

Міністерство освіти і науки України
Національний технічний університет України
"Київський політехнічний інститут"

На правах рукопису

СОЛОШЕНКО ОЛЕКСАНДР МИКОЛАЙОВИЧ

УДК 303.732.4:519.237.5

**МОДЕЛІ І МЕТОДИ ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ
ФІЗИЧНИХ ОСІБ**

01.05.04 – системний аналіз і теорія оптимальних рішень

Дисертація на здобуття наукового ступеня кандидата технічних наук

Науковий керівник:
Бідюк Петро Іванович
доктор технічних наук, професор

Київ – 2016

ЗМІСТ

Вступ.....	7
Розділ 1 Місце та роль системної методології в системних задачах аналізу життєвого циклу роздрібного кредитування з використанням моделей оцінювання кредитоспроможності.....	14
1.1 Актуальність дослідження кредитного ризику та його місце у загальній системі класифікації фінансових ризиків	14
1.2 Кредитний скоринг як методологія класифікації позичальників у ризик-менеджменті.....	17
1.3 Поняття життєвого циклу роздрібного кредитування та можливості застосування різних видів скорингу.....	18
1.4 Поняття аналізу відхилених заявок у кредитному скорингу.....	20
1.5 Поняття інтелектуального аналізу даних як області досліджень системного аналізу.....	22
1.6 Місце моделей оцінювання кредитоспроможності в інтелектуальному аналізі даних. Застосування поза межами ризик-менеджменту.....	24
1.7 Властивості системних задач, властивості та принципи системної методології та інші поняття системного аналізу на прикладі розробки систем оцінювання кредитоспроможності.....	26
1.8 Існуючі сучасні програмні засоби для побудови моделей оцінювання кредитоспроможності клієнтів. Актуальність розробки рішень прогностичної аналітики	35
1.9 Схематичне узагальнення структури системної методології моделювання кредитних ризиків, що відповідає подальшій постановці задач.....	39
1.10 Постановка задач дисертаційного дослідження	41
1.11 Висновки до першого розділу.....	42

Розділ 2	Методи формування вибірки та перетворення параметрів, методи аналізу вхідних характеристик та вибору множини незалежних змінних	43
2.1	Особливості формування вибірки на прикладі аплікаційного скорингу. Період вибірки. Індикатор простроченої заборгованості. Показовий період.....	44
2.2	Методи використання категоріальних змінних. Концепція ваги категорії змінної та інформаційної статистики як показника предикативності	56
2.3	Доведення взаємозв'язку відстані Кульбака-Лейблера з показником інформаційної статистики та показником ваги значення змінної.....	62
2.4	Категоризація змінних. Розробка методу оптимальної дискретизації неперервних змінних за допомогою динамічного програмування Беллмана.....	64
2.5	Індекс Джині та статистика Колмогорова-Смирнова для аналізу однієї вхідної змінної	74
2.6	Розробка методу розрахунку статистики Колмогорова-Смирнова, ваги категорії змінної та інформаційної статистики при відомому розподілі категорій та умовному розподілі цільової змінної.....	76
2.7	Кореляційний аналіз змінних.....	81
2.8	Факторний аналіз. Метод головних компонент	82
2.9	Виключення змінних за допомогою мультиваріаційного аналізу коефіцієнтів побудованої моделі	91
2.9.1	Оцінювання математичного сподівання та дисперсії коефіцієнтів моделі на прикладі логістичної регресії	92
2.9.2	Обчислення рівня статистичної значимості p -value за допомогою функції розподілу Хі-квадрату Вальда	93
2.9.3	Розробка алгоритму обчислення нормованих ваг змінних моделі	96
2.10	Висновки до другого розділу	97
Розділ 3	Методи побудови попередніх моделей, методи включення відхилених заявок, методи оцінювання якості прогнозів та аналіз стійкості моделей.....	99
3.1	Методи побудови попередніх скорингових моделей	99

3.1.1 Регресійні методи моделювання.....	100
3.1.1.1 Лінійна регресія. Метод найменших квадратів. Основний недолік лінійної моделі.....	100
3.1.1.2 Узагальнена ймовірнісна нелінійна регресія. Метод максимальної правдоподібності.....	102
3.1.1.2.1 Пробіт-регресія на основі нормального розподілу	103
3.1.1.2.2 Логістична регресія. Аналітичні формули для градієнта та матриці Гессе.....	103
3.1.1.3 Формула Фробеніуса в лінійній та нелінійній регресії.....	104
3.1.2 Кластерний аналіз та машинне навчання на основі пам'яті.....	105
3.1.2.1 Метод k -середніх	105
3.1.2.2 Метод найближчого сусіда	106
3.1.2.3 Метод k -найближчих сусідів.....	106
3.1.3 Древа рішень	107
3.1.4 Нейронні мережі. Алгоритм оберненого розповсюдження помилки.....	109
3.1.5 Генетичні та еволюційні алгоритми.....	110
3.1.6 Наївний байєсівський класифікатор. Інші методи моделювання	111
3.1.7 Розробка алгоритму прискорення збіжності вектору коефіцієнтів логістичної регресії	111
3.1.8 Вдосконалення, а саме комплексне узагальнення логістичної регресії та ваги категорії змінної для ймовірнісної цільової змінної	112
3.1.9 Вдосконалення методу k -найближчих сусідів	114
3.1.10 Калібрування карти балів відносно логарифму відношення шансів.....	116
3.2 Методи включення та аналізу відхилених заявок.....	117
3.2.1 Метод присвоєння нульового результату відхиленим заявкам	117
3.2.2 Метод присвоєння аналогічної пропорції	118
3.2.3 Метод повного ігнорування відхилених заявок.....	118
3.2.4 Метод тимчасового погодження всіх заявок для збору даних.....	118

3.2.5	Метод використання аналогічних даних банку або кредитних бюро	118
3.2.6	Метод доповнення для експертного процесу прийняття рішень	119
3.2.7	Метод простого доповнення («жорстке відсікання»)	120
3.2.8	Метод доповнення на основі ймовірностей погодження заявок	120
3.2.9	Метод розбиття груп ризику відхилених заявок	120
3.2.10	Метод нечіткого доповнення	122
3.2.11	Метод ітеративної класифікації	124
3.2.12	Метод найближчого сусіда (кластеризація)	125
3.2.13	Метод виводу на основі пам'яті	125
3.2.14	Інші методи аналізу відхилених заявок	126
3.2.15	Вдосконалення методу ітеративної класифікації	127
3.3	Методи оцінювання якості прогнозів та аналіз стійкості моделей	129
3.3.1	Оцінка якості прогнозів скорингових моделей	129
3.3.1.1	Методи розрахунку показника Джині	130
3.3.1.1.1	Операційна характеристика приймача (ROC-крива)	130
3.3.1.1.2	Крива Лоренца для накопичення відсортованої вибірки	133
3.3.1.1.3	Крива Лоренца функцій кумулятивного розподілу класів	135
3.3.1.2	Статистика Колмогорова-Смирнова	136
3.3.1.3	Відстань Махаланобіса	138
3.3.1.4	Статистика Хосмера-Лемешоу	139
3.3.1.5	Розробка алгоритму розрахунку показника Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса засобами мови SQL	140
3.3.1.6	Вдосконалення, а саме узагальнення індексу Джині та статистики Колмогорова-Смирнова для ймовірнісної цільової змінної	141
3.3.1.7	Cross-валідація скорингової карти. Метод «leave-one-out»	143
3.3.2	Аналіз стійкості скорингових моделей	143
3.3.2.1	Індекс стійкості розподілу	144
3.3.2.2	Звіти розрахунку зміщення скорингового балу	144

3.3.2.3 Статистика Колмогорова-Смирнова та стійкість розподілів	145
3.4 Висновки до третього розділу	146
Розділ 4 Система підтримки прийняття рішень для побудови довільних моделей оцінювання кредитоспроможності та результати проведених експериментів	148
4.1 Архітектура системи підтримки прийняття рішень	148
4.2 Середовище розробки системи та формат вхідних даних	149
4.3 Впровадження та технічні вимоги для коректної роботи програми	149
4.4 Результати обчислювальних експериментів для споживчого кредитування з наведенням етапів моделювання у рамках графічного інтерфейсу системи	150
4.4.1 Завантаження навчальних даних на прикладі MS SQL Server	150
4.4.2 Аналіз предикативної сили вхідних змінних і кореляційний аналіз ..	151
4.4.3 Побудова моделі логістичної регресії. Аналіз відхилених заявок	152
4.4.4 Аналіз якості прогнозів на тестовій вибірці	154
4.4.5 Порівняння з класичним методом ітеративної класифікації	155
4.4.6 Представлення скорингової карти балів	156
4.4.7 Збереження налаштувань змінних	157
4.5 Результати застосування вдосконалення методу k -найближчих сусідів ..	157
4.6 Деякі характеристики моделі для індивідуального кредитування	158
4.6.1 Завантаження навчальних даних з електронної таблиці MS Excel	158
4.6.2 Дослідження статистичної значимості коефіцієнтів моделі	159
4.6.3 Результати перехресного тестування моделі	159
4.7 Приклад дискретизації при побудові моделі колекторського скорингу ...	160
4.8 Висновки до четвертого розділу	162
Висновки	163
Список використаних джерел	166
Додаток А Акт впровадження результатів дисертаційної роботи	178
Додаток Б Принцип збереження чисел у форматі IEEE 754	179
Додаток В Інформаційна статистика та кореляції на всій множині заявок	180

ВСТУП

Актуальність теми. Дисертаційна робота присвячена застосуванню універсальної прикладної наукової методології системного аналізу для вирішення системних задач ризик-менеджменту, зокрема задач математичного моделювання та прогнозування кредитоспроможності клієнтів банку на різних етапах життєвого циклу роздрібного кредитування в умовах неповноти, невизначеності, нечіткості, зашумленості та суперечливості вхідної інформації. Основною сучасною системною методологією прогнозування кредитних ризиків є кредитний скоринг, що полягає у розробці математичних моделей спеціального типу – скорингових моделей, які також називаються скоринговими картами, метою яких є прогнозування майбутнього стану обслуговування позичальником заборгованості або прогнозування довільних поведінкових показників по договору, клієнту, виходячи з соціально-демографічних характеристик, параметрів кредитного продукту, минулих поведінкових індикаторів.

Теоретичні засади концептуальної парадигми системного аналізу закладені, зокрема, академіком В.І. Вернадським, О.О. Богдановим, Л. фон Берталанфі, Н. Вінером, Т. Котарбінським, а також розвинені в сучасних роботах академіка М.З. Згуровського, Н.Д. Панкратової, В.Д. Романенка, О.А. Павлова, В.Я. Данилова, М.А. Айзермана, Н.З. Шора, Т. Сааті, О.В. Антонова, В. Кінга, Д. Кліланда, М.М. Моїсеєва, Р.Л. Акоффа, Дж. Кліра та інших вчених.

Відомими сучасними теоретиками і практиками в області управління ризиками є професори Ю.П. Зайченко, В.М. Подладчиков, Джонатан Н. Крук, Лін С. Томас, Девід Дж. Хенд, Л.М. Любчик, доктор Елізабет Мейз, дослідники Наїм Сіддікі, Девід Б. Едельман. Значний внесок у дослідження задач бінарної класифікації за допомогою логістичної регресії зробили Девід В. Хосмер, Стенлі Лемешоу, Пол Д. Елісон. Першим вченим, хто застосував підхід класифікації популяції на прикладі рослин був Рональд Ейлмер Фішер у 1936 р., а першим

дослідником, який застосував дану методику для бінарної класифікації кредитів у 1941 р., будучи таким чином основоположником кредитного скорингу, є Девід Дюран, що написав фундаментальну книгу «Елементи ризику у фінансуванні споживчої розстрочки». Системна методологія кредитного скорингу з точки зору системного аналізу відповідає фундаментальним принципам процедурної відкритості та раціональної доповнюваності, зокрема. До недоліків, подолання яких має найвищу актуальність, насамперед відносяться: відсутність чітких обмежень, критеріїв оптимальності для основних методів дискретизації змінних, неможливість забезпечення глобального оптимуму для множини таких методів, незастосовність множини таких методів для випадку ймовірнісної цільової змінної, відсутність формул обчислення ваг категорій та інформаційної статистики вхідної змінної в термінах її безумовного розподілу та умовного розподілу цільової змінної, визначеність ваг категорій вхідних змінних та їх інформаційної статистики, а також класичної бінарної логістичної регресії лише для випадку бінарної цільової змінної, бінарна визначеність проміжних та фінальних оцінок класів у методі ітеративної класифікації та необхідність застосування порогу відсікання на ітераціях методу, визначеність ключових показників якості прогнозів – індексу Джині, статистики Колмогорова-Смирнова – лише для випадку бінарних фактичних значень цільової змінної, присутність фактору масштабу змінних у регресійних моделях, що не дозволяє реально оцінювати ступені впливу змінних, відсутність рекомендацій щодо ініціалізації початкового вектору коефіцієнтів у методі максимальної правдоподібності, низька застосовність популярних класичних методів чисельного інтегрування при оцінюванні рівнів статистичної значимості коефіцієнтів логістичної регресії, існуючі недоліки методу k -найближчих сусідів.

Зв'язок роботи з науковими програмами, планами, темами

Дисертаційна робота виконана на кафедрі математичних методів системного аналізу Навчально-наукового комплексу «Інститут прикладного

системного аналізу» Національного технічного університету України «Київський політехнічний інститут» у відповідності до планів науково-дослідних робіт: (1) д/б НДР «Розробка інформаційної технології моделювання та оцінювання фінансово-економічних ризиків із врахуванням невизначеностей різної природи (на основі байєсівських моделей)» (№ ДР: 0113U000650 2622-п), 2013 – 2014 рр.; (2) д/б НДР «Розробка методології системного аналізу, моделювання та оцінювання фінансових ризиків» (№ ДР: 0115U000356 2813-п), 2015 – 2016 рр.

