підтримки та організації робочого процесу для великої кількості технічних розрахунків.

Система розроблена таким чином, що здатна забезпечити зручну роботу з декількома документами. Це надає неймовірні можливості під час розв'язання багатьох символьних задач, а завдяки спеціальній структурі вона має можливість об'єднувати розрахунки одним ядром, у якому зберігаються дані під час роботи з кожним документом. Якщо поглянути на Mathematica 11.0 з точки зору програмування, то ця система відноситься до інтерпретуючих, тобто аналіз (інтерпретація) відбувається для кожного виразу і одразу ж відбувається його виконання.

Ще однією перевагою пакета Mathematica 11.0 є те, що вона має дуже високий рівень інтеграцій з різними мовами програмування за допомогою системних бібліотек, таких як «Wolfram.dll». Це дозволяє інтегрувати її у інші мови для розв'язання задач за допомогою алгоритмів Mathematica чи, навпаки, пере-

класти задачу, відправивши її на виконання Mathematica і лише отримати і опрацювати результат.

3.2 Python як середа розробки програмного продукту

Python - це інтерпретуюча об’єктно орієнтована мова програмування. Ця мова досить проста і містить не дуже велику кількість ключових слів, але при цьому є досить гнучкою і швидкою. За рівнем ця мова займає більш високу сходинку, ніж Pascal, C++, C, що було досягнене за допомогою використання структур даних високого рівня (списки, словники, кортежі).

Однією з найголовніших переваг цієї мови програмування є розширюваність і, як зазначив створювач даної мови, вона розроблялась саме як розширювана. Це означає, що можливість вдосконалення мови стала публічною, і кожен зацікавлений програміст може внести свій вклад у розвиток. Інтерпретатор було написано на мові C, і увесь код було розміщено у вільному доступі для надання можливості до будь-яких маніпуляцій. Також ця мова має високий рівень інтеграції, і, у разі необхідності, файли можна використовувати при розробці програм, які об’єднують декілька різних мов програмування. Це також надає можливість перекласти всі необхідні обчислювальні дії на спеціальні бібліотеки, написані мовою Python. Це дозволяє зменшити час розробки програмного забезпечення чи інтегрування, бо в Python є багато вбудованих засобів для обробки даних різних типів, які можуть стати у нагоді для будь-якого програмного забезпечення.

Завдяки цьому Python, як мова програмування, набула популярності і можна навіть сказати, що популяризувала аналіз даних, значно зменшивши поріг входу до нього. Завдяки такій гнучкості і відкритості зараз у Python є дуже багато вбудованих функцій, які можуть значно полегшити аналіз даних чи будь-які обчислювальні дії, але при цьому зберігають можливість для реалізації специфічних методів власноруч.

3.3 Опис програми

Програма, яка надає змогу розв’язати задачу бінарної класифікації клієнтів банку, що бажають отримати кредитну позику чи планують здійснити протиправні дії у фінансових операціях задля власної вигоди, зокрема, спрогнозувати ймовірність повернення кредитної позики клієнтом банку чи ймовірність того, що він не шахрай, була виконана за допомогою можливостей системи Python 3.

Як тільки користувач відкриє файл з програмою, то перед ним з’явиться програмний код і можливість змінити вхідні данні, які з легкістю імпортуються з Excel чи csv формату.

Після завантаження вхідних даних можна переходити до реалізації алгоритму, який підготовлює дані для використання програмою. Зменшуємо розмірність підготовлених даних. Також дані потрібно розділити на дві частини, для використання їх відокремлено один від одного, аби уникнути ситуацій з перенавчанням. Перший набір даних будемо використовувати як тренувальний, а другий для прогнозування і оцінки якості навчання класифікатора.

Далі розраховуються основні значення, що використовуються для побудови класифікаторів та будуються графіки, за допомогою яких можна наочно побачити точність отриманого розв’язку.

Для контролю якості будемо використовувати ROC-криву, яка добре демонструє точність роботи алгоритму.

Дана програма дозволяє провести аналіз клієнтів банку й розрахувати ймовірність шахрайських дій або повернення кредиту на основі даних стосовно цього клієнта. Це може полегшити роботу співробітників банку щодо прийняття рішень з можливості співпраці з окремими клієнтами, знизивши кількість не ризикових позик чи кредитних махінацій.

Лістинг програми приведено у додатку В.

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ

4.1 Задача кредитного скорінгу

Розглянемо задачу визначення кредитної платоспроможності (кредитного скорінгу). На вході ми маємо дані стосовно клієнтів банку, які звернулись для отримання кредитної позики. Об’єктами є клієнти, а ознаками є їх стать, кількість дітей, рівень заробітної платні, освіта, кредитна історія і т.д. (всього 15 ознак).

Виділеною ознакою (відповідь) виступає інформація про те, чи була повернена кредитна позика клієнтом у банк чи ні. Потрібно навчитися прогнозувати за цими даними, чи поверне кредит новий клієнт, який звертається до банківської установи.

Таким чином, постає питання щодо вирішення задачі класифікації: потрібно визначити, до якого з класів: позитивного (кредитну позику буде повернено) чи негативного (кредитну позику не буде повернено) – належить клієнт.

У якості модельних даних візьмемо Credit Approval Data Set з колекції UCI Machine Learning Repository.

Стисло охарактеризуємо ці дані. Маємо 690 записів стосовно клієнтів, які характеризуються за 15-тьма критеріями. Оскільки інформація є приватною, то дані було знеособленно для збереження конфіденційності. Останній стовпчик містить символи «+» і «–», які є мітками класу, і показують, чи повернув даний клієнт кредитну позику видану банківською установою.

Нажаль, неможливо одразу використати ці дані, оскільки вони можуть містити пусті значення, які можуть не коректно вплинути на роботу алгоритму.

З’ясуємо, скільки стовпців з пропущеними чи не коректними значеннями для кожного рядка ми маємо. Можливими варіантами рішення є видалення рядків з незаповненими даними чи заповнення їх середніми значеннями за стовпчиками. Так як видалення даних може призвести до нестачі матеріалу для навчання, було прийнято рішення щодо заповнення пропущених значень середні-

ми за стовпцями.

Наступним кроком буде виділення підмножини з даних, у якій значення є символьними, а її, у свою, чергу розділимо ще на 2 групи: ті, які у стовпчику набувають лише 2 значення, і ті, котрі набувають більше ніж 2 значення. Для першої підмножини замінимо перший символ на одиницю, другий на нуль. Для другої ж доведеться застосувати метод векторизації для того, щоб привести дані до потрібного формату.

Останнім кроком обробки даних буде нормалізація, для того щоб подати їх у числовому вигляді, і після цього дані будуть готові для використання алгоритмом.

Для набуття максимальної точності буде використано декілька методів навчання нейронної мережі, а результати будуть порівняні між собою.

Для виключення ситуації, коли класифікатор буде прогнозувати на даних, на яких він навчався, розділимо дані на 2 частини: тренувальну і тестову. Перша слугує для навчання, на ній ми будемо корегувати вагові коефіцієнти нашого класифікатора. Друга вибірка слугує невідомими даними для навченого класифікатора і використовується для тестування якості його роботи.

