УДК: 004.89:336.77

А. І. Мендус
С. О. Жуков

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ПЕРЕДБАЧЕННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання в інформаційно-аналітичній системі оцінювання кредитоспроможності позичальників і мінімізації фінансових ризиків комерційного банку методами машинного навчання. Проведено комплексний аналіз структури фінансового датасету та його ознак, а також досліджено просторово-часові залежності між ризиком дефолту, встановленим кредитним лімітом, віковими характеристиками та платіжною дисципліною клієнтів.

Ключові слова: кредитний скоринг, машинне навчання, аналіз даних, передбачення дефолту, кредитоспроможність, розвідувальний аналіз.

Abstract

The paper is devoted to the preparation and exploratory data analysis for further use in the information technology of assessing borrowers' creditworthiness and minimizing financial risks of a commercial bank using machine learning methods. A comprehensive analysis of the structure of the financial dataset and its features was conducted, and the spatio-temporal dependencies between default risk, the established credit limit, age characteristics, and the payment discipline of clients were investigated..

Keywords: credit scoring, machine learning, data analysis, default prediction, creditworthiness, exploratory data analysis.

Вступ

Сучасні інформаційні технології інтелектуального аналізу даних відіграють домінуючу роль у фінансово-економічному секторі, зокрема у сфері автоматизації контуру ризик-менеджменту комерційних банків. Кредитний ризик є ключовим фактором деструктивного впливу на стійкість фінансових компаній, оскільки дефолти клієнтів призводять до значних чистих збитків та зниження ліквідності установи.

Алгоритми машинного навчання відкривають нові можливості для побудови гнучких скорингових систем, здатних враховувати складні нелінійні залежності. Проте першочерговим та есенціальним етапом проектування таких технологій є розвідувальний аналіз даних (Exploratory Data Analysis, EDA). Він дозволяє формалізувати вхідні масиви інформації, ідентифікувати латентні патерни поведінки споживачів, виявити проблему асиметрії розподілів та підготувати надійну основу для фази безпосереднього математичного моделювання.

Розвідувальний аналіз

Для проведення аналізу в межах кваліфікаційної роботи було обрано реальний знеособлений набір даних «Default of Credit Card Clients Dataset», доступний у відкритому репозиторії обчислювальної платформи Kaggle [1]. Зазначений масив інформації консолідує відомості про 30 000 клієнтів банківських установ та містить 25 первинних ознак.

Інформаційний контур датасету включає демографічні профілі позичальників, фінансові показники схвалених кредитних лімітів, історію щомісячної платіжної дисципліни, суми виставлених рахунків та фактично здійснених виплат за шестимісячний період. Приклад початкової структури матриці даних із ключовими ознаками наведено на рисунку 1.

Файл завантажено: UCI_Credit_Card.csv
Розмір датасету: 30000 рядків та 25 колонок

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5
0	1	20000.0	2	2	1	24	2	2	-1	-1	...	0.0	0.0
1	2	120000.0	2	2	2	26	-1	2	0	0	...	3272.0	3455.0
2	3	90000.0	2	2	2	34	0	0	0	0	...	14331.0	14948.0
3	4	50000.0	2	2	1	37	0	0	0	0	...	28314.0	28959.0
4	5	50000.0	1	2	1	57	-1	0	-1	0	...	20940.0	19146.0

5 rows × 25 columns

BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.payment.next.month
0.0	0.0	689.0	0.0	0.0	0.0	0.0	1
3261.0	0.0	1000.0	1000.0	1000.0	0.0	2000.0	1
15549.0	1518.0	1500.0	1000.0	1000.0	1000.0	5000.0	0
29547.0	2000.0	2019.0	1200.0	1100.0	1069.0	1000.0	0
19131.0	2000.0	36681.0	10000.0	9000.0	689.0	679.0	0

Рис. 1. Фрагмент структури вхідного набору даних кредитного скорингу

Цільовою змінною дослідження є бінарна ознака default_next_month, яка фіксує настання дефолтної події в наступному звітному місяці (1 – наявність дефолту, 0 – стабільне погашення). Важливою системною характеристикою досліджуваного об'єкта є гострий дисбаланс класів. Графічний розподіл цільової величини відображено на рисунку 2.

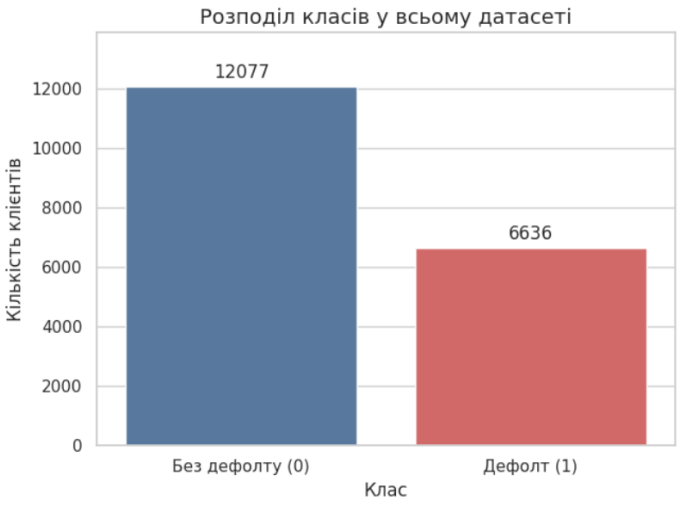

Рис. 2. Графік розподілу цільової змінної дефолту позичальників

Статистичний розподіл свідчить, що 23 364 клієнти (77.9%) належать до мажоритарного стабільного класу, тоді як міноритарна ризикова група налічує лише 6 636 спостережень (22.1%). Співвідношення класів 3.52:1 вимагає подальшого застосування зважування функцій втрат або балансування для уникнення зсуву моделей у бік максимізації загальної точності за рахунок ігнорування ризиків.