Мета і завдання дослідження

Мета дослідження – вдосконалення системної методології побудови моделей оцінювання кредитоспроможності, зокрема її цілісне узагальнення на всіх рівнях для випадку ймовірнісної цільової змінної, яке забезпечує суттєве вдосконалення аналізу відхилених заявок; розробка та вдосконалення методів і алгоритмів обчислення ключових показників, методів моделювання; розробка оригінального програмного продукту. Для досягнення мети потрібно вирішити такі завдання:

1. розробити формалізований метод дискретизації неперервних вхідних змінних на основі динамічного програмування у моделях оцінювання кредитоспроможності;

2. розробити метод розрахунку статистики Колмогорова-Смирнова і ваг категорій змінних та інформаційної статистики за умови відомого розподілу категорій та умовному розподілі цільової змінної;

3. розробити алгоритм обчислення рівнів статистичної значимості для оцінок моделі логістичної регресії шляхом інтегрування розкладу в ряд Тейлора;

4. вдосконалити модель логістичної регресії, методи її аналізу та оцінювання ваг категорій змінних, інформаційної статистики, індексу Джині, статистики Колмогорова-Смирнова для випадку ймовірнісної цільової змінної; вдосконалити метод ітеративної класифікації з метою подолання його головних недоліків;

5. розробити алгоритми нормування ваг змінних регресійної моделі з урахуванням варіації вхідних параметрів і прискорення процесу збіжності вектора коефіцієнтів логістичної регресії;

6. удосконалити модель на основі методу k -найближчих сусідів для задач машинного навчання моделей для оцінювання кредитоспроможності фізичних осіб;

7. розробити алгоритм розрахунку показника Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса засобами мов програмування четвертого покоління (4GL) на прикладі мови SQL, а також архітектуру системи підтримки прийняття рішень для побудови скорингових моделей і здійснити програмну реалізацію мовою програмування Visual C#.

Об'єкт дослідження: база даних з поведінковими та аплікаційними характеристиками клієнтів.

Предмет дослідження: моделі і методи оцінювання кредитоспроможності фізичних осіб.

Методи дослідження: регресійні методи моделювання, непараметричний метод машинного навчання на основі пам'яті – вдосконалений метод k -найближчих сусідів, методи теорії інформації для оцінювання нерівності розподілів, метод динамічного програмування, статистичні методи обчислення рівня значимості, методи теорії випадкових процесів, методи математичного аналізу, методи теорії оптимізації, методи теорії мов програмування, методи системного аналізу: метод відкриття, метод постулювання та їх комбінації (при побудові та валідації моделей).

Наукова новизна одержаних результатів. Проведені у дисертаційній роботі дослідження дозволили суттєво вдосконалити цілісну, багаторівневу, багатоетапну методологію побудови моделей оцінювання кредитоспроможності фізичних осіб, зокрема в умовах узагальнення на випадок ймовірнісної цільової змінної.

Уперше:

- розроблено метод дискретизації неперервних вхідних змінних на основі динамічного програмування Беллмана з критерієм максимізації інформаційної статистики за умовами трьох обмежень: (1) необхідна кількість інтервалів, (2) мінімально допустима частка інтервалу, (3) кратність кроку дискретизації;
- розроблено метод розрахунку значень статистики Колмогорова-Смирнова, ваг категорій змінної та інформаційної статистики при відомому розподілі категорій та умовному розподілі цільової змінної.

Удосконалено:

- метод комплексного моделювання кредитного ризику за моделлю логістичної регресії, вагами категорій змінних, інформаційною статистикою, оцінками ступенів впливів змінних, рівнями статистичної значимості, індексом Джині, статистикою Колмогорова-Смирнова, а також методами включення і аналізу відхилених заявок;
- метод k -найближчих сусідів для розв'язання задач бінарної класифікації.

Набуло подальшого розвитку:

- інтерпретація відстані Кульбака-Лейблера.

Практичне значення одержаних результатів

1. У результаті дослідження вдосконалено та програмно реалізовано цілісну системну методологію побудови довільних скорингових моделей як для бінарного, так і для неперервного ймовірнісного значення цільової змінної. Розроблена система підтримки прийняття рішень застосовна для побудови довільних скорингових карт з метою аналізу різноманітних етапів життєвого циклу роздрібного кредитування.

2. Результати дисертації у складі моделей, методів та програмних засобів були впроваджені в ПАО «БАНК ФОРВАРД» (акт впровадження вих. № 15/4-05-

960 від 07.11.2014). Впровадження пропонованої методології та програмних засобів в ПАО «БАНК ФОРВАРД» дали можливість вирішити проблеми побудови системи аплікаційного скорингу для споживчого кредитування, побудови поведінкового скорингу для прогнозування прострочення найближчого платежу в рамках проекту організації попереднього збору заборгованості.

Особистий внесок здобувача

Всі основні наукові положення та результати, що складають основний зміст роботи та становлять наукову новизну, отримані автором самостійно.

У доповіді [1], опублікованій у співавторстві, здобувачеві належить розробка та програмна реалізація алгоритму розрахунку індексу Джині та статистики Колмогорова-Смирнова засобами мови структурованих запитів SQL.

У доповіді [2], опублікованій у співавторстві, здобувачеві належать налаштування та аналіз процедури логістичної регресії, а також розробка аналітичних додатків засобами SAS Enterprise Guide 4.3.

У праці [3], опублікованій у співавторстві, здобувачеві належить обґрунтування формули зміщеної ймовірності.

Апробація результатів роботи

Основні положення були представлені на наукових конференціях та семінарах:

– V^й всеукраїнській науково-практичній конференції «Інформаційно-комп'ютерні технології в економіці, освіті та соціальній сфері» (Україна, Сімферополь, 2010);

– міжнародній науково-практичній конференції молодих учених і студентів «Інформаційні процеси і технології «Інформатика – 2012» (Україна, Севастополь, 2012);

– 9^й міжнародній науково-практичній конференції «ІНТЕРНЕТ-ОСВІТА-НАУКА-2014 «ІОН-2014» (Україна, Вінниця, 2014);

– міжнародній конференції «Розвиток інформаційно-ресурсного забезпечення освіти і науки в гірничо-металургійній галузі і транспорті 2014» (Україна, Дніпропетровськ, 2014);

– XI^й міжнародній науково-практичній конференції «Актуальні питання й організаційно-правові основи міжнародного співробітництва в сфері високих технологій» (Україна, Київ, 2014);

– науково-технічній конференції «Інформатика, математика, автоматика «ІМА :: 2015» (Україна, Суми, 2015);

– науковому семінарі «Системні дослідження та інформаційні технології» при навчально-науковому комплексі «Інститут прикладного системного аналізу» (Україна, Київ, 09 грудня 2015 року).

Публікація результатів

За матеріалами дисертаційного дослідження опубліковано 13 наукових праць: серед них 6 статей у провідних наукових фахових виданнях (у тому числі 1 – в іноземному виданні; у тому числі 1 – в українському виданні, що входить до міжнародних наукометричних баз даних), 1 статтю у електронному виданні, 6 праць у матеріалах доповідей міжнародних та національних конференцій.

Структура дисертаційної роботи

Дисертаційна робота складається зі вступу, змісту, чотирьох основних розділів, висновків, списку використаних джерел, трьох додатків. Робота викладена на 180 сторінках і містить 165 сторінок основної частини, 53 рисунки (серед них 4 у додатках), 16 таблиць, список використаних джерел із 114 найменувань, 3 додатки.

РОЗДІЛ 1

МІСЦЕ ТА РОЛЬ СИСТЕМНОЇ МЕТОДОЛОГІЇ В СИСТЕМНИХ ЗАДАЧАХ АНАЛІЗУ ЖИТТЄВОГО ЦИКЛУ РОЗДРІБНОГО КРЕДИТУВАННЯ З ВИКОРИСТАННЯМ МОДЕЛЕЙ ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ

Мета першого розділу полягає у висвітленні актуальності дослідження кредитного ризику та в ознайомленні з теорією, щоб означити основні поняття, зокрема кредитного ризику, його складових, дефолту та пов'язаних з ними сутностей. У даному розділі вводиться означення поняття скорингу, життєвого циклу роздрібного кредитування, аналізу відхилених заявок, інтелектуального аналізу даних, системної методології, системної задачі, складної ієрархічної системи. Розглядаються особливості побудови скорингових моделей в умовах мінливої української дійсності та здійснюється аналіз світового ринку інформаційних технологій в області забезпечення рішеннями для скорингового моделювання. Методологія побудови скорингових моделей аналізується з точки зору системного аналізу як системна методологія, що відповідає фундаментальним властивостям та принципам, а задачі побудови скорингових карт розглядаються як сукупності системних задач з відповідними властивостями та особливостями. Також рішення в області побудови скорингових моделей розглядаються як складні ієрархічні системи з визначеним рівнем організаційної ієрархії та складності прийняття рішення. Наприкінці розділу формулюється постановка задач дисертаційного дослідження. У висновках даного розділу наводиться остаточний узагальнюючий аналіз використаного матеріалу з джерел щодо основних понять наведених у попередніх підрозділах.

1.1 Актуальність дослідження кредитного ризику та його місце у загальній системі класифікації фінансових ризиків

У фінансовому ризик-менеджменті поняття ризику (risk) переважно означає можливість втрати частини власних ресурсів, недоотримання доходів або виникнення додаткових затрат в результаті здійснення підприємницької діяльності, що відповідає поняттю чистої невизначеності (uncertainty), що передбачає лише можливість негативних (збиткових) відхилень кінцевого результату діяльності [4]. Також, згідно з неокейнсіанським напрямком економічної науки, категорія ризику вважається ширшою, ніж категорія невизначеності, завдяки наявності кількісної оцінки ймовірності реалізації певних подій, що забезпечується наявністю статистичних даних за попередні періоди, а неокласична школа розглядає ці два поняття як тотожні [4, 5].

З точки зору системного аналізу, ризик (risk), у сенсі загального поняття, характеризується як можливість виникнення та результат дії небажаних ситуацій, умов та факторів, що визначаються випадковими та хаотичними процесами [6]. При цьому ризики створюються багатьма різноманітними внутрішніми та зовнішніми факторами, тому механізми їх впливів являються багатофакторними [6]. У рамках термінології системного аналізу визначаються два основні показники ризику: ступінь ризику і рівень ризику [6, 7]. Ступінь ризику визначається як ймовірність появи події, що призводить до небажаних наслідків, а рівень ризику визначається як розмір потенційного збитку у разі впливу факторів ризику [6, 7].

У фінансовому ризик-менеджменті, у відповідності зі стандартною класифікацією, головними загрозами для благополуччя фінансового інституту є: ринкові ризики (зокрема валютний ризик та відсотковий ризик), кредитні ризики (зокрема ризик контрагента, ризик дефолту та ризик дострокового погашення), операційні ризики (включаючи модельний ризик та ризик неадекватності методів оцінки та управління ризиками), ризики ліквідності (зокрема ризик ринкової ліквідності та ризик балансової ліквідності), ризики події (зокрема юридичні

ризика, бухгалтерські ризики, податкові ризики, ризики репутації, ризики дій регулюючих органів) [4].

Ще в 1997 році Базельський комітет по банківському нагляду в своєму документі «Основоположні принципи ефективного банківського нагляду» назвав кредитний ризик основним видом фінансового ризику, з яким стикаються фінансові інститути в своїй діяльності. Будучи найбільш поширеним, а отже й актуальним, видом фінансового ризику, кредитний ризик є елементом невизначеності при виконанні контрагентом своїх договірних зобов'язань, пов'язаних з поверненням позикових засобів. Іншими словами, кредитний ризик – це можливість втрат унаслідок нездатності контрагента виконати свої контрактні зобов'язання. Для кредитора наслідки невиконання цих зобов'язань вимірюються втратою основної суми заборгованості, неоплачених відсотків, затрат збору заборгованості і т.д. за вирахуванням суми відновлених грошових коштів. Кредитний ризик включає ризик країни і ризик контрагента [4].

Найбільш яскравим проявом кредитного ризику є дефолт (default) – невиконання контрагентом через нездатність або небажання умов кредитної угоди або ринкової операції. Тому до категорії кредитного ризику відносяться, в першу чергу, втрати, пов'язані з оголошенням контрагентом дефолту. Крім того, до кредитного ризику відносяться також і втрати, пов'язані з пониженням кредитного рейтингу позичальника, оскільки це зазвичай призводить до пониження ринкової вартості його зобов'язань, а також втрати у вигляді недоотриманого прибутку унаслідок дострокового повернення позики позичальником [4]. Найчастіше постановкою задачі для скорингу в поняттях ризику контрагента є прогнозування показників, що є ранніми індикаторами по відношенню до дефолту, але з високою ймовірністю до нього призводять: прикладом таких індикаторів може бути ознака досягнення певної кількості днів прострочення заборгованості протягом всього певного показового періоду

спостереження або в його кінцевій точці, починаючи з точки початку спостереження (наприклад, дати видачі ануїтетного кредиту) [5, 8–10].

При управлінні кредитним портфелем у фінансовому ризик-менеджменті поняттю ступеню ризику системного аналізу відповідає, наприклад, поняття ймовірності дефолту (probability of default, PD), а поняттю рівня ризику системного аналізу відповідає, наприклад, поняття відносних втрат у випадку дефолту (loss given default, LGD) [4, 8, 9]. Також при управлінні портфелем у ризик-менеджменті вводиться поняття схильності кредитному ризику (credit exposure, CE) або величини кредитної вимоги, що підлягає ризику дефолту (exposure at default, EAD), тобто величина суми кредитного запиту або фактичного балансу. Тоді очікувані втрати (expected loss, EL) або очікувані кредитні втрати (expected credit loss, ECL), що можуть виступати як розмір резервування за портфелем, записуються таким чином [4, 8, 9]:

$$ECL = EL = \sum_{i=1}^N CE_i \cdot PD_i \cdot LGD_i = \sum_{i=1}^N EAD_i \cdot PD_i \cdot LGD_i,$$

де N – кількість кредитних угод або кредитних запитів.