Після перетворення даних на числові кількість ознак зросла до 42. Навчивши нашу модель за усіма цими ознаками, ми отримали точність прогнозування 87% і 90% на тренувальній і тестовій виборці відповідно. Щоб зрозуміти, наскільки така модель підходить для даної задачі, побудуємо спеціальну ROC-криву (рисунок 4.1) й, обчисливши площу під нею, отримаємо якість даного класифікатора.

Площа дорівнює 0,937, що свідчить про те, що наш класифікатор був досить якісно навчений і добре підходить для роботи з задачею такого типу. Але для графічної інтерпретації результатів класифікації 42 критерії – це дуже багато, тому було б добре зменшити їх кількість. Для зменшення розмірності можна скористатися методом головних компонент, зменшивши кількість компонент до двох, або спробувати використати логістичну регресію з вибором найбільш впливових критеріїв.

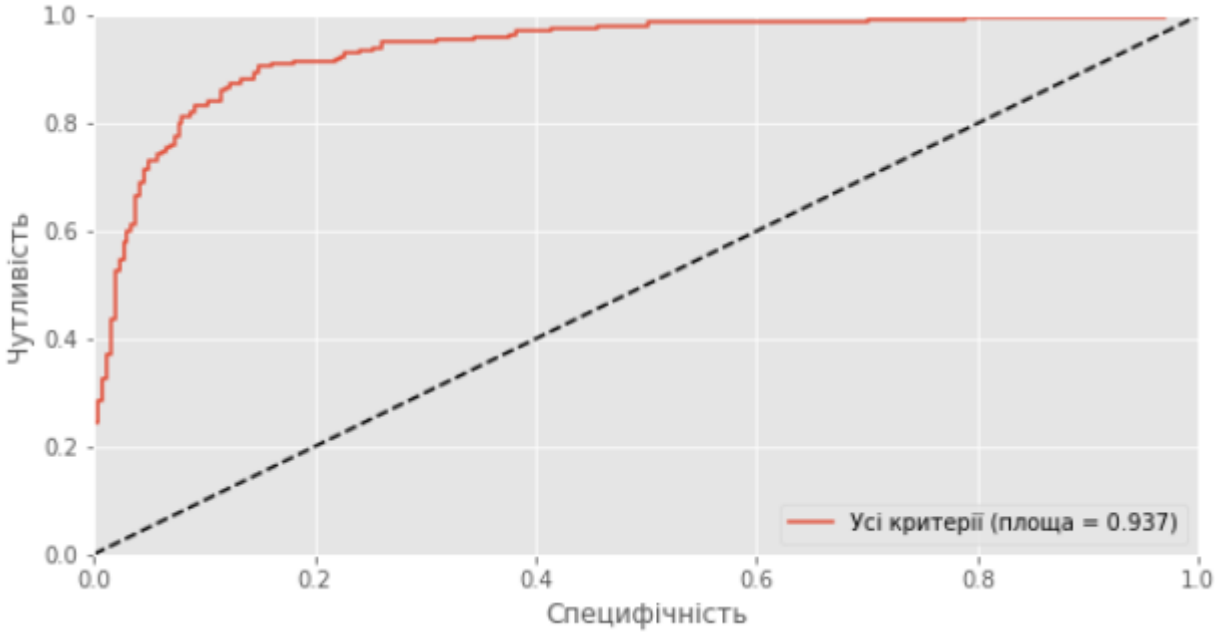

Рисунок 4.1 – ROC-крива

Спочатку спробуємо використати метод головних компонент для зменшення розмірності до двох критеріїв. Навчання класифікатора на наборі даних, отриманому методом головних компонент, дало точність 77% і 81% для тренувальної і тестової вибірки відповідно. Побудуємо ROC-криві (рисунок 4.2) і порівняємо обидва методи.

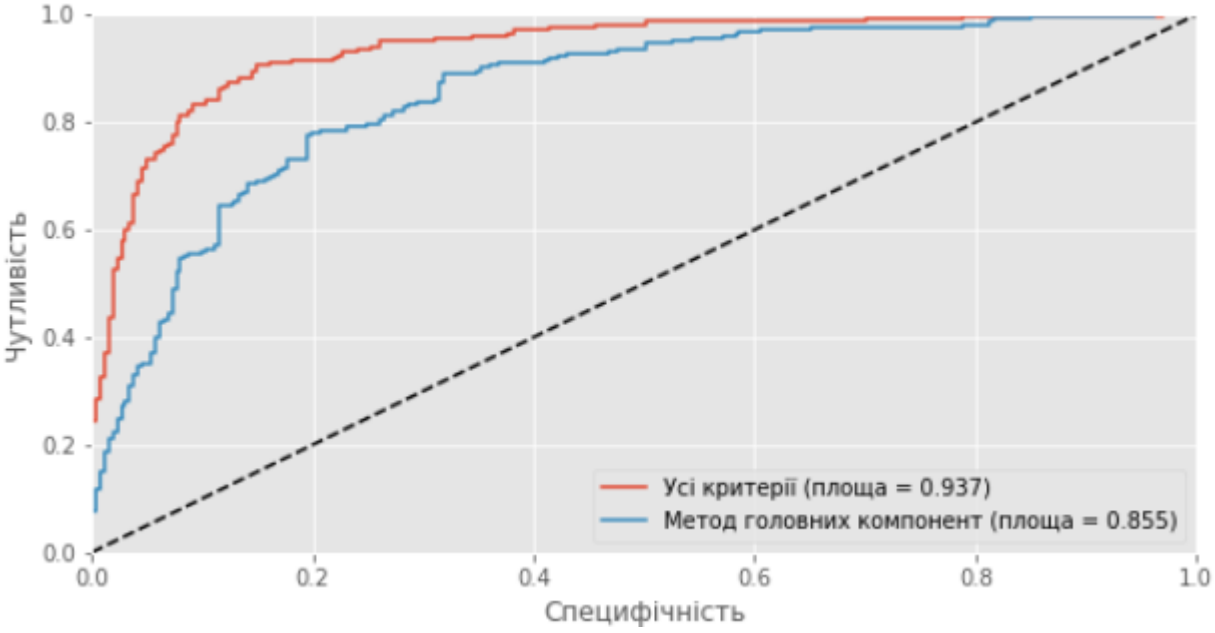

Рисунок 4.2 – ROC-крива порівняння двох методів

Наведемо графік, який ілюструє, як саме класифікатор проводить розподіл клієнтів до різних класів, відносно отримання кредитних позик (рисунок 4.3). На цьому графіку у лівій області зображені клієнти банку, які повернули кредит, а у правій – ті, хто не повернув.