У межах двофакторного аналізу було досліджено синергетичний вплив вікових характеристик та фінансових лімітів на динаміку дефолтних випадків (рис. 3).

Рис. 3. Теплова карта щільності дефолтів залежно від віку та ліміту

Аналіз матриці ризиків (рис. 3) дозволив локалізувати критичні зони деструктивного стану системи. Найвища концентрація дефолтів (36.5%) зосереджена в сегменті клієнтів старшої вікової групи (від 61 року), яким було схвалено мінімальний початковий кредитний ліміт (інтервал 10тис–50тис). Загалом спостерігається чітка лінійна тенденція: зі зростанням кредитного ліміту загальний рівень ризику для всіх вікових груп знижується до мінімальних 12–13.8%, що свідчить про вищу фінансову спроможність або ретельнішу попередню верифікацію таких клієнтів банком.

Для формалізації лінійних взаємозв'язків та відсікання колінеарних чинників було побудовано оптимізовану матрицю кореляцій (рис. 4).

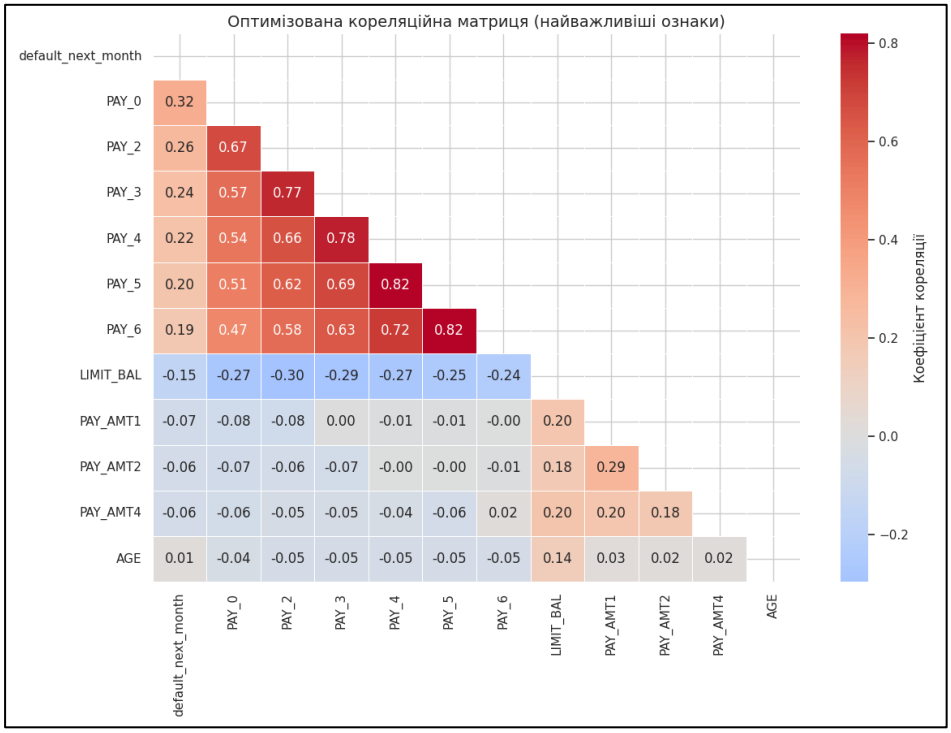

Рис. 4. Оптимізована трикутна матриця кореляцій найважливіших ознак

З аналізу матриці кореляцій (рис. 4) встановлено, що найбільш значущий прямий зв'язок із дефолтною подією мають предиктори платіжної дисципліни клієнта, а саме статус затримки погашення за останній звітний місяць PAY_0 (r = 0.32). Кредитний ліміт LIMIT_BAL має стійкий зворотний зв'язок (r = -0.15).

Крім того, виявлено високу взаємну кореляцію між сумами щомісячних рахунків (BILL_AMT* на рівні 0.80–0.95), що свідчить про ефект мультиколінеарності та обумовлює доцільність порівняння лінійних моделей із ансамблевими методами на основі дерев рішень, стійкими до надлишкових ознак.

Висновки

У ході розвідувального аналізу даних було повністю досліджено структуру та внутрішні закономірності фінансового набору даних «Default of Credit Card Clients Dataset». Встановлено, що ключовими факторами формування кредитного ризику є ретроспективна платіжна дисципліна позичальника (ознаки стану прострочення) та обсяг схваленого банком ліміту. Виявлений гострий дисбаланс класів та сильна внутрішня колінеарність фінансових матриць рахунків обґрунтовують необхідність переходу від класичних статистичних правил до побудови гнучких нелінійних контурів моделювання. Отримані аналітичні результати формують репрезентативну та очищену базу для подальшого проектування аналітичного ядра інформаційної технології скорингу з використанням алгоритмів градієнтного бустингу.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Default of Credit Card Clients Dataset. [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset
2. Pandas Tutorial. [Електронний ресурс] – Режим доступу: https://www.w3schools.com/python/pandas/default.asp
3. Matplotlib Pyplot Documentation. [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
4. Seaborn: Statistical Data Visualization. [Електронний ресурс]. Режим доступу: https://seaborn.pydata.org/

Мендус Андрій Ігорович – студент групи СА-22б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: mendusandriy@gmail.com

Жуков Сергій Олександрович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Mendus Andrii – student of Faculty of Intellectual Information Technologies and Automation, SA-22b, Vinnytsia National Technical University, Vinnytsia, e-mail mendusandriy@gmail.com

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com