Актуальність розробки, розвитку та впровадження рішень прогностичної аналітики в цілому (включаючи кредитний скоринг та інші види скорингу) описується в підрозділі 1.8.

1.2 Кредитний скоринг як методологія класифікації позичальників у ризик-менеджменті

Кредитний скоринг (credit scoring, від англ. score – рейтинг) або аплікаційний скоринг (application scoring) – це методологія оцінювання кредитоспроможності потенційних позичальників у ризик-менеджменті [11], або

методологія класифікації потенційних клієнтів (контрагентів) банку по ступеню (рівню) ризику [4], або набір моделей прийняття рішень та основоположних технік, що допомагають кредиторам в процесі вирішення питання надання споживчого кредиту [5]. Скоринг – методологія оцінювання кредитоспроможності або майбутньої поведінки на рівні клієнтів або договорів, як потенційних, так і існуючих [11], тому існує багато категорій скорингу: кредитний (аплікаційний) скоринг, поведінковий скоринг, скоринг виявлення та попередження шахрайства, колекторський скоринг, інші численні категорії скорингу [5, 8–11].

1.3 Поняття життєвого циклу роздрібного кредитування та можливості застосування різних видів скорингу

Під життєвим циклом (life cycle) [5, 12] роздрібного кредитування вводиться означення множини концептуальних послідовних етапів роботи банку з клієнтами у процесі функціонування системи споживчого кредитування, наприклад:

1) рекламні або маркетингові акції по залученню потенційних клієнтів схильних до отримання кредиту на основі наявних даних про них [8–10] (область застосування скорингу схильності до відгуку – propensity scoring, response scoring);

2) акції по попередній аплікаційній оцінці потенційних клієнтів при наявних анкетних, фінансово-економічних, історичних даних про них (область застосування скорингу попереднього погодження – pre-approval scoring) [5];

3) визначення і відсікання шахрайських аплікаційних заявок на отримання кредиту від потенційних клієнтів (область застосування скорингу виявлення та попередження шахрайства – fraud scoring або fraud prevention scoring) [5, 8, 9];

4) визначення і відхилення заявок некредитоспроможних клієнтів (область застосування кредитного скорингу – application \ credit scoring) [5, 8–11, 13];

5) формування відсоткових ставок і комісій згідно балу, а саме скорингової групи, до якої належить потенційний позичальник (область застосування скорингу ціноутворення на основі ризику – risk-based pricing scoring) [5, 8, 9];

6) оцінка подальшої платоспроможності при поточному нульовому значенні днів прострочення (days past due, DPD) для існуючих клієнтів або угод (область застосування поведінкового скорингу – behavioral scoring) [5, 8–10];

7) оцінка ймовірності виникнення заборгованості по існуючих угодах, починаючи з найближчої дати вимоги платежу, з метою здійснення превентивних дій нагадування про найближчий платіж (область застосування скорингу попереднього збору заборгованості – pre-collection scoring);

8) оцінка подальшого погіршення днів прострочення заборгованості (DPD) до певного критичного значення за певний період часу – Performance Window або його збіжність до нуля по існуючих клієнтах або договорах з метою організації ефективного збору проблемної кредитної заборгованості (область застосування скорингу збору заборгованості або колекторського скорингу – collection(s) scoring і також скорингу судового стягнення – litigation scoring) [5, 10];

9) оцінка досягнення певного рівня прибутку банку за існуючими рахунками клієнта (область застосування скорингу прогнозування потенційного прибутку – profit scoring) [5];

10) організація маркетингових кампаній та перехресних продажів для існуючих прибуткових клієнтів (область застосування скорингу перехресних продажів і скорингу збільшення лімітів – cross-sell \ up-sell scoring) [5, 8, 9];

11) утримання існуючих клієнтів у разі їх відтоку та відмови від послуг банку (область застосування скорингу утримання клієнтів та скоринг прогнозування зміни кредитора – retention \ *churn* \ attrition scoring) [5, 8, 9];

12) прогнозування портфельного показника відносних втрат у випадку дефолту (loss given default, LGD) як відсотку втрат спричинених реалізацією події дефолту (область застосування скорингу прогнозування відносних втрат у випадку дефолту – LGD scoring) [10]. Значний теоретичний та практичний внесок у даний вид скорингу буде описано у наступних розділах.

На рисунку 1.1 зображено різні види скорингу у розрізі життєвого циклу.



Рисунок 1.1 – Застосування видів скорингу на життєвому циклі кредитування

Комбіноване використання скорингових моделей різного типу та призначення застосовується, наприклад, при використанні мультискорингового підходу до управління портфелем (multi-score approach for portfolio management) [10], наприклад, за допомогою матричного комбінування (тобто поєднання у системі ортогональних координат) [10]. Прикладом мультискорингового підходу може бути виділення кредитоспроможних позичальників, схильних до маркетингових впливів, де для визначення кредитоспроможності та ступеню схильності до відгуку на рекламу використовуються дві окремі оціночні скорингові карти [10]. Можливості скорингу у сфері банківської діяльності виходять за межі кредитування, популярним застосуванням скорингу утримання клієнтів (retention scoring) є прогнозування та моделювання відтоку депозитних вкладів фізичних осіб [2] з метою проведення превентивних заходів з утримання.

1.4 Поняття аналізу відхилених заявок у кредитному скорингу

Аналіз відхилених заявок (reject inference) – надзвичайно важлива та перспективна новітня системна методологія у рамках системної методології скорингового моделювання, що полягає у включенні у навчальну вибірку заявок, що були відхилені згідно з кредитно-ризиковою політикою або з попередньою

скоринговою картою (scorecard) у процесі прийняття рішення щодо можливості кредитування, тому по яким невідомий фактичний історичний бінарний цільовий результат (performance) – невідома бінарна цільова змінна (binary target variable), а відомі аплікаційні характеристики [5, 8–10, 13–15]. На рисунку 1.2 представлено проблему невизначеності якості обслуговування кредиту у разі відмови [13, 14].

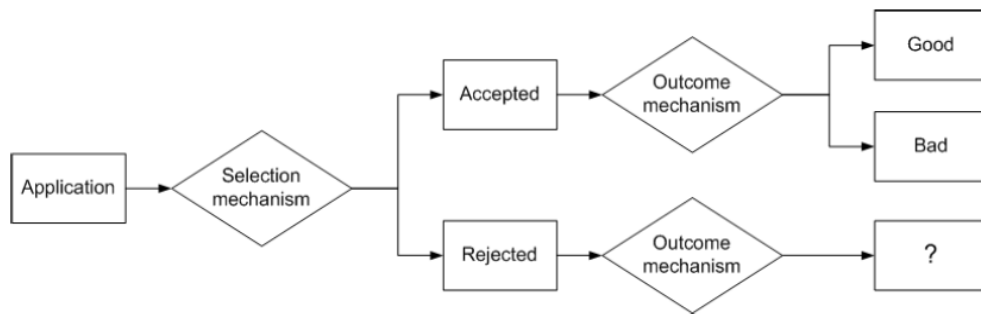


Рисунок 1.2 – Проблема невизначеності цільової змінної в аплікаційному кредитному скорингу для відхилених заявок [13, 14]

Потреба аналізу відхилених заявок виникає при необхідності забезпечити відсутність зміщення (*bias*) розподілів та статистичних оцінок [8, 9, 14, 15], тобто при необхідності забезпечити стійкість розподілів вхідної множини (або генеральної сукупності заявників або заявок) відносно навчальної вибірки (як у розрізі фінального скорингового балу, так і в розрізі розподілу кожної окремої змінної моделі) [8–10, 13], та задля адекватної екстраполяції [5, 10, 14, 15] скорингової моделі на заявки, що були відхилені у процесі кредитного аналізу. Аналіз відхилених заявок притаманний лише задачам побудови аплікаційних моделей (типу моделей кредитного скорингу). Конкретні методи моделювання з урахуванням аналізу відхилених заявок детально описуються в підрозділі 3.2. Суть більшості методів аналізу відхилених заявок полягає у розширенні значень цільової змінної на її невизначені значення, тобто на підмножину відхилених заявок, в межах генеральної сукупності (рисунок 1.3), що означає присвоєння певних класів відхиленим заявкам [8, 9].

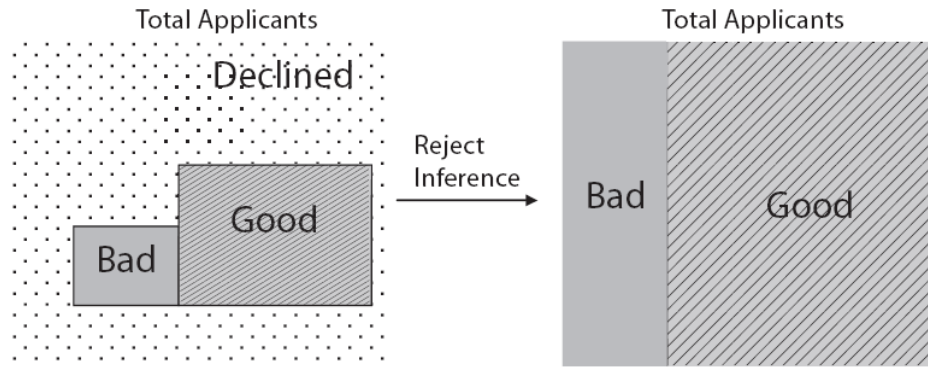


Рисунок 1.3 – Розширення значень цільової змінної на множину відхилених заявок – присвоєння певних класів відхиленим заявкам [8, 9]

1.5 Поняття інтелектуального аналізу даних як області досліджень системного аналізу

Термін «інтелектуальний аналіз даних» (ІАД) походить від англomовного поняття «data mining» [5, 8, 9, 13, 16, 17], що отримав свою назву від двох понять: (1) дані – «data» та (2) видобуток гірської руди – «mining», тому термін «data mining» перекладається як видобуток даних, витяг інформації, розкопка даних, інтелектуальний аналіз даних, засіб пошуку закономірностей, витягування знань, аналіз шаблонів, «витягування зерен знань з гір даних», розкопка знань у базах даних, інформаційна проходка даних, «промивання» даних [16]. Сучасне поняття «виявлення знань у базах даних» (KDD – Knowledge Discovery in Databases) можна вважати синонімом інтелектуального аналізу даних [16]. З ІАД також тісно пов'язане поняття «Data Science» (наука про дані). Поняття інтелектуального аналізу даних з'явилося в 1989 році, але високу популярність у сучасному трактуванні набуло приблизно в першій половині 1990-х років [16]. До цього часу обробка та аналіз даних здійснювалися в рамках прикладної статистики, при цьому в основному вирішувалися завдання обробки невеликих баз даних [16]. Інтелектуальний аналіз даних – це процес підтримки прийняття рішень, який ґрунтується на пошуку в даних прихованих закономірностей [16]. Досить

точне означення запропонував Григорій П'ятецький-Шапіро (Gregory Piatetsky-Shapiro) – один із засновників напрямку: «Інтелектуальний аналіз даних – це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних та доступних інтерпретацій знань, необхідних для прийняття рішень у різних сферах людської діяльності» [16, 18]. Тобто суть та ціль технології полягають у пошуку неочевидних, об'єктивних та корисних на практиці закономірностей у великих обсягах даних [16]. Неочевидних – означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом [16]. Об'єктивних [16] – означає, що виявлені закономірності будуть повністю відповідати дійсності, на відміну від експертної думки, яка завжди суб'єктивна. Практично корисних – означає, що висновки мають конкретне значення, якому можна знайти практичне застосування [16]. Знання – сукупність відомостей, що утворить цілісний опис, який відповідає деякому рівню поінформованості про предмет, проблему або питання, що розглядається [16]. Використання знань означає реальне застосування віднайдених знань для досягнення конкретних переваг (наприклад, у конкурентній боротьбі за ринок) [16]. Інтелектуальний аналіз даних – це процес виділення з даних неявної та неструктурованої інформації, а також її подання у вигляді, придатному для використання [16]. Інтелектуальний аналіз даних можна віднести до системного аналізу як наукового методу пізнання, оскільки інтелектуальний аналіз даних дозволяє вирішувати довільні міждисциплінарні задачі шляхом параметричної та структурної ідентифікації на основі поглибленого аналізу накопичених емпіричних даних. З точки зору системного аналізу, інтелектуальний аналіз даних – це міждисциплінарна область, що поєднує в собі щонайменше такі науки як [16, 17]: (1) прикладна статистика; (2) розпізнавання образів; (3) машинне навчання; (4) штучний інтелект; (5) теорія баз даних; (6) теорія алгоритмів та структур даних. До основних задач ІАД належать [16, 17]: (1) класифікація; (2) кластеризація; (3) асоціація; (4)

послідовна асоціація; (5) прогнозування; (6) визначення відхилень або викидів; (7) оцінювання; (8) аналіз зв'язків; (9) візуалізація; (10) підведення підсумків.