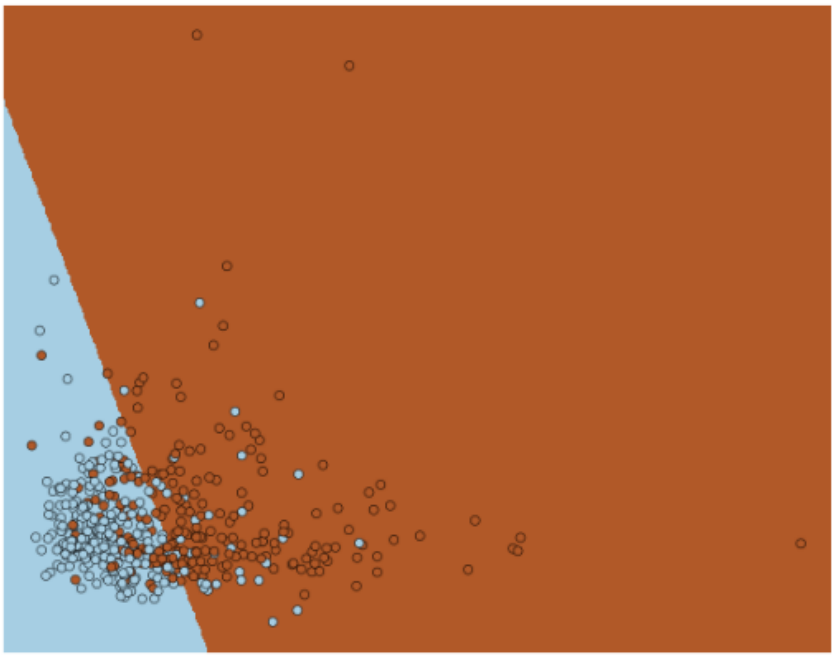

Рисунок 4.3 – Класифікація клієнтів

Тепер спробуємо зменшити кількість критеріїв за допомогою покрокової логістичної регресії, яка дозволяє виділити найбільш вагомі критерії. На перших двох кроках цього алгоритму критеріями, що вносили найбільший внесок, виявились критерії 14-й і 9-й. Навчений за цими критеріями логістичний класифікатор дозволив отримати точність у 85% для тренувальної вибірки і 88% для тестової. Порівняння цього методу навчання класифікатора з іншими наведено на рисунку 4.4.

Слід також спробувати довчити модель за допомогою критеріїв, які дають найбільший приріст до якості класифікатору. У результаті цього процесу алгоритмом покрокової логістичної регресії було обрано п’ять критеріїв (з номерами 14, 9, 6, 2, 13), які суттєво збільшували точність нашого класифікатору, доки це не стало неможливим. Модель, навчена за допомогою цих критеріїв, при

класифікації отримала точність у 86% для тренувального набору і 88% – для тестового. ROC-криві порівняння усіх методів навчання наведені на рисунку 4.5.