1.6 Місце моделей оцінювання кредитоспроможності в інтелектуальному аналізі даних. Застосування поза межами ризик-менеджменту

Кредитний скоринг можна вважати одночасно основоположним витоком, найуспішнішою областю застосування і методологічною підмножиною інтелектуального аналізу даних (ІАД) [5]. У рамках інтелектуального аналізу даних, побудову скорингових моделей можна віднести щонайменше до семи класичних задач ІАД: класифікації, прогнозування, оцінювання, аналізу зв'язків, візуалізації, асоціації та послідовної асоціації. Задачі аналізу зв'язків, візуалізації, асоціації та послідовної асоціації розв'язуються на етапах кореляційного аналізу та аналізу предикативної (прогностичної) сили вхідних змінних. Задача класифікації використовується як при постановці задачі моделювання, так і при побудові моделі та оцінюванні якості прогнозів, на ряду з задачами прогнозування та оцінювання. Основна область застосування скорингового моделювання – управління ризиками (ризик-менеджмент), але в цілому скорингові моделі застосовні для довільних задач бінарної класифікації, діагностики, прогнозування ймовірностей виникнення певної довільної події, виявлення певної прихованої ознаки через призму спостережуваних ознак з певною ймовірністю. Прикладами інших областей використання скорингових моделей є медицина [5], наприклад, рання діагностика захворювання через ряд спостережуваних симптомів, або аналіз фінансового ринку та аналіз часових рядів, наприклад, короткострокове прогнозування знаку зміни ціни активу в залежності від спостережуваної кон'юнктури ринку. Також важливою областю застосування є розпізнавання образів: візуальних образів, звуків, тексту, опрацювання цифрових сигналів. Одним з прикладів областей тісно пов'язаних з

методологією скорингу є методи колаборативної фільтрації при аналізі користувачів в інтернеті, що дозволяють роздрібним торговцям рекомендувати товари чи мультимедійну продукцію [19]. Ще однією областю суміжною з методологією скорингу є виявлення груп, наприклад, з метою систематизації блогерів в інтернеті або з метою виявлення кластерів вподобань в інтернеті за допомогою методів кластеризації [19]. Не менш важливими областями суміжними зі скорингом є пошук та ранжування при розробці пошукових машин, а також машинне навчання нейронних мереж на основі дій користувачів [19]. Також з методологією скорингу пов'язані оптимізаційні задачі групових подорожей та підбору авіарейсів, оптимізація з урахуванням вподобань при розподілі обмежених ресурсів, оптимізація візуалізації мереж (задача про оптимальне розміщення) [19]. Пов'язаною з методологією скорингу є розробка антиспамних фільтрів за допомогою ймовірнісного наївного байєсівського класифікатора [19]. Також важливою областю суміжною з методологією скорингу є прогнозування кількості платних реєстрацій в інтернеті, моделювання ступеню привабливості та моделювання цін на нерухомість за допомогою дерев рішень (decision trees) [19]. Не менш важливою областю застосування методів суміжних з методологією скорингу є підбір пар на сайтах знайомств та у соціальних мережах (типу Facebook) за допомогою ядерних методів та машин опорних векторів (Support Vector Machines, SVM) [19]. Ідентифікація основних тем в масивах новин за допомогою невід'ємної матричної факторизації (Non-negative Matrix Factorization, NMF) з метою виявлення незалежних ознак також пов'язана з опціональними етапами побудови скорингових моделей, що описані в дисертаційній роботі (факторний аналіз) [19]. У даній дисертаційній роботі надалі наводяться нові можливості скорингового моделювання, що виходять за рамки задач бінарної (або категоріальної) класифікації, розширюючи задачу бінарного дискретного вибору на регресійну задачу з ймовірнісною цільовою змінною або відповідно з неперервною, але обмеженою інтервалом значень.

1.7 Властивості системних задач, властивості та принципи системної методології та інші поняття системного аналізу на прикладі розробки систем оцінювання кредитоспроможності

Системний аналіз – це прикладна наукова методологія, що спирається на широке різноманіття системно організованих, структурно взаємопов'язаних та функціонально взаємодіючих евристичних процедур, методичних прийомів, математичних методів, алгоритмічних програмних та обчислювальних засобів, що забезпечує формування цілісних, міждисциплінарних знань про досліджуваний об'єкт як про сукупність взаємопов'язаних процесів різної природи для подальшого прийняття рішення відносно його подальшого розвитку та поведінки об'єкта з урахуванням множини конфліктуючих критеріїв та цілей, наявності факторів ризику, неповноти та недостовірності інформації [6, 7, 20]. Системний аналіз – універсальна міждисциплінарна загальнонаукова прикладна методологія, наступний крок у розвитку сучасної науки, а саме перехід до багатовимірної науки [6, 7, 20]. Системна задача – це задача аналізу сукупності властивостей об'єкта з єдиної позиції цілісного підходу для досягнення заданих цілей за наявних умов [6, 7]. Задачам системного аналізу властиві: (а) багатокритеріальність; (б) багатофакторність; (в) багатопараметричність [6, 7, 21]. Також системні задачі характеризуються рядом факторів [6, 7, 22]: (1) структурна впорядкованість та взаємозалежність зв'язків між множинами вихідних даних задачі та множинами кінцевих результатів її розв'язання; (2) узгодженість та впорядкованість вимог до вказаних множин; (3) рівень обчислювальної складності задачі; (4) ступінь потенційної структурованості та рівень потенційної формалізованості задачі. Перші два фактори характерні також для задач в рамках аксіоматичних дисциплін, а останні два – є специфічними ознаками системних задач [6, 7, 22]. Системні задачі характеризуються неповнотою, неточністю, суперечливістю, невизначеністю вихідної інформації [6, 7, 23]. Системний аналіз є науковим процесом або власне

методологією [24]. Фундаментальні властивості системної методології: (а) результативність; (б) ефективність; (в) масштабність [6, 7]. Фундаментальні принципи системної методології [6, 7]: (1) принцип системної узгодженості; (2) принцип процедурної повноти; (3) принцип функціональної ортогональності; (4) принцип інформаційної взаємозалежності; (5) принцип цілеспрямованої відповідності; (6) принцип функціональної раціональності; (7) принцип багатоцільової загальності; (8) принцип багатофакторної адаптивності; (9) принцип процедурної відкритості; (10) принцип раціональної доповнюваності. Основні методи системного аналізу – методи дослідження систем: (а) метод відкриття, (б) метод постулювання та їх комбінації [6, 7]. Щодо додаткової термінології, що використовується в рамках дисертаційної роботи, поняття «шар» у системному аналізі означає термін, що визначає рівень складності прийняття рішення, а саме: ієрархічна структура, що відповідає поняттю «шар», передбачає, що процедуру ухвалення рішення реалізують у вигляді послідовності часткових процедур, де кожна забезпечує одержання розв'язку з ступенем обґрунтованості, вірогідності за різних рівнів неповноти, невизначеності, нечіткості і суперечливості вхідної інформації [6, 7]. Таку ієрархічну структуру прийнято називати багатошаровою, багаторівневою або ієрархічною системою прийняття рішень [6, 7]. Щодо управління ризиками з точки зору системного аналізу, то особливість задачі системного аналізу ризиків полягає в тому, що вона є основою послідовності задач кількох типів, які традиційно вивчають у різних наукових напрямках [6, 7]: (1) багатофакторна класифікація ризику – задача кластерного аналізу; (2) багатофакторне розпізнавання ризиків – задача теорії розпізнавання; (3) ранжування ситуацій ризику – задача теорії прийняття рішень; (4) багатоцільова мінімізація ризиків – головна задача аналізу ризиків. Щодо аналізу стійкості моделей, у рамках системного аналізу вводиться поняття «стійкості системи» – властивості виконувати свої функції при виході параметрів середовища за певні обмеження [25]. У таблиці 1.1 проведено аналогії між термінологіями системного аналізу [6, 7] і методології скорингу на прикладі, зокрема, кредитного скорингу.

Таблиця 1.1 – Проведення аналогій між термінологіями системного аналізу та методології побудови скорингових моделей на прикладі кредитного скорингу

Група термінів системного аналізу	Термін системного аналізу	Термін або суть проблеми кредитного скорингу
1	2	3
Показники ризику	Ступінь ризику	Ймовірність дефолту (probability of default, PD)
	Рівень ризику	Відносні втрати у випадку дефолту (loss given default, LGD)
Якісні властивості інформації	Невизначеність	Наявність у базах даних альтернативних анкетних даних досліджуваного клієнта (без різниці у часі заповнення або без можливості однозначного зіставлення з остаточним рішенням про видачу кредиту)
	Неточність	Похибки та неточності в анкетних даних клієнтів та (або) подальших обчисленнях
	Неповнота	Проблема врахування та аналізу відхиленних заявок (reject inference) у кредитному скорингу
	Нечіткість	Суб'єктивна розпливчастість трактування притаманна деяким якісним параметрам клієнта, що не мають кількісної оцінки або точного індикатора стану
	Несвоєчасність	Обмежена можливість відслідковування подальшої динаміки зміни вихідних даних про клієнта від моменту видачі
	Недостовірність	Використання недостовірних даних про клієнта (в т.ч. наданих з метою шахрайства)
	<i>Суперечливість</i>	(а) суперечливість вхідних анкетних даних; (б) <i>розв'язок несумісних систем рівнянь при статистичному моделюванні</i>

Продовження таблиці 1.1

1	2	3
Фундаментальні принципи системної методології	принцип системної узгодженості	Методи кредитного скорингу структурно взаємопов'язані та функціонально взаємозалежні. Наприклад, p -value та знак коефіцієнта логістичної регресії можуть свідчити про погану якість відсікання по кореляційному аналізу чи аналізу предикативності
	принцип процедурної повноти	Методологія кредитного скорингу охоплює весь цикл побудови та супроводу моделі: від формалізації задачі (означення «good\bad») до верифікації і контролю стійкості
	принцип функціональної ортогональності	При програмній реалізації процесу моделювання, функції одного етапу не залежать від функцій іншого
	принцип інформаційної взаємозалежності	Кожен етап моделювання залежить інформаційно від попередніх, в т.ч. ROC-аналіз від логістичної регресії
	принцип цілеспрямованої відповідності	<u>Всі етапи</u> моделювання прямо або опосередковано <u>спрямовані на максимізацію роздільної здатності</u> (через Gini, K.-S., Likelihood, I.V.)
	принцип функціональної раціональності	Етапи моделювання скорингових карт не виконують дублювання функцій, не мають надлишковості
	принцип багатоцільової загальності	Методи кредитного скорингу застосовні для побудови моделей довільного скорингу і класифікації
	принцип багатофакторної адаптивності	Методи моделювання не мають суттєвих обмежень кола задач, мають високий ступінь гнучкості
	принцип процедурної відкритості	<i>Існуючі процедури та методи скорингу підлягають повній заміні, вдосконаленню або агрегуванню</i>
	принцип раціональної доповнюваності	<i>Методологія кредитного скорингу відкрита до розширення області застосування та арсеналу методів</i>

Продовження таблиці 1.1

1	2	3
Фундаментальні властивості системної методології	Результативність	Різноманітні методи побудови скорингових моделей дозволяють дослідникам отримати прийнятний результат (якщо це можливо, виходячи з наявності прихованих взаємозв'язків) як з позиції прийнятної роздільної здатності моделі, так і з бажаного масштабу шкали та способу інтерпретації скорингового балу, так і з позиції обґрунтованості локальних трендів на рівні окремих змінних. Також багато залежить від типу обраної моделі, наприклад, досить зручними, але дещо відмінними, є інтерпретації дерева рішень та логістичної регресії, але зовсім інший та складніший спосіб кінцевої інтерпретації результатів має, наприклад, метод k -найближчих сусідів
	Ефективність	Затрати часу та обчислювальних ресурсів при побудові моделей кредитного скорингу на надзвичайно великих даних (big data) можуть бути знижені шляхом вибору найбільш швидкодіючих методів з доступного арсеналу методів
	Масштабність	Методологія кредитного скорингу застосовна до розв'язання широкого кола інших задач, в т.ч.: медична діагностика, розпізнавання візуальних образів, звуків, тексту, аналіз часових рядів, обробка неструктурованих текстів, соц. мережі, маркетинг, веб-аналітика, опрацювання цифрових сигналів

Продовження таблиці 1.1

1	2	3
Складні ієрархічні системи	Шар (у певному сенсі)	<p>Загальну процедуру побудови скорингових моделей на прикладі кредитного скорингу можна здебільшого описати у вигляді послідовності часткових процедур:</p> <ol style="list-style-type: none"> (1) формування вибірки; (2) дискретизація неперервних змінних та оцінка предикативної сили характеристик; (3) кореляційний аналіз змінних; (4) побудова моделі та мультиваріаційний аналіз змінних; (5) аналіз відхилених заявок (при необхідності); (6) оцінювання якості прогнозів; (7) аналіз стійкості скорингової моделі; (8) калібрування балу скорингу; (9) визначення порогу відсікання, скорингових груп, стратегій
Основні методи системного аналізу (методи дослідження систем)	Метод відкриття	<p>Побудова моделі кредитного скорингу по суті є виводом зашумленої породжувальної системи (системи вищого рівня) шляхом <i>виводу</i> системи (моделі) <i>із заданої системи даних</i> – навчальної вибірки, що складається з завершених кредитних історій позичальників. Мають місце індуктивні висновки, що базуються на системі даних, при побудові дворівневої моделі на прикладі логістичної регресії: (1) рівень ваг категорій змінних (Weight of Evidence, WoE) окремої змінної; (2) рівень коефіцієнтів логістичної регресії для множини змінних</p>

Продовження таблиці 1.1

1	2	3
Основні методи системного аналізу (методи дослідження систем)	Метод постулювання	Застосовується при валідації (тестуванні) скорингової моделі, коли проводиться етап оцінювання якості прогнозів моделі. При цьому гіпотетична породжувальна система або система вищого рівня (скорингова модель), що, наприклад, пов'язує параметри клієнта або угоди з ймовірністю виникнення дефолту, постулюється, а потім її правильність перевіряється порівнянням породжених нею прогнозних даних з емпіричними даними за допомогою оцінювання якості ймовірнісних прогнозів відносно бінарної цільової змінної (індекс Джині, статистика Колмогорова-Смирнова, відстань Махаланобіса, площа під ROC-кривою (Area Under Curve, AUC), логарифм правдоподібності і т.д.)
Загальна термінологія системного аналізу	Багатофакторний ризик	Ризик виникнення дефолту або ризик досягнення певного рівня прострочення заборгованості створюється багатьма різноманітними внутрішніми та зовнішніми факторами, тому механізм впливу ризику є багатофакторним. При побудові моделей, зокрема кредитного скорингу, найбільш предикативні, доступні та відносно взаємно незалежні фактори включаються в модель на етапах оцінювання предикативної сили, кореляційного аналізу (іноді використовуючи ще й факторний аналіз) та мультиваріаційного аналізу змінних

Продовження таблиці 1.1

1	2	3
Набір факторів відмінностей системних задач	Структурна впорядкованість і взаємозалежність зв'язків між множинами вихідних даних задачі та множинами остаточних результатів її розв'язання	<p>Методи структурної і параметричної ідентифікації дають детермінований точний або визначений з заданою точністю результат, що при фіксованих налаштуваннях методів залежить лише від матриці вимірів та вектору вимірів цільової змінної.</p> <p>Після структурної ідентифікації аналітичні обчислення та рекурсивні процедури дають детерміновану модель у процесі вже параметричної ідентифікації, а чисельні методи – детерміновану модель з заданою точністю. Генетичні та еволюційні алгоритми, методи машинного навчання теж з високою ймовірністю сходяться до визначеного навчальною множиною оптимального результату</p>
	Узгодженість та впорядкованість вимог до множин вихідних даних задачі та множин остаточних результатів її розв'язання	<p>Методи моделювання передбачають, що формат вхідних змінних є або категоріальним зі скінченної множини значень або числовим неперервним або цілочисельним. Цільова змінна зазвичай вважається бінарною. Обрана підмножина вхідних змінних характеризується (а) предикативністю; (б) відносною незалежністю. Також, наприклад, до класу моделей типу логістичної регресії побудованої на показниках ваг категорій змінних (WoE) часто ставиться вимога позитивної визначеності до її результуючих коефіцієнтів-множників з метою запобігання інверсій трендів ризику</p>
	Рівень обчислювальної складності задачі	Обчислювальна складність побудови моделі залежить від обраних методів моделювання, дискретизації і т.д.

Продовження таблиці 1.1

1	2	3
Набір факторів відмінностей системних задач	Ступінь структурованості та рівень потенційної формалізованості задачі	Постановка задачі побудови моделей кредитного скорингу та моделей скорингу довільного типу зазвичай слабо структурована та слабо формалізована, зокрема, мають місце елементи загальності, невизначеності, неточності, неповноти, нечіткості, суперечливості або навіть технічної недоскопості за наявних даних (наприклад, рівня індексу Джині) у постановці вимог та формулюванні технічного завдання (ТЗ)

Даний досить детальний та обширний оригінальний список аналогій між термінологіями системного аналізу та методологією скорингу не охоплює повністю по суті безмежний перелік спільних елементів міждисциплінарної науки з методологією скорингового моделювання. Також це ще раз підтверджує, що керування кредитним ризиком має здійснюватися на засадах системного аналізу та адекватного врахування множини керованих і некерованих чинників – спиратися, зокрема, на якісний та кількісний аналіз кредитного ризику [26], оскільки системний підхід до розв’язання довільних задач прогнозування дозволяє організувати процес обробки даних в умовах наявності невизначеностей структурного, параметричного і статистичного характеру [27], що власне й притаманні, зокрема, процесу побудови моделей кредитного скорингу (а також довільного скорингу). Впровадження системного підходу до прогнозування обсягів втрат внаслідок реалізації банківських ризиків [28] також дає змогу ефективно керувати портфелем, зокрема процесом резервування за кредитними операціями згідно з міжнародними стандартами. Також з системним підходом пов’язаний сучасний процес побудови інформаційних систем

підтримки прийняття рішень (ІСППР або інформаційних СППР) [29]. У книзі творця системи Wolfram Mathematica[®] та вільної бази знань Wolfram|Alpha Стівена Вольфрама «Новий вид науки» («A New Kind of Science») [30] згадується про зростаючу роль загальної теорії систем та статистичного аналізу у новітній науці майбутнього з новими методами та ідеями.

1.8 Існуючі сучасні програмні засоби для побудови моделей оцінювання кредитоспроможності клієнтів. Актуальність розробки рішень прогнозу аналітики

Основними існуючими конкуруючими програмними рішеннями в області побудови скорингових моделей на сучасному ринку інформаційних технологій є такі програмні засоби: (1) сучасні ефективні рішення компанії SAS Institute або SAS Institute Inc. [8–10, 17, 31–35], що займають значну долю на сучасному ринку аналітичного банківського програмного забезпечення (<http://www.sas.com/>): (1.1) SAS[®] Enterprise Miner[™]; (1.2) SAS[®] Credit Scoring for Banking; (2) рішення однієї з найбільш відомих і найперших американських компаній, спеціалізованих саме на кредитному скорингу, яку часто вважають першою (1950-ті роки) та найуспішнішою консалтинговою компанією в області кредитного скорингу – FICO, попередня назва: Fair Isaac Corporation (<http://www.fico.com/>) [5, 10]: (2.1) FICO[™] Model Builder, зокрема FICO[™] Model Builder's Scorecard module (MBS); (3) спеціалізоване для побудови довільних скорингових моделей рішення бізнес-відділу Plug&Score компанії Scorto Corporation, яка має досить численний та важливий штат співробітників у м. Харків, Україна, а саме ключовий дослідницько-розробницький (девелоперський) центр: Scorto R&D Center (веб-сторінки: <http://www.plug-n-score.com/> та <http://www.scorto.com/>): (3.1) Plug&Score Modeler; (4) рішення IBM (раніше SPSS Inc. [10, 34, 35], <http://www.ibm.com>), що може використовуватись для побудови ефективних скорингових моделей: (4.1) IBM[®] SPSS Modeler, і, зокрема, додатковий адаптер, який потребує встановлення

з метою впровадження побудованих за допомогою основної програми скорингових моделей безпосередньо в базу даних – IBM® SPSS Modeler Server Scoring Adapter; (5) рішення SAP SE (раніше KXEN Inc., <http://go.sap.com/>): (5.1) програмний продукт «*KXEN: Scoring*», зокрема модуль KXEN Model Export, що дозволяє експортувати побудовані моделі; (6) рішення австралійської компанії *Tiberius Data Mining* (<http://www.tiberius.biz/>): (6.1) *Scorecard Builder*, де перевага віддається сплайнам лінійної регресії перед логістичною регресією; (7) рішення компанії StatSoft – тепер дочірня компанія компанії Dell (<http://www.statsoft.com/>): (7.1) *STATISTICA Scorecard*; (8) частково деякі рішення від компанії *Experian* (<http://www.experian.com/>); (9) рішення британської компанії Paragon Business Solutions (<http://www.credit-scoring.co.uk/>): (9.1) *Paragon Modeller*; (10) рішення канадської компанії Angoss (<http://www.angoss.com/>): (10.1) *KnowledgeSTUDIO™*; (11) рішення російської компанії BaseGroup Labs (<http://basegroup.ru/>): (11.1) *Deductor Credit Scorecard Modeler*. Також до потужних статистичних засобів, що можуть бути налаштовані та використані в цілях побудови скорингових моделей, можна віднести проект вільної, потужної статистичної мови програмування R [34–36] (при цьому існує додаток з графічним інтерфейсом користувача – RapidMiner, що також, окрім R, використовує напрацювання системи Weka для машинного навчання). До засобів, що можуть бути налаштованими для побудови скорингових моделей, можна віднести статистичний пакет EViews (Econometric Views), щонайменше три бібліотеки мови програмування Python: SciPy, NumPy [19, 37, 38], pandas [37, 38], щонайменше одну бібліотеку мови Java: Apache Mahout [39]. До переліку варто додати продукти компанії Wolfram Research, зокрема систему Wolfram Mathematica® з інтерпретатором мови програмування [30], а також систему Wolfram Finance Platform на основі базової системи. Перевагами системи підтримки прийняття рішень, що розроблюється в рамках дисертаційної роботи, є: (1) максимальна простота, зручність та зрозумілість графічного інтерфейсу з мінімальною необхідністю залучення користувача в

процес побудови довільної скорингової моделі; (2) максимально простий процес використання програми без необхідності її встановлення для сімейства операційних систем типу Microsoft Windows™, єдиною вимогою є лише наявність встановленої платформи .NET Framework (не нижче версії 3.5); (3) залежність швидкодії, продуктивності системи лише від фізичних характеристик ПК, обчислювальної платформи (оперативна пам'ять, частота, ядерність процесора і т.д.); (4) реалізація новітніх методів, алгоритмів побудови скорингових моделей та обчислень індикаторів, що запропоновані в дисертаційному дослідженні. Актуальність розробки, розвитку та впровадження рішень прогнозу аналітики (Predictive Analytics), до якої відноситься в першу чергу кредитний скоринг, доводиться масою досліджень перспективності та зрілості сучасних високих технологій. Компанія Gartner [17] у 2013 році поставила на перше місце прогнозу аналітику як технологію, що вийшла на плато продуктивності (рисунок 1.4), яка має переваги щодо вирішення широкого кола задач, є стійкою технологією, перевіреною часом, застосовується в різних областях науки, бізнесу, виробництва.

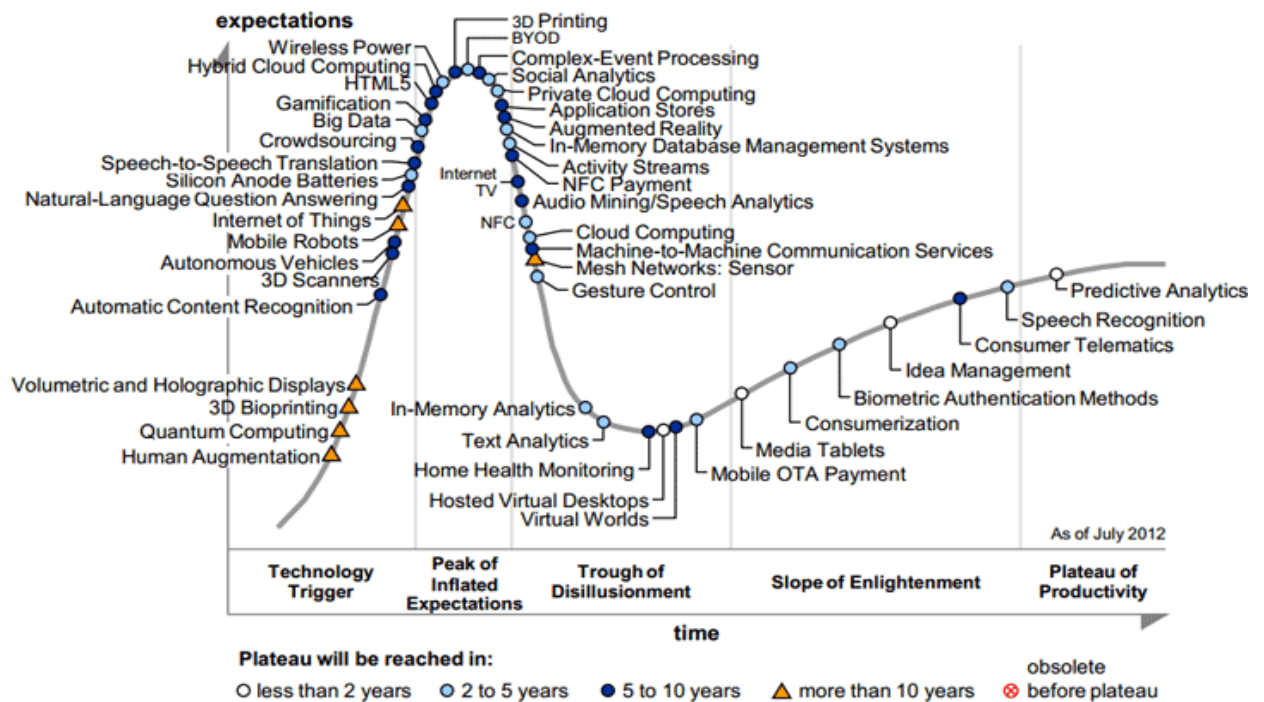


Рисунок 1.4 – Цикл зрілості технології на 2013 рік за даними Gartner, Inc. [17]

Кожна технологія у певний фіксований момент часу знаходиться на певній стадії (фазі) або переходить з однієї стадії (фазі) на іншу згідно з рисунком 1.4:

1) технологічний тригер (Technology Trigger) – період виникнення інновації, перші публікації про нову технологію;

2) пік надмірних очікувань (Peak of Inflated Expectations) – період очікування від нової технології революційних властивостей, часто не сумісних з майбутньою реальністю, етап різкого стрибка популярності та досягнення статусу предмету широкого обговорення в суспільстві (навіть людьми, які не спеціалізуються на даній технології);

3) позбавлення від ілюзій (Through of Disillusionment) – період виявлення недоліків технології та розчарування у надмірних та завищених очікуваннях від технології;

4) схил просвітлення (Slope of Enlightenment) – період усунення недоліків технології та повернення інтересу до неї; етап перших вагомих ефективних впроваджень технології у комерційних проектах;

5) плато продуктивності (Plateau of Productivity) – період зрілості та всезагального визнання технології, усвідомлюючи її суттєві значимі переваги та деякі недоліки, які ще не вдалося подолати.

Вважається, що не кожна технологія здатна пройти всі стадії, зокрема, багато технологій завершують своє існування (або активне використання) на етапі фази позбавлення від ілюзій.