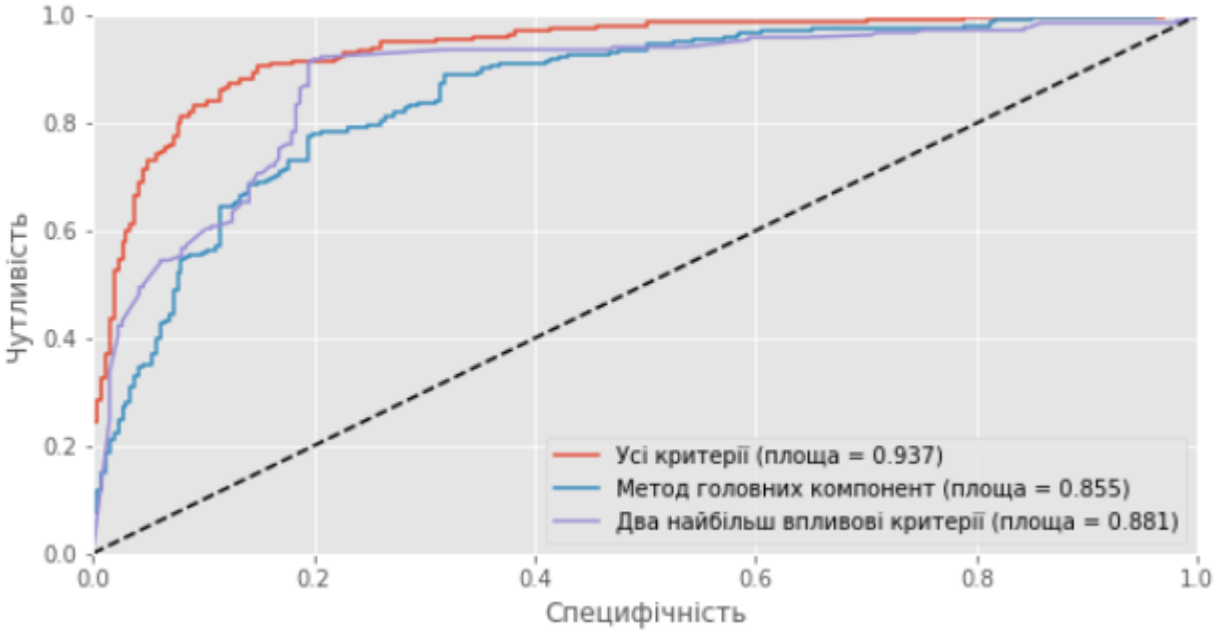

Рисунок 4.4 – Порівняння ROC-кривих для трьох методів

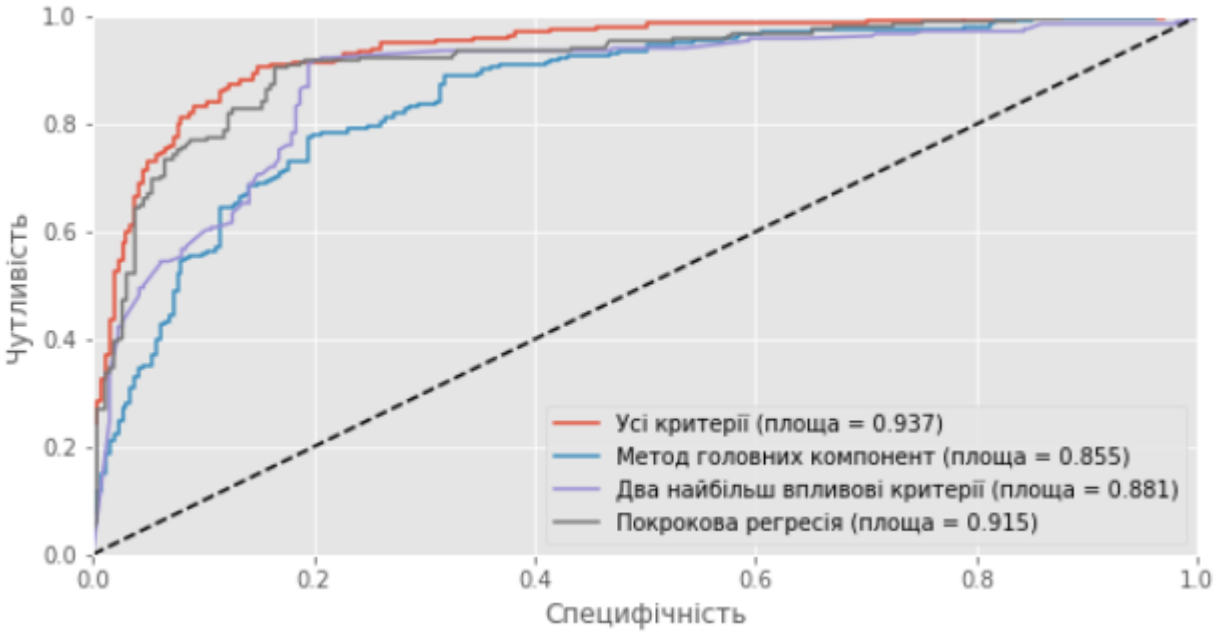

Рисунок 4.5 – ROC-криві порівняння чотирьох методів навчання

Як ми бачимо з графіків, логістична регресія досить не погано впоралась з класифікацією клієнтів банку за усією чи частковою інформацію. Для кращо-

го розуміння слід також порівняти результати не лише з методами, у основі яких лежить логістична регресія. Для цього порівняємо логістичну регресію з найліпшими показниками з методами SVM та деревами рішень.

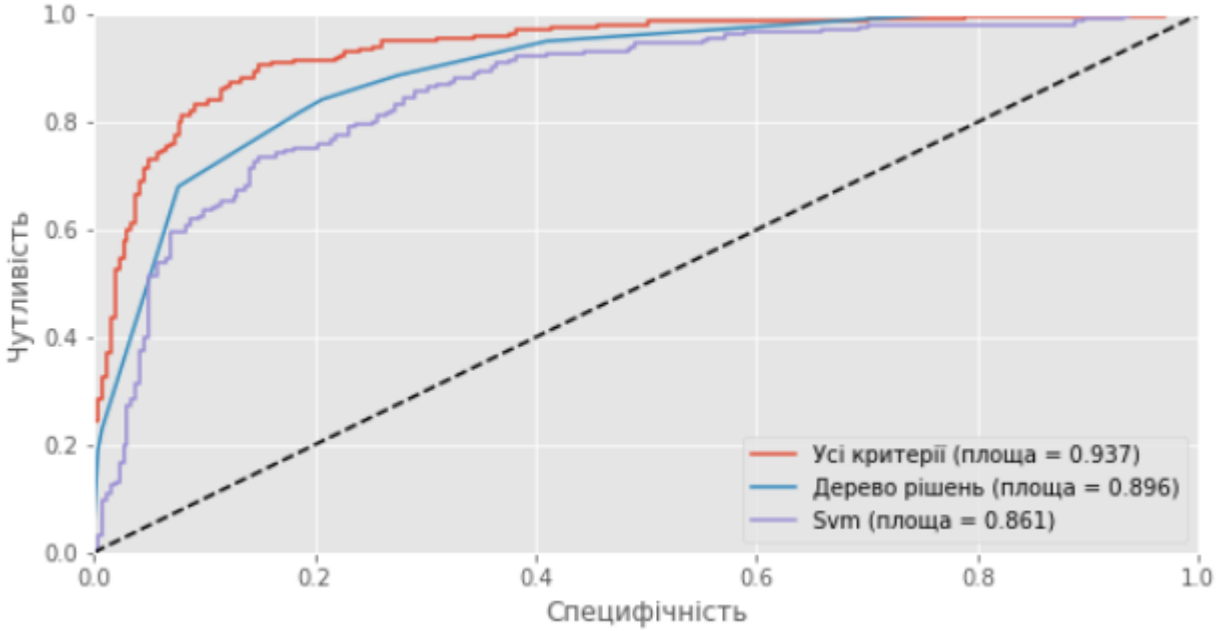

Рисунок 4.6 – ROC-криві порівняння різних методів

Як ми бачимо з графіків, вони також мають гарну точність, але все ще поступаються класифікатору на основі логістичної регресії, навченому за допомогою усіх ознак об’єктів. Це може свідчити про те, що ці класифікатори можуть не дуже добре працювати конкретно з даним типом задач.

Спробуємо створити комбінацію (ансамбль) методів і навчити його за допомогою нашої вибірки даних. Візьмемо декілька комбінацій. Перший ансамбль складається з двох логістичних регресії з різними конфігураціями ядра і різними обмеженнями ітерацій. У другому ансамблі скомбінуємо логістичну регресію і метод дерева рішень.

Порівнюючи результати, отримані у ході досліджень, бачимо, що найкращим класифікатором для даної задачі виявився не один класифікатор, а їх комбінація, яка змогла покращити слабкі сторони кожного з методів і тим самим підвищити якість класифікації.

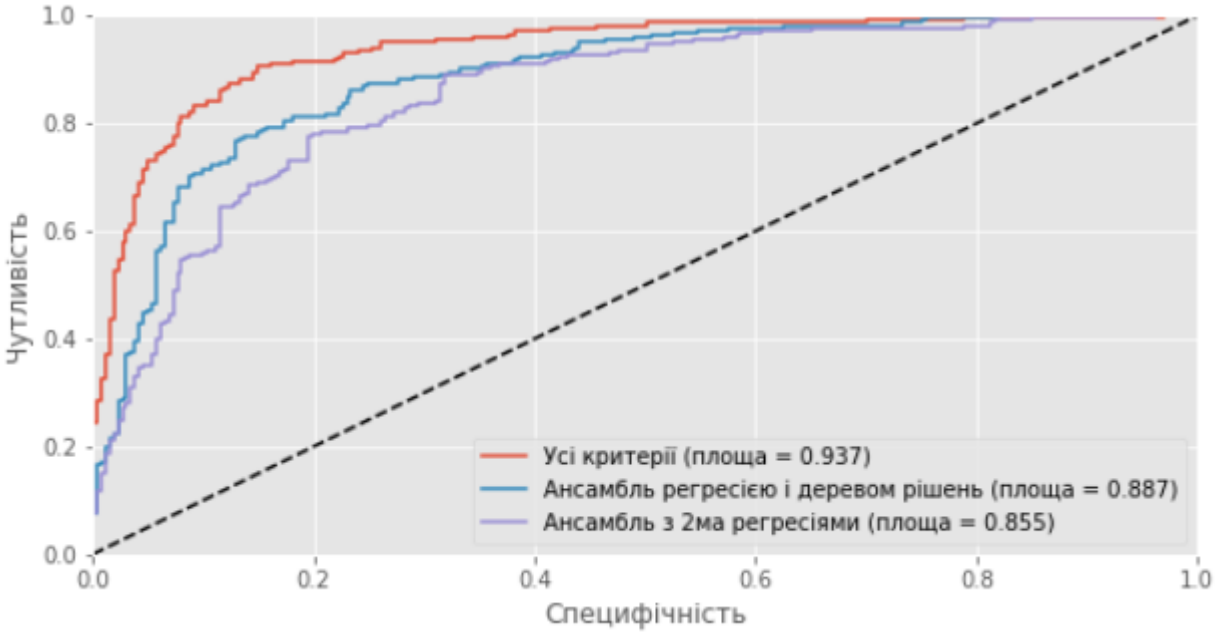

Рисунок 4.7 – ROC-криві порівняння логістичної регресії і ансамблів

Для більшої наочності ROC-криві кожного з методів скомбіновані на наступному графіку. Результати класифікацій різними алгоритмами зведені до таблиці 4.1.