Компанія Gartner ще у 1995 році запропонувала даний «Цикл зрілості технологій» (hype cycle) [17]. Дана крива описує фактично залежність очікувань від часу й притаманна життєвому циклу більшості технологій до їх становлення (якщо воно врешті відбувається). При цьому швидкість перебігу фаз та співвідношення тривалості перебігу фаз у кожній технології різна. Рисунок 1.4 відображає стадію на якій перебуває кожна технологія у певний фіксований момент часу. Очевидно, що рисунок 1.4 у певний момент залежить також і від

моментів виникнення зображених технологій. Як висновок, можна узагальнити, що залежність очікувань від часу для кожної технології подібна (рисунок 1.5).



Рисунок 1.5 – Крива циклу зрілості технології, що досягає період становлення

Враховуючи, що крива циклу зрілості у кожній технології (рисунок 1.5) може варіюватися в масштабі, кривизні та пропорціях, виникає висновок, що зображений на попередньому рисунку (рисунок 1.4) стан зрілості технологій у фіксований момент часу є ідеалізованим, тобто умовним, але, тим не менше, таке зображення стану зрілості технологій є схематично коректним та правильно відображає реальні стадії зрілості кожної зображеної технології у момент часу (без відображення конкретних абсолютних значень по осях координат).

1.9 Схематичне узагальнення структури системної методології моделювання кредитних ризиків, що відповідає подальшій постановці задач

Блок-схему етапів побудови моделей оцінювання кредитоспроможності зображено на рисунку 1.6 з наведенням відповідних критеріїв оптимальності. На рисунку 1.7 зображено структуру системної методології моделювання кредитних ризиків (узагальнення підрозділу 1.7) у розрізах етапів, методів, критеріїв оптимальності. Відповідно до структури ставиться постановка задач дослідження.



Рисунок 1.6 – Етапи побудови моделей оцінювання кредитоспроможності

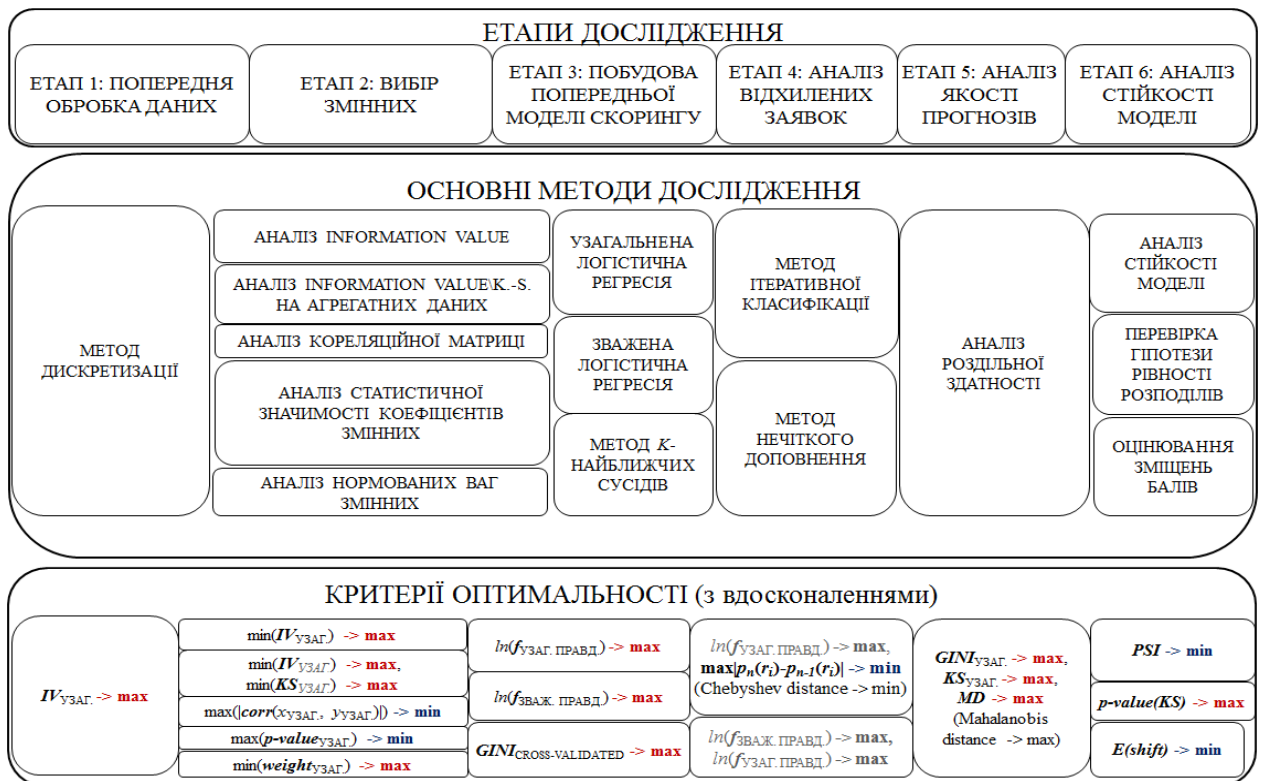


Рисунок 1.7 – Структура системної методології моделювання кредитних ризиків

1.10 Постановка задач дисертаційного дослідження

Мета дослідження – вдосконалення системної методології побудови моделей оцінювання кредитоспроможності, зокрема її цілісне узагальнення на всіх рівнях для випадку ймовірнісної цільової змінної, яке забезпечує суттєве вдосконалення аналізу відхилених заявок; розробка та вдосконалення методів і алгоритмів обчислення ключових показників, методів моделювання; розробка оригінального програмного продукту. Для досягнення мети потрібно вирішити такі завдання:

1. розробити формалізований метод дискретизації неперервних вхідних змінних на основі динамічного програмування у моделях оцінювання кредитоспроможності;

2. розробити метод розрахунку статистики Колмогорова-Смирнова і ваг категорій змінних та інформаційної статистики за умови відомого розподілу категорій та умовному розподілі цільової змінної;

3. розробити алгоритм обчислення рівнів статистичної значимості для оцінок моделі логістичної регресії шляхом інтегрування розкладу в ряд Тейлора;

4. вдосконалити модель логістичної регресії, методи її аналізу та оцінювання ваг категорій змінних, інформаційної статистики, індексу Джині, статистики Колмогорова-Смирнова для випадку ймовірнісної цільової змінної; вдосконалити метод ітеративної класифікації з метою подолання його головних недоліків;

5. розробити алгоритми нормування ваг змінних регресійної моделі з урахуванням варіації вхідних параметрів і прискорення процесу збіжності вектора коефіцієнтів логістичної регресії;

6. удосконалити модель на основі методу k-найближчих сусідів для задач машинного навчання моделей для оцінювання кредитоспроможності фізичних осіб;

7. розробити алгоритм розрахунку показника Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса засобами мов програмування четвертого покоління (4GL) на прикладі мови SQL, а також архітектуру системи підтримки прийняття рішень для побудови скорингових моделей і здійснити програмну реалізацію мовою програмування Visual C#.

1.11 Висновки до першого розділу

Продемонстровано, що методологія побудови скорингових моделей відповідає фундаментальним принципам та властивостям системної методології, а термінологія системного аналізу пояснює термінологію скорингу.

Наведено можливості застосування видів скорингу на етапах життєвого циклу роздрібного кредитування. Не обмежуючись ризик-менеджментом у сенсі управління кредитними ризиками та управлінням клієнтськими відносинами (Customer Relationship Management, CRM), вказано задачу прогнозування відтоку депозитних вкладів. Не обмежуючись діяльністю банків, наведено перелік інших областей застосування, описано суміжні області. У висновках варто додати, що областю застосування є також сегментація у галузі телекомунікацій, страхування. Сучасний стан циклу зрілості технологій теж доводить актуальність дослідження. Відкритим питанням досі залишається, зокрема, узагальнення методології для випадку ймовірнісної цільової змінної.

РОЗДІЛ 2

МЕТОДИ ФОРМУВАННЯ ВИБІРКИ ТА ПЕРЕТВОРЕННЯ ПАРАМЕТРІВ, МЕТОДИ АНАЛІЗУ ВХІДНИХ ХАРАКТЕРИСТИК ТА ВИБОРУ МНОЖИНИ НЕЗАЛЕЖНИХ ЗМІННИХ

Мета другого розділу полягає у розгляді двох груп тісно взаємопов'язаних методів: (1) методів формування вибірки для скорингового моделювання та подальшого перетворення параметрів при попередній обробці вхідних даних (в т.ч. методів WoE-перетворень, використання фіктивних змінних, дискретизації); (2) методів аналізу предикативної сили характеристик та вибору множини відносно незалежних змінних (в т.ч. кореляційного та факторного аналізу, аналізу предикативності, мультиваріаційного аналізу), а також у доведенні взаємозв'язку відстані Кульбака-Лейблера з показником інформаційної статистики та показником ваги значення змінної (WoE), а також у розробці власного методу дискретизації неперервних вхідних змінних на інтервали за допомогою динамічного програмування Беллмана.

Паралельно з економічними показниками для формування вибірки, такими як період вибірки, індикатор простроченої заборгованості та показовий період, продемонстровано застосування елементів теорії інформації та прикладної статистики при перетворенні параметрів, аналізі вхідних характеристик та виборі множини відносно незалежних змінних (в т.ч. описано застосування перевірки статистичних гіпотез на прикладі рівня статистичної значимості p -value з метою забезпечення мультиваріаційного аналізу).

Хоча розділ виділено окремо, термінологія та методологія даного розділу нерозривно взаємопов'язана у цілісній системі, зокрема, з підрозділом 3.3, що стосується валідації (тестування або оцінювання якості прогнозів) скорингових карт, та підрозділом 3.1, що присвячений власне методам побудови скорингових

моделей, тому деякі важливі методологічні аспекти (методи) описуються у відповідних наступних підрозділах, на які вказано конкретні посилання у тексті.

2.1 Особливості формування вибірки на прикладі аплікаційного скорингу. Період вибірки. Індикатор простроченої заборгованості. Показовий період

Класична теорія аплікаційного кредитного скорингу у напрямку побудови ймовірнісних моделей ідейно призначена для прогнозування ймовірності дефолту PD (див. підрозділ 1.1), де дефолт визначається як певний шаблон поведінки: наприклад, означення дефолту по платежах згідно з Базельською угодою II зазвичай передбачає прострочення на 90 днів (DPD) [8, 9]. Мають місце щонайменше три зауваження (на прикладі споживчого кредитування та індивідуального кредитування):

(1) для перевірки на наявність майбутнього досягнення заборгованості, наприклад, не менше 90 днів прострочення платежу (цільовий бінарний індикатор) зазвичай для зручності встановлюється фіксований показовий період (Performance Window, PW) – період «визрівання» прострочення або період спостереження або показове вікно і т.д. – від моменту видачі кредиту у місяцях [8, 9, 40];

(2) спостережуване досягнення певного рівня прострочення за показовий період може мати характер [8, 9]: (а) будь-коли за період спостереження (Ever) з можливістю подальшого спостережуваного пониження кількості днів прострочення; (б) на кінець періоду спостереження – «поточний» показник (Current);

(3) класичний показник 90+ днів прострочення заборгованості часто замінюється на більш ранні показники прострочення: 60 днів, 30 днів і т.д. [8, 9].

Виходячи з означень вище, для навчальної та тестової вибірок (для загальної початкової вибірки) вводяться означення нульової або «поганої»

(«bad» definition) та одиничної або «хорошої» («good» definition) угоди (або клієнта) [8, 9]. Прикладом означення нульової («поганої») угоди може бути «90+ днів прострочення на кінець 12-місячного показового періоду» («Current 90+ DPD after 12 months»), що часто називають показником NPL (net present loss), або «будь-коли 30+ днів прострочення за 6 місяців» («Ever 30+ within 6 months») і т.д. Прикладом найбільш популярного означення одиничних («хороших») угод є означення виданих договорів, що не підлягають означеному критерію нульової («поганої») угоди, тобто означаються по принципу простого логічного заперечення: «Good = NOT Bad» (тобто всі інші видані договори). Таким чином означаються бінарні класи угод (або клієнтів), що утворюють повну групу подій. Також іноді означається «сіра зона» («gray zone»), що зазвичай повністю виключається з вибірки, наприклад, у результаті такого розбиття:

- (1) одиничний клас («good»): «Ever 0 DPD within 6 months»;
- (2) невизначена «сіра зона» («gray zone»): «Ever 1-29 DPD within 6 months»;
- (3) нульовий клас («bad»): «Ever 30+ DPD within 6 months».

Недоліком виключення невизначеної зони є деяке спотворення ймовірності дефолту, оскільки вона буде оцінена в процесі навчання моделі лише з урахуванням двох класів з трьох попередньо означених, тобто без врахування «сірої зони».

Також окрім трьох класичних показників «не менше ніж 90 (або 60 або 30) днів прострочення» часто використовуються зміщені на одиницю: «не менше ніж 91 (або 61 або 31) день прострочення». Останній підхід має перевагу у сенсі портфельного аналізу, коли кредитний портфель зручно розбивати на кратні тридцяти нестрогі межі: (0) 0 днів прострочення; (1) 1-30 днів прострочення; (2) 31-60 днів прострочення; (3) 61-90 днів прострочення; (4) 91-120 днів прострочення; (5) 121-180 днів прострочення (для зручності часто пропускають границю в 150 днів); (6) 180+ днів прострочення (важливе зауваження: у даному випадку символ плюса вже інтерпретується як «строго більше», а у попередніх

позначеннях використовувалось як «не менше» – у подібних випадках необхідно оговорювати заздалегідь умовні позначення).

Період вибірки (Sample Window, SW; Observation Window, OW; Application Window, AW) – це період, що відповідає моментам фіксування стану значення вхідних змінних (предикторів) – «точкам відліку» для показового періоду [8, 9, 40]; це період видачі кредитів (у випадку аплікаційного скорингу для споживчого або індивідуального кредитування), що розглядається у вибірці, де предиктори – це, зазвичай, анкетні соціально-демографічні дані про клієнта та його попередня кредитна історія або дані перевірок по бюро кредитних історій (при наявності).

У випадку аплікаційного скорингу для кредитних карт період вибірки (період «точок відліку» для показового періоду) зазвичай відповідає даті першій транзакції – даті виникнення першої заборгованості. Також можуть вводитися обмеження щодо розміру утилізації (відношення суми транзакції до встановленого кредитного ліміту) для включення у вибірку (для означення «хороших» та «поганих» клієнтів).

Щодо методів визначення оптимального індикатора DPD для «поганого» клієнта (або угоди), то найчастіше використовуються такі методи [8, 9]:

- (1) метод консенсусу;
- (2) аналітичні методи:
 - (2.1) метод аналізу міграцій прострочень (roll rate analysis);
 - (2.2) порівняльний аналіз поточної заборгованості з найгіршою.

Метод консенсусу [8, 9] є суто логічним методом експертної оцінки з залученням, наприклад, спеціалістів з департаменту контролю ризиків, з відділу маркетингу та з операційного відділу для досягнення спільного консенсусу по найкращому означенню «поганого» клієнта (угоди), виходячи з досвіду, оперативної діяльності та результатів аналізу розподілів (наприклад, достатньої

кількості означених «поганих» клієнтів або угод, або задовільного співвідношення «хороших» та «поганих» елементів вибірки).

Перший аналітичний метод – метод аналізу міграцій прострочень (roll rate analysis) – полягає у підрахунку умовних ймовірностей переходу з одного рівня прострочення (статусу) в інший. Зазвичай, аналіз міграцій прострочень проводиться для переходів типу зі статусу «будь-коли за x попередніх місяців» («Ever within previous x months») у статус з «будь-коли за x наступних місяців» («Ever within next x months»), тобто всюди по найгіршому (максимальному) простроченню [8, 9]. При цьому, використовуючи фіксовану кількість можливих статусів, сума умовних ймовірностей переходу в інші статуси, включаючи також ймовірність зберігання статусу, для кожного статусу дорівнює 100%, тобто має місце аналіз класичної квадратної матриці міграцій:

$$R(x, d) = \begin{pmatrix} R_{11}(x, d) & R_{12}(x, d) & \cdots & R_{1n}(x, d) \\ R_{21}(x, d) & R_{22}(x, d) & \cdots & R_{2n}(x, d) \\ \vdots & \vdots & \ddots & \vdots \\ R_{n1}(x, d) & R_{n2}(x, d) & \cdots & R_{nn}(x, d) \end{pmatrix} = (R_{ij}(x, d))_{i=1, j=1}^{n, n},$$

де, у даному випадку, $R_{ij}(x, d)$ – фактична ймовірність переходу (roll rate) з минулого i -того рівня прострочення у майбутній j -тий рівень прострочення при використанні минулих та майбутніх спостережень в x місяців відносно певної дати d ; n – кількість станів прострочення, що розглядаються.

Визначальні властивості матриці міграцій описуються системою умов:

$$\begin{cases} \forall i \in \{1, \dots, n\} \forall j \in \{1, \dots, n\} : R_{ij}(x, d) \in [0; 1], \\ \forall i \in \{1, \dots, n\} : \sum_{j=1}^n R_{ij}(x, d) = 1. \end{cases}$$

тобто сума ймовірностей про кожному рядку дорівнює одиниці (100%).

Нижче наводиться приклад представлення матриці міграцій прострочень у графічному вигляді за допомогою нормованої діаграми (рисунок 2.1) [8, 9]. Максимальне значення днів прострочки за 12 місяців округлюється тільки вниз до найближчого числа кратного 30 дням (по принципу «менше або дорівнює»), розглядається рівень кредитних рахунків (тобто рівень угод). Значення кожного підсумкового стовпця дорівнює 100%. Умовне позначення «Curr/x day» означає відносно непрострочену будь-коли заборгованість за діапазон, що розглядається, тобто *DPD менше 30 днів* всюди на діапазоні. При цьому відносно непрострочена заборгованість часто називається «поточною» («current») – це означення надалі буде використовуватися рідко з метою уникнення недорозумінь зі звичайним класичним загальним значенням слова «поточна», що, наприклад, використовується у наступному методі для порівняння поточного прострочення.

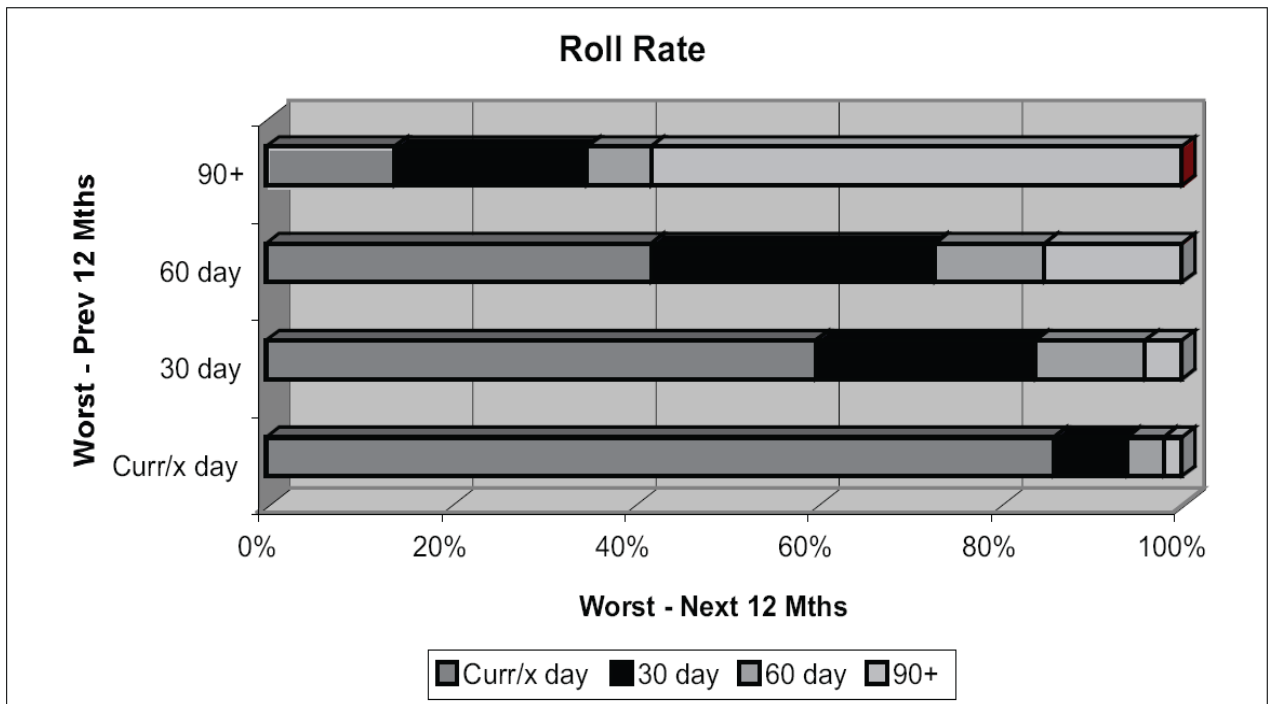


Рисунок 2.1 – Приклад діаграми аналізу міграцій прострочень (roll rate chart)

[8, 9]

У даному випадку перший статус позначено як «Curr/x day», другий – як «30 day», третій – як «60 day», четвертий – як «90+».

Прикладом переходу для деякого рахунку з статусу «90+» в «60 day» для ситуації, що відповідає рисунку 2.1, є таблиця 2.1 [8, 9]. Стан прострочення представлено у повних місяцях. Описана ситуація по деякому рахунку в таблиці 2.1 відповідає поняттю «перетікання» назад (roll back або roll backward).

Таблиця 2.1 – Приклад історії заборгованості для деякого рахунку [8, 9]

Попередні 12 місяців												
Місяць	1	2	3	4	5	6	7	8	9	10	11	12
Прострочення (м.)	0	0	1	1	0	0	0	1	2	3	0	0
Наступні 12 місяців												
Місяць	13	14	15	16	17	18	19	20	21	22	23	24
Прострочення (м.)	0	0	1	2	0	0	0	1	0	1	0	0

Метою методу аналізу міграцій прострочень є визначення «точки неповернення» («point of no return») – мінімального рівня прострочення заборгованості (delinquency), при якому переважна більшість боргів стають безнадійними, тобто відбувається переважне «перетікання» вперед (roll forward) [8, 9]. Багато світових досліджень [8, 9] свідчить про те, що досягнення 90+ днів прострочення будь-коли за 12 місяців має ймовірність переходу у не нижчі статуси прострочення на майбутньому 12-місячному діапазоні рівну близько 60%-70% (рисунок 2.1).

Також бажано поваріювати значення довжини досліджуваних діапазонів (змінна x) поки оптимальне значення показового періоду (Performance Window) ще не обрано.

У подальшому вже при обраному показовому періоді метод аналізу міграцій статусів прострочення може бути використаний для комплексного

підтвердження коректності означення дефолту, зокрема означення критичного індикатора днів прострочення (DPD) та періоду прогнозу (Performance Window). При цьому тоді період прогнозу відноситься скоріше до першого діапазону, який надалі підлягає міграціям на другому діапазоні.

Метод порівняння поточного прострочення з найгіршим будь-коли (порівняльний аналіз поточної заборгованості з найгіршою) по суті схожий на метод аналізу міграцій прострочень. Суттєвими перевагами даного ефективного методу є: (1) простота; (2) відсутність параметру довжини досліджуваних діапазонів (параметру x), бо використовується просто повністю вся доступна історія по рахунку або клієнту на момент аналізу, тому метод дає змогу безпосередньо оцінити критичне значення днів прострочення (індикатор простроченої заборгованості) без будь-якого явного впливу чи залучення показового періоду (Performance Window); (3) за момент аналізу зручно брати поточну дату, що також найбільш точно відображає поточну ситуацію (аналогічно хорошій стійкості моделі – підрозділи 3.3.2, 3.3.2.1 та 3.3.2.3).

Суть методу порівняння поточного прострочення з найгіршим будь-коли полягає у розкладі множини угод (клієнтів) для кожного статусу за всю історію, тобто згрупованих по значенню найгіршого прострочення будь-коли, на поточні стани. Тобто відбувається представлення умовних розподілів поточних статусів для історичних статусів, обчислених по принципу максимального прострочення.

Таблиця 2.2 [8, 9] відображає приклад застосування методу порівняльного аналізу простроченої заборгованості з найгіршою за історію. При цьому відносно непрострочена будь-коли заборгованість у сенсі днів прострочення по договору або по всіх договорах клієнта (тобто прострочення, що жодного разу *не досягало 30 днів*) також зображується: для відносно непростроченої заборгованості будь-коли за означенням очевидно, що вона на 100% відносно непрострочена й на момент аналізу (як логічний наслідок в один бік, тобто поняття «будь-коли» включає в аналіз також поточний момент).

Таблиця 2.2 – Порівняння днів поточної заборгованості з максимальними днями прострочення за історію по кредитних рахунках (приклад) [8, 9]

		Найгірша заборгованість по рахунку за історію (Ever)					
		< 30 днів	30 днів	60 днів	90 днів	120 днів	Списання
Поточна заборгованість (на момент аналізу)	< 30 днів	100%	84%	53%	16%	7%	
	30 днів		12%	28%	10%	8%	
	60 днів		4%	11%	14%	10%	
	90 днів			8%	44%	13%	
	120 днів				16%	62%	
	Списання						100%

У прикладі, що відповідає таблиці 2.2, для всіх кредитних рахунків, що коли-небудь досягали прострочення на 30 днів, переважна більшість – 84% – на момент аналізу не мають значимої простроченої заборгованості, тобто не мають *щонайменше 30 днів* прострочення. Також 60% всіх рахунків (16% + 44%), які досягали за свою історію 90 днів прострочення, переходять на гірший рівень прострочення або залишаються на тому ж самому. Цей факт для прострочення на 90 днів повністю узгоджується зі спостереженням, наведеним при розгляді попереднього методу.

Англійською мовою метод порівняння поточної заборгованості з найгіршою називається «Current versus Worst Delinquency Comparison» [9].

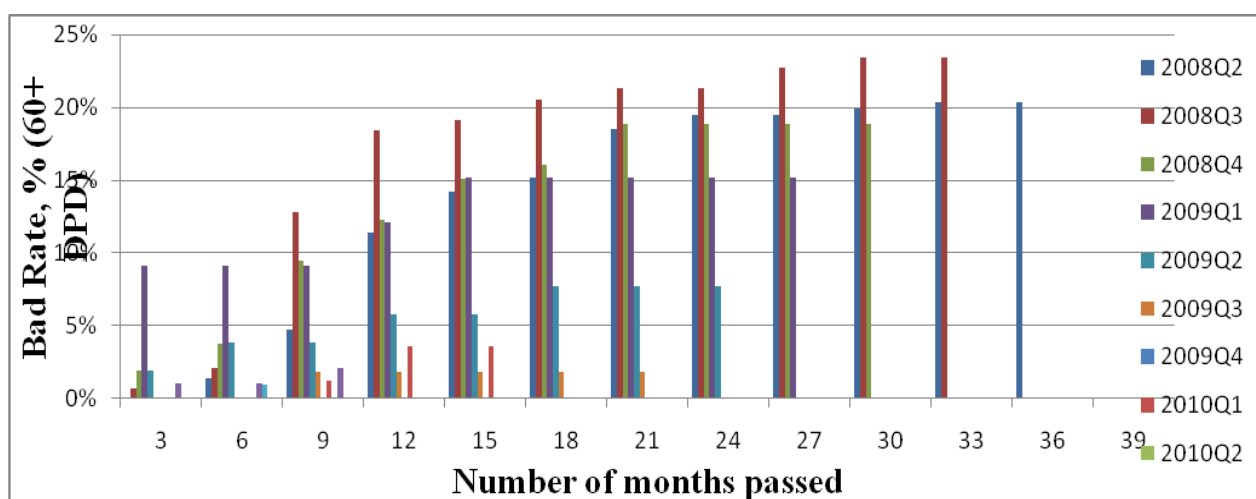
Щодо методу визначення оптимального показового періоду (Performance Window, PW), то зазвичай використовується звіт аналізу поколінь (vintage analysis), який ще називається звітом динаміки прострочення (dynamic delinquency).