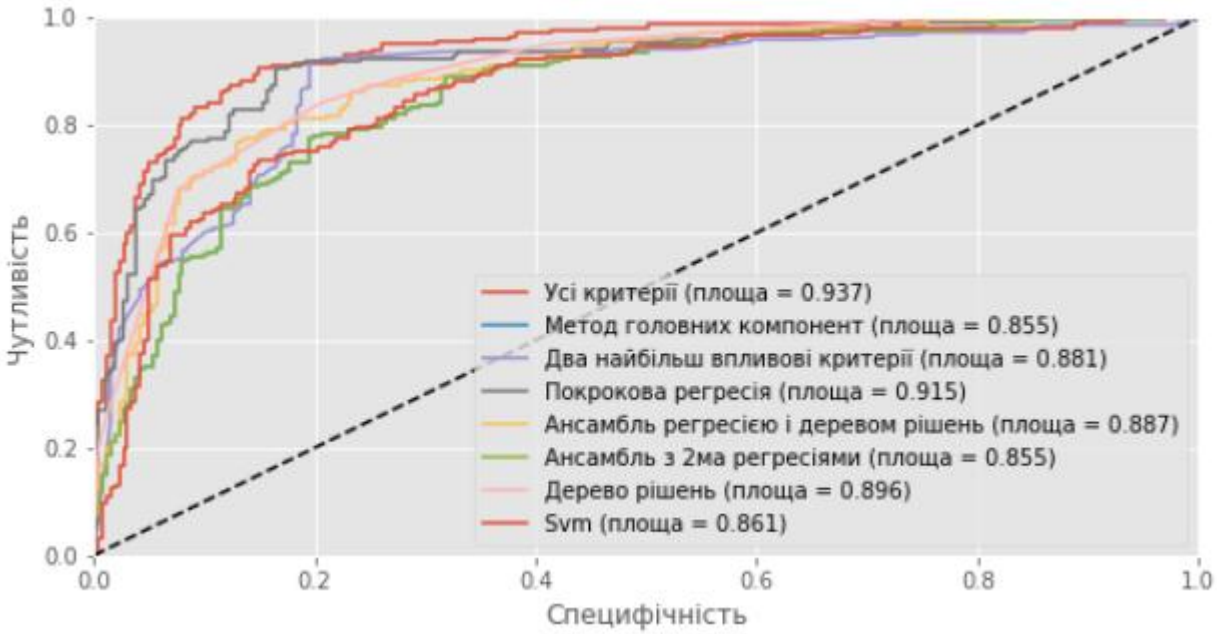

Рисунок 4.8 – ROC-криві порівняння усіх методів

Таблиця 4.1 – Порівняння методів навчання першої задачі

	Тренувальна вибірка	Тестова Вибірка
Усі критерії	88%	90%
Дві головні інтегральні компоненти, сформовані методом компонентного аналізу	77%	81%
Два найбільш значущі критерії, обрані покроковою логістичною регресією	85%	87%
Усі значущі критерії обрані методом покрокової логістичної регресії	86%	88%
Ансамбль з деревом рішень	93%	77%
Ансамбль з двома регресіями	78%	81%
Метод дерева рішень	82%	76%
Метод SVM	78%	77%

Як видно з таблиці 4.1, найточніші результати було отримано за допомогою логістичної регресії з використанням усіх критеріїв, а найгірший за допомогою ансамблю з деревом рішень.

4.2 Задача Fraud скорінгу

Нажаль, на сьогоднішній день не є рідкістю випадки шахрайства під час спроби отримання кредиту. Зазвичай зловмисники заволодівають конфіденційними даними клієнта і використовуючи їх намагаються отримати гроші.

У пункті 4.1 ми розглянули розв’язання задачі кредитного скорінгу, коли клієнта банку було класифіковано до одного з двох класів, чи зможе він повернути кредит, чи ні. Та нажаль, навіть якщо система дозволить видати цій людині кредитну позику, це зовсім не гарантує, що він поверне її, а тим паче, що

зловмисники не обрали цю людину за добру кредитну історію і зараз намагаються отримати кошти шахрайським шляхом.

Ціллю цього пункту є поліпшення результатів кредитного скорінгу, зменшення ризиків для банківських установ і максимізація їх прибутків.

Зазвичай банки збирають величезні обсяги інформації стосовно своїх клієнтів чи можливих клієнтів. Уся ця інформація є конфіденційною, але деякі установи видають обмежені набори даних для допомоги розвитку галузі аналізу даних, але ці дані знеособлюються і модифікуються за для безпеки користувачів. Та усе одно ці дані дуже цінні для нас, бо там є необхідні для класифікації мітки класів і безліч даних, які характеризують клієнта. А оскільки комп'ютеру все одно, що саме означають ті чи інші стовпчики у даних, бо для нього це лише цифри, то можна створити класифікатор, який буде аналізувати дані клієнтів і, аналогічно тому, як це робилося для кредитного скорінгу, буде ставити мітку класу на клієнта, показуючи нам, чи не шахрай він.

Об'єднавши результати обох програм, можна разом підвищити безпеку і прибуток багатьох банківських установ, чи, принаймні, зменшити кількість випадків, коли люди, які отримали дані клієнта, отримують кредит на його ім'я.

Припустимо, що після першого етапу відбору бажаючих отримати кредитну позику, наша вибірка даних зменшилась, і у ній залишились тільки ті, хто за прогнозом класифікатора зможе повернути кошти банку. Нажаль серед них є невеликий процент шахраїв, яких треба викрити до видачі кредиту. Банки зазвичай зберігають різну інформацію стосовно своїх клієнтів, серед яких точно знайдуться спільні критерії, що будуть добре вказувати на шахрайство. Добрим прикладом такого критерію може бути час, який було затрачено на оформлення кредитної заявки через Інтернет-сайт банку, бо людина, яка займається цим постійно, витратить значно менше часу на заповнення, навіть менше, ніж людина, яка вже не один раз отримувала кредитну позику.

Набір даних з відкритих джерел для розв'язання цієї задачі складається з 284807 записів, кожен з яких характеризується 30-ма критеріями.

Для початку навчання першим етапом потрібно обробити вхідні дані. Га-

рною новиною є те, що усі дані представлені у вигляді чисел, а це означає, що не потрібно буде перетворювати категоріальні дані на чисельні, що значно зменшить кінцевий розмір вибірки.

Також вибірку треба перевірити на некоректні значення і на пусті рядки. Проаналізувавши дані, можна побачити, що пропусків даних немає.

Останнім кроком обробки даних буде нормалізація, для того щоб подати їх у числовому вигляді, зручному для роботи алгоритму.

Тепер розділимо дані на 2 частини, щоб використовувати одну з них для навчання нашого класифікатора, а інші дані, на яких класифікатор не мав змоги прогнозувати, будемо використовувати для оцінки якості класифікатора.

Класифікація буде виконуватися методами, які були використані у пункті 4.1 для забезпечення можливості порівняння результатів і моніторингу поліпшення чи погіршення кінцевої точності класифікації.

Перший класифікатор побудуємо на основі логістичної регресії з усіма критеріями. Зазвичай такі класифікатори показують одні з найякісніших класифікацій. Якість класифікації можна оцінити за допомогою ROC-кривої, наведеної на рисунку 4.9.

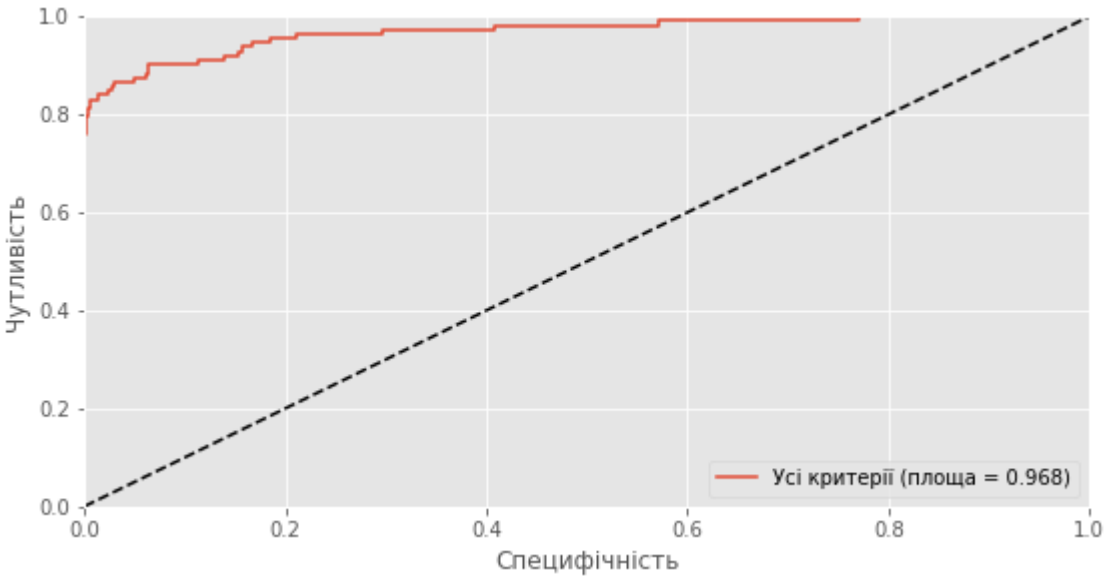

Рисунок 4.9 – ROC-крива класифікатора за усіма ознаками

Як бачимо з графіку, даний класифікатор добре працює с даним типом вибірки. Зауважимо, що точність його класифікації складає 99% для тестової та тренувальної вибірки.

Оскільки критеріїв, за якими потрібно класифікувати клієнтів, більше 30, то було б непогано зменшити розмірність. Для цього, як і у попередній задачі, будемо використовувати метод головних компонент і метод покрокової регресії. Побудуємо ROC-криві для методу головних компонент і покрокової регресії і порівняємо їх (рисунок 4.10).

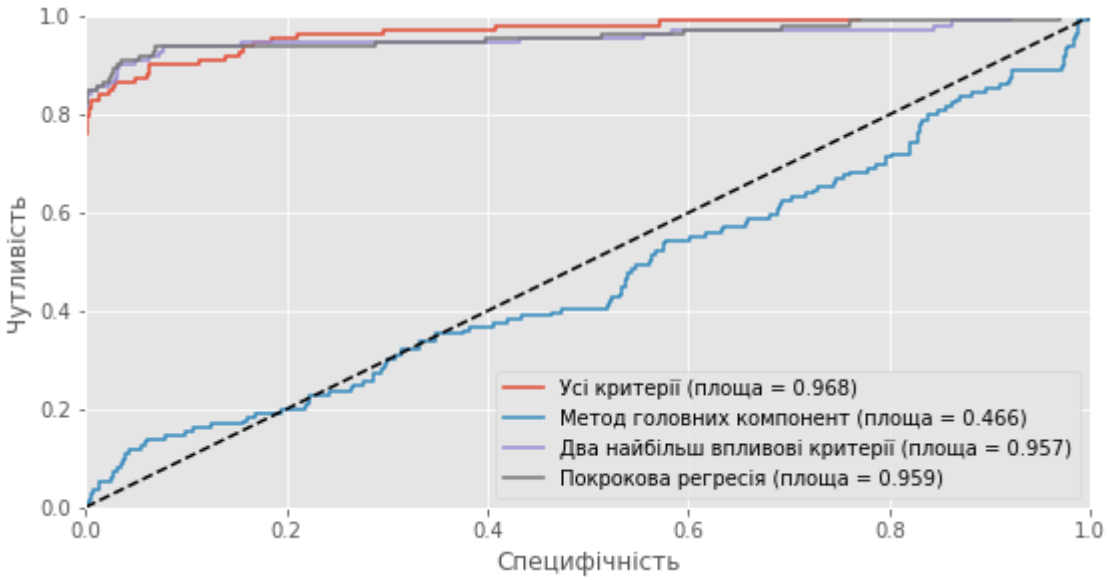

Рисунок 4.10 – ROC-криві порівняння чотирьох методів

За допомогою кожного з цих методів навчання було отримано гарні результати класифікації на рівні 99%, але, як можна побачити з графіку, метод головних компонент не дає гарного результату для даної задачі. Не дивлячись на те, що він, як і усі інші, при прогнозуванні видав високий відсоток точності, ROC-крива показує, що для даної задачі це не найбільш вдалий метод, бо, скоріш за все, стиснути дані з тридцятьох критеріїв до двох не вдається дуже якісно без втрати важливої інформації.

Побудувавши класифікатори, основою котрих не є логістичні регресії, і порівнявши їх з вже отриманими результатами, можна побачити, що логістична

регресія працює з даними такого типу найкраще серед усіх розглянутих класифікаторів (рисунок 4.11).

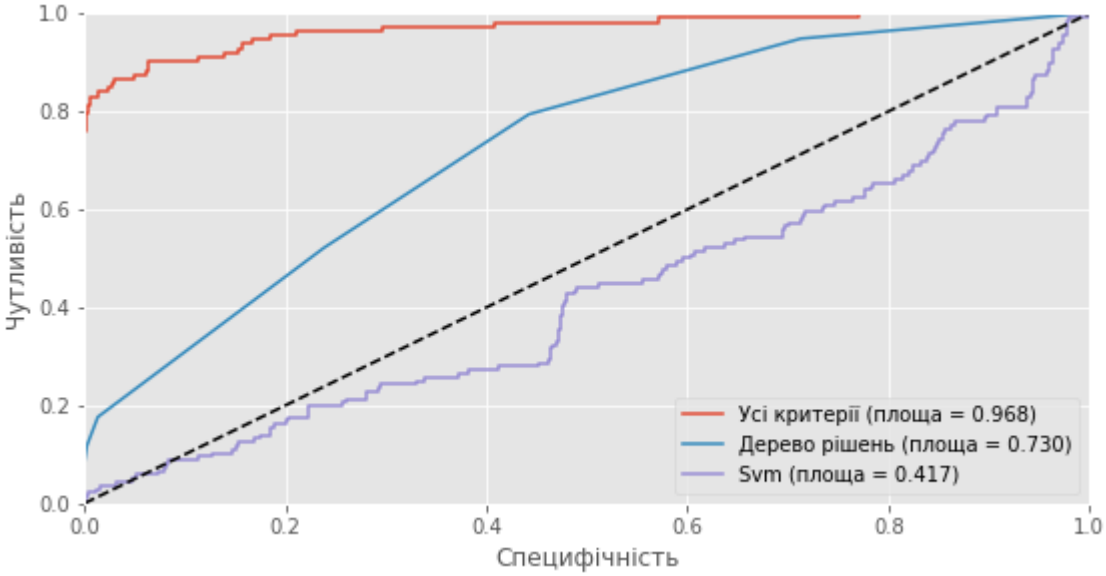

Рисунок 4.11 – ROC-криві порівняння різних методів

Порівняємо також класифікатор, побудований за допомогою логістичної регресії, з ансамблями класифікаторів. Використаємо такі ж комбінації, які були використані у першій задачі (рисунок 4.12).

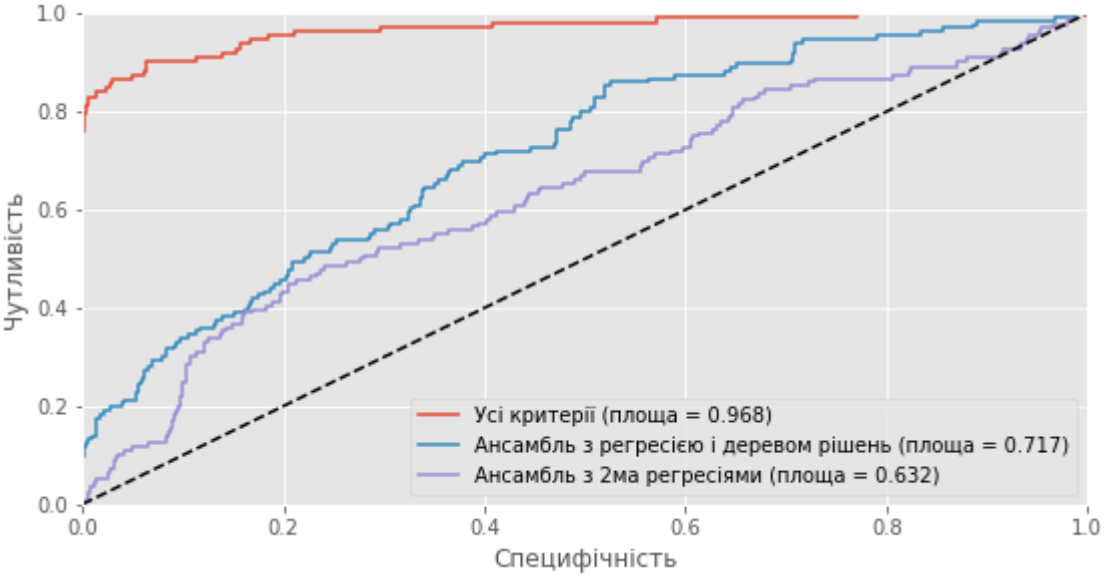

Рисунок 4.12 – ROC-криві порівняння логістичної регресії і ансамблів

У випадку використання двох регресій для даного набору даних точність прогнозування все ще тримається на рівні 99%, але, як видно з графіку, ROC-крива свідчить про те, що зі зміненням вибірки якість може суттєво змінюватися. Якщо звернути увагу на комбінацію логістичної регресії і дерева рішень, то можна побачити картину зовсім протилежну, що свідчить про те, що комбінація різних класифікаторів може значно поліпшити слабкі сторони кожного з класифікаторів для даної вибірки даних.

Скомбінуємо усі ROC-криві на одному графіку для поліпшення порівняння (рисунок 4.13), а також підсумуємо точність класифікації кожного з розглянутих методів у таблиці 4.2.

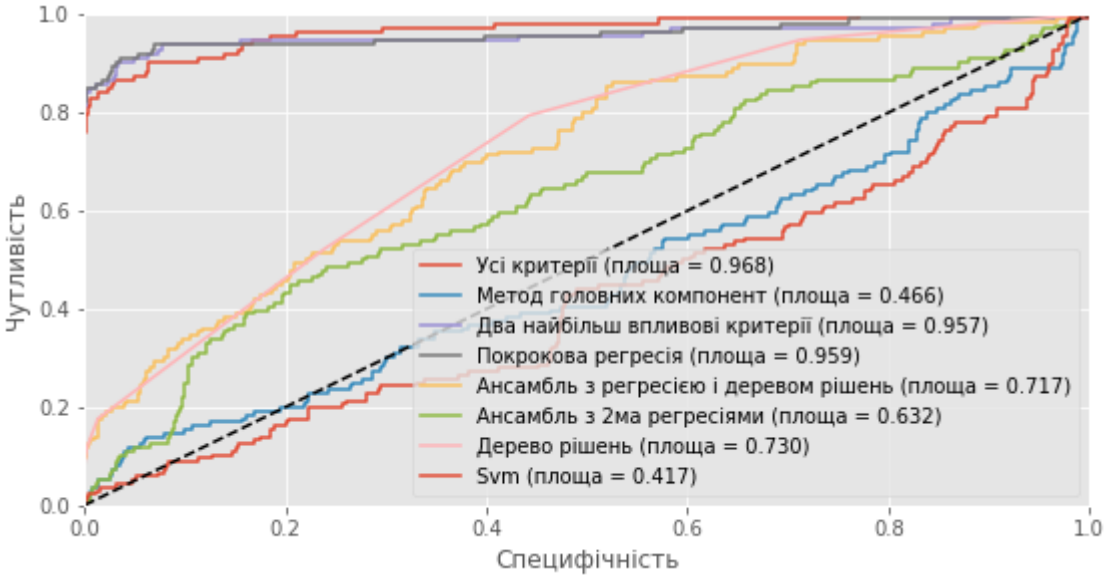

Рисунок 4.13 – Комбінація усіх ROC-кривих

Як бачимо з таблиці, усі методи навчання та класифікатори для даної задачі дозволяють отримати дуже гарні результати класифікації. Але якщо взяти до уваги не тільки відсотки, а й якість класифікаторів, яку можна оцінити за допомогою ROC-кривих, то можна побачити, що не кожен з них є універсальним, і при незначних змінах у даних точність класифікації може змінюватися, що не є доброю властивістю, бо банківські установи потребують стабільних результатів класифікації, що б бути впевненими, що вони не понесуть додаткових збитків.

Таблиця 4.2 – Таблиця порівняння методів навчання другої задачі

	Тренувальна вибірка	Тестова Вибірка
Усі критерії	99,9280%	99,9051%
Дві головні інтегральні компоненти, сформовані методом компонентного аналізу	99,8103%	99,8103%
Два найбільш значущі критерії, обрані покроковою логістичною регресією	99,9227%	99,9823%
Усі значущі критерії обрані методом покрокової логістичної регресії	99,9260%	99,9016%
Ансамбль з деревом рішень	99,8139%	99,9981%
Ансамбль з двома регресіями	99,8139%	99,8103%
Метод дерева рішень	99,8139%	99,9981%
Метод SVM	99,8103%	99,8103%

Таким чином, обробивши дані у два етапи, нам вдалося відсіяти випадки, коли кредитна позика не повернулася банку, а також виключити можливість того, що серед клієнтів, яким видача кредитної позики була дозволена, немає шахраїв.

5 АНАЛІЗ МОЖЛИВИХ ЗАСТОСУВАНЬ

Актуальність задачі, розв’язаної у атестаційній роботі, зумовлена тим, що необхідність класифікації різноманітних об’єктів виникає не лише у банківській діяльності, але ще й у багатьох інших галузях, зокрема, у психології, медицині, поліції тощо. Для поліції даний алгоритм можна з легкістю переробити для класифікації потенційних злочинців, що може значно підвищити якість роботи правоохоронної системи. Злочинця можна буде направити на додаткову реабілітацію, психічну чи медичну, аби уникнути погіршення його стану у майбутньому, щоб не спонукати його до нових злочинів.

Також має місце застосування класифікації у роботі психологів. Можна налаштувати класифікатор таким чином, щоб проводити аналіз за показниками емоційного стану пацієнта і порівнювати з емоційними станами вже відомих системі пацієнтів, і прогнозувати варіанти можливих психічних захворювань.

Не тільки психічні захворювання можна визначити за допомогою класифікатора. Таким же самим чином можна класифікувати людей за їх життєвими показниками і прогнозувати можливі хвороби.

Програма розроблена таким чином, щоб робити усі дії автоматично, незалежно від користувача. Під час використання програми користувачу необхідно зробити наступні дії:

– скласти вибірку статистичних даних;

– завантажити зібрані дані у програму;

– на основі аналізу ROC-кривих зробити висновок про те, наскільки даний класифікатор підходить для класифікації у розглядуваній задачі;

– за умови доброї узгодженості даного класифікатора з розглядуваною задачею бінарної класифікації, прийняти до уваги рекомендації програми щодо прийняття рішення про класифікацію об’єктів.

ВИСНОВКИ

У даній атестаційній роботі було розглянуто проблему застосування методів бінарної класифікації у банківській діяльності для класифікації клієнтів банку за можливістю повернення кредитних позик та за можливістю кредитних махінацій.

За допомогою методу аналізу ієрархій було визначено, що для даної задачі у банківській діяльності найкращим буде класифікатор, заснований на логістичній регресії.

Програмну реалізацію методу аналізу ієрархій було створено за допомогою математичного пакету символьної математики Mathematica 11. Встановивши вагові коефіцієнти для кожного з критеріїв, можна отримати рекомендації для обирання альтернативи за допомогою даного програмного продукту.

Для аналізу статистичних даних клієнтів банку та розв’язання задачі класифікації було створено програмний продукт з використанням мови програмування Python 3. Даний програмний продукт дозволив навчити різні види класифікаторів та обрати найкращий з них для проведення класифікації клієнтів банківської установи відносно можливості повернення кредитних позик чи здійснення фінансових махінацій. Під час роботи алгоритму вдалося досягти точності розподілення клієнтів за класами на рівні 99%. Проблемою під розв’язання задачі стало те, що у відкритих джерелах немає вибірки даних, яка б комбінувала необхідні для обох етапів скорінгу. Тому бажано було б провести додаткові дослідження для уточнення отриманих у роботі результатів на узгодженій для обох етапів класифікацій вибірці даних.

Розроблений у атестаційній роботі програмний продукт можна з легкістю налаштувати для роботи з іншими задачами класифікації в різноманітних галузях медицини, економіки, психології та інших дослідженнях, де виникає потреба у бінарній класифікації об’єктів, які характеризуються великою кількістю критеріїв.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Эконометрия // В. И. Суслов, Н. М. Ибрагимов, Л. П. Талышева, А. А. Цыплаков. Новосибирск : Новосибирский государственный університет, 2005. 742 с.

2. Катренко А. В., Пасічник В. В., Пасько В. П. Теорія прийняття рішень. Київ : Видавнича група BHV, 2009. 448 с.

3. Андрейчиков А. В., Андрейчикова О. Н. Анализ, синтез, планирование решений в экономике. Москва : Финансы и статистика, 2001. 364 с.

4. Катренко А. В. Системний аналіз. Львів : "Новий світ – 2000", 2011. 396 с.

5. Лямец В. И., Тевяшев А. Д. Системний анализ. Вводный курс : учеб. пос. Харьков : ХНУРЭ, 2004. 448 с.

6. Alexander J. McNeil, Rudiger Frey, Paul Embrechts Quantitative Risk Management. New Jersey : Princeton University Press, 2005. 720 с.

7. Бардасов С. А. Эконометрика. Тюмень : Издательство Тюменского государственного университета, 2010. 264 с.

8. Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. Москва : Финансы и статистика, 1989. 607 с.

9. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян. М. С. Куприянов, В. В. Степаненко, И. И. Холод. Санкт-Петербург : БХВ-Петербург, 2007. 384 с.

10. Altman N. S. An introduction to kernel and nearest-neighbor nonparametric regression // The American Statistician. 1992. Vol. 46, № 3. С. 175-185.

11. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer, 2009. 745 с.

12. Brett L. Machine Learning with R. Birmingham – Mumbai : Packt Publishing, 2013. 365 с.

13. Deep Learning. URL : http://www.deeplearningbook.org (дата звернення: 12.03.2019).

14. Флах П. Машинное обучение. М.: ДМК Пресс, 2015. 400 с.

15. Рашка С. Python и машинное обучение. М.: ДМК Пресс, 2017. 418 с.

16. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и Ten-sorFlow. СпБ.: ООО Альфа-книга, 2018. 688 с.

17. Волошин О. Ф., Мащенко С. О. Моделі та методи прийняття рішень. Київ : Видавничо-поліграфічний центр "Київський університет", 2010. 336 с.

18. Нікольський Ю.В., Пасічник В.В., Щербина Ю.М. Системи штучного інтелекту. Львів: Магнолія-2006, 2013. 279 с.