Приклад звіту аналізу поколінь для суб'єктів малого та середнього бізнесу наведено відносно дати першого кредитування наведено в таблиці 2.3.

Таблиця 2.3 – Аналіз поколінь (vintage analysis) суб'єктів малого та середнього бізнесу по індикатору «Ever 60+ DPD»
(станом на II квартал 2011 – 2011Q2)

Квартал видачі	Показовий період (Performance Window) в місяцях												
	3	6	9	12	15	18	21	24	27	30	33	36	39
2008Q2	0%	1%	5%	11%	14%	15%	18%	19%	19%	20%	20%	20%	
2008Q3	1%	2%	13%	18%	19%	21%	21%	21%	23%	23%	23%		
2008Q4	2%	4%	9%	12%	15%	16%	19%	19%	19%	19%			
2009Q1	9%	9%	9%	12%	15%	15%	15%	15%	15%				
2009Q2	2%	4%	4%	6%	6%	8%	8%	8%					
2009Q3	0%	0%	2%	2%	2%	2%	2%						
2009Q4	0%	0%	0%	0%	0%	0%							
2010Q1	0%	0%	1%	4%	4%								
2010Q2	0%	0%	0%	0%									
2010Q3	1%	1%	2%										
2010Q4	0%	1%											
2011Q1	0%												
2011Q2													

У графічному вигляді таблицю 2.3 можна представити у вигляді рисунку 2.2.



Рисунком 2.2 – Аналіз поколінь (vintage analysis) суб'єктів малого та середнього бізнесу по індикатору «Ever 60+ DPD»
(станом на II квартал 2011 – 2011Q2)

При означенні оптимального значення показового періоду для встановленого раніше критичного значення днів прострочення (індикатора простроченої заборгованості) вивчається аналіз поколінь з метою визначення такої мінімальної кількості місяців, при якій починає глобально (в ідеалі) стабілізуватися графік аналізу поколінь. При цьому зручно будувати один графік для аналізу об'єднаних поколінь (в ідеалі для результату об'єднання – одного великого покоління, й аналізувати його до такого місяця включно, де у всіх елементів вибірки вже відомий результат). Також можна аналізувати набір гістограм (рисунок 2.2), кожна наступна з яких доступна на зменшену на один період аналізу множину показових періодів. У прикладі на рисунку 2.2 та у відповідній таблиці 2.3 «плато визрівання» для індикатора 60+ (Ever) досягається у відносно далекому майбутньому для суб'єктів з довгостроковими кредитами, тобто, враховуючи вимоги до швидкості приросту, визначається стабілізаційне «плато», яке починається з 21-го місяця спостереження, тому оптимальний показовий період у наведеному прикладі обирається рівним 21-му місяцю. У даному прикладі було встановлено обмеження на пошук точки розбиття, починаючи з якої приріст долі «поганих» клієнтів (аналогічно першій похідній) для всіх поколінь не перевищує 2% (для прикладу) – див. таблицю 2.4. На практиці часто використовуються значення 12 [8, 9] та 18 місяців у якості оптимального показового періоду. У наведеному прикладі при використанні обмеження рівним 3% замість 2% оптимальним показовим періодом буде період 12 місяців (таблиця 2.4). Таким чином можна означити більш нестроге «плато», яке може використовуватись при побудові скорингової моделі з можливістю включення більш свіжих даних у період вибірки (Observation Window). Можна зробити висновок, що при виборі оптимального показового періоду необхідно налаштовувати метод оснований на аналізі поколінь таким чином, щоб мати змогу надалі означити такий період вибірки, що може містити відносно нові дані.

Таблиця 2.4 – Максимальна емпірична перша похідна по всіх поколіннях в залежності від показового періоду (за даними таблиці 2.3)

	<i>Показовий період (Performance Window) в місяцях</i>												
	3	6	9	12	15	18	21	24	27	30	33	36	39
Макс. перша похідна	2%	11%	6%	3%	2%	3%	1%	2%	1%	0%	0%	-	-
Макс. по першому розбиттю	11%						2%						
Макс. по другому розбиттю	11%			3%									

Період вибірки (Observation Window), як було вище зазначено, відповідає періоду видачі кредитних продуктів (на прикладі споживчого або індивідуального кредитування) або періоду здійснення першої утилізації кредитного ліміту для кредитних карт, або ж періоду фіксування моменту стану існуючої кредитної угоди або клієнта в кредитному портфелі (для поведінкових скорингових моделей, скорингу попереднього збору заборгованості, скорингу збору заборгованості і т.д.). Для моделей типу поведінкового скорингу, скорингу попереднього збору заборгованості, скорингу збору заборгованості та всіх інших моделей, що призначені для оцінювання існуючих угод або клієнтів, період вибірки може спричиняти коректне включення угоди або клієнта у вибірку декілька разів, оскільки важливий саме стан («стадія») угоди або клієнта, що відповідає елементу вибірки.

Методами визначення оптимального періоду вибірки є насамперед аналіз стійкості (підрозділи 3.3.2, 3.3.2.1 та 3.3.2.3) всіх включених часових періодів, наприклад, відносно всієї вибірки, тобто аналіз незалежності розподілів змінних від часу, що дозволяє, наприклад, виключити старі періоди, які мають інші розподіли, та аналіз варіації долі «поганих» клієнтів (bad rate) відносно періодів,

яка повинна бути досить низькою. Варіацію можна оцінювати, наприклад, використовуючи класичний коефіцієнт варіації (variation coefficient):

$$V(x) = \frac{\sigma_x}{\bar{x}},$$

де σ_x – середньоквадратичне відхилення вибірки, \bar{x} – середнє значення по вибірці, тобто:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$\sigma_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2},$$

де n – розмір вибірки, x_i – елемент вибірки.

У даному випадку елементи вибірки – це числові долі «поганих» клієнтів, а розмір вибірки – кількість періодів (зазвичай місяців), що складають період вибірки.

Варіювання вважається слабким, якщо $V(x) < 10\%$, якщо $V(x) \in (10\%; 25\%]$, то середнім і значним при $V(x) > 25\%$ [41].

Очевидно, що кінець періоду вибірки не може перевищувати дату аналізу (наприклад, сьогоднішню) зменшену на значення показового періоду. Іноді використовують nereкомендований прийом включення ранніх «поганих» клієнтів (угод) з проміжку після закінчення періоду вибірки й до дати аналізу. Такий прийом доповнення вибірки нульовими клієнтами (угодами) не рекомендується у зв'язку зі штучним збільшенням долі нульових (по цільовій змінній) елементів вибірки, що буде відобразитися на прогностичних значеннях.

Щодо всіх наведених цифр у даному підрозділі, то, враховуючи різноманітність кредитних продуктів, регіональні відмінності, часові відмінності, впливи пов'язані з системою збору та прийняття рішень та низку багатьох інших факторів, при застосуванні у різних інших ситуаціях на різних вхідних даних вказані цифри також можуть суттєво відрізнятися.

У рамках дисертаційного дослідження використовується вибірка для аплікаційного скорингу для споживчого кредитування з параметрами зображеними на рисунку 2.3.

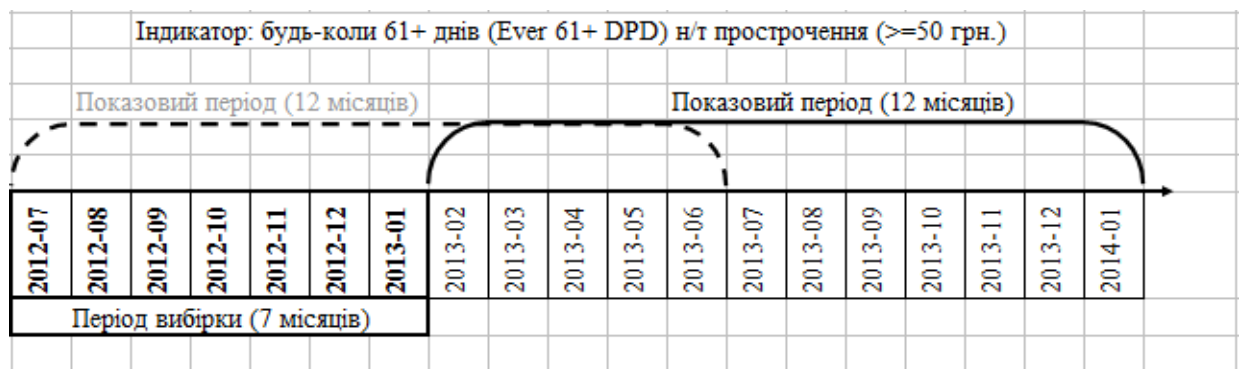


Рисунок 2.3 – Характеристики вибірки для побудови аплікаційного скорингу для споживчого кредитування

Щодо рисунку 2.3, то для означення «bad / good» також використовується поняття нетехнічного (н/т) прострочення, тобто має місце мінімальне обмеження на значимий прострочений баланс – тут не менше 50 грн., тобто під час проведення аналізу DPD вважається нульовим, якщо прострочений баланс менше 50 грн.

Зазвичай загальна сформована вибірка розбивається випадковим чином на навчальну та тестову вибірки у співвідношенні «80% / 20%» або «70% / 30%».

2.2 Методи використання категоріальних змінних. Концепція ваги категорії змінної та інформаційної статистики як показника предикативності

Категоріальні змінні часто потребують перетворення у числовий формат. Нагальна необхідність коректного перетворення текстових значень з фіксованого набору, тобто категорій, у числовий формат обумовлюється потребою забезпечення можливості використання категоріальних змінних у моделях типу логістичної регресії [8–10, 32–34] або інших типах моделей, де в кінці кінців використовуються числові вхідні змінні. Для прикладу, моделі типу дерев рішень можуть оперувати з категоріями безпосередньо, тому не потребують такого перетворення.

Категоріальні (categorical) змінні також називають факторами (factor), змінними класу (class) або номінальними (nominal) змінними [32–34]. Ключовою особливістю номінальних змінних є те, що їх значення між собою неупорядковані [34], тобто операції порівняння між окремими значеннями номінальної змінної невизначені [34]. У якості значень категоріальної змінної можуть використовуватися числові значення зі скінченного набору, де введена впорядкованість числових значень не несе в собі ніякого принципового змісту. Кожне унікальне значення (категорія) представляє собою логічно окремий зміст. Розглядаються два методи числового оперування категоріальними змінними.

Перший метод називається методом фіктивних змінних (dummy variables, design variables, indicator variables) [8–10, 32, 33]. Суть методу полягає у розбитті кожної категоріальної змінної на множину бінарних змінних (таблиця 2.5). При цьому одна з категорій зазвичай не кодується [33] з метою забезпечення функціональної незалежності множини створених бінарних змінних. Таким чином, категоріальна змінна перетворюється в набір $k - 1$ бінарних змінних, де k – кількість взаємовиключних категорій (унікальних значень) даної категоріальної змінної, які утворюють повну групу подій. Ознака наявності категорії, що не кодується («нульової»), може бути обчислена таким чином:

$$D_0 = \neg(D_1 \vee D_2 \vee \dots \vee D_{k-1}).$$

Таблиця 2.5 – Приклад декомпозиції змінної методом фіктивних змінних

Змінна сімейного стану	Значення $k - 1$ фіктивних змінних		
	D_1	D_2	D_3
Значення номінальної змінної (категорія)	D_1	D_2	D_3
<i>Неодружений/Незаміжня</i> (включаючи відсутність наявних даних)	0	0	0
<i>Одружений/Заміжня</i>	1	0	0
<i>Вдівець/Вдова</i>	0	1	0
<i>Розвідний / Розвідна</i>	0	0	1

При використанні даного методу порядок створених бінарних змінних неважливий. Серед множини недоліків даного методу можна виділити зокрема такі: (а) суттєве збільшення кількості змінних при заміщенні кожної категоріальної змінної на множину фіктивних змінних; (б) метод не враховує ні характеристики розподілу категорій, ні умовні ймовірності цільової змінної (або будь-який взаємозв'язок з цільовою змінною).

Розглядаючи методи приведення категоріальних змінних до числового формату, необхідно зазначити, що найбільшої популярності [8, 9, 11, 13, 42–45] набула концепція ваг категорій змінної (Weight of Evidence) та інформаційної статистики (Information Value).

Метод використання ваг категорій змінної оснований на використанні сумісного розподілу категоріальної вхідної змінної та бінарної цільової змінної, а власне вага категорій змінної (вага атрибуту змінної) – це логарифм відношення ймовірностей окремих розподілів для категорії [8, 9, 11, 13, 42–45]:

$$WoE_i = \ln \left(\frac{g_i}{b_i} \right),$$

де g_i – доля числа *одиничних* (*good*) клієнтів, що відповідають i -й категорії, по відношенню до всіх *одиничних* клієнтів у навчальній вибірці, аналогічно, показник b_i – доля числа *нульових* (*bad*) клієнтів, що відповідають i -й категорії, по відношенню до всіх *нульових* клієнтів у навчальній вибірці; WoE_i – вага категорії змінної, що дорівнює натуральному логарифму відношення g_i до b_i .

Таким чином, мають місце два дискретні розподіли відносно категорій:

$$\sum_{i=1}^k g_i = 1,$$

$$\sum_{i=1}^k b_i = 1.$$

Як зазначено вище, через абсолютні числа *одиничних* (*good*) клієнтів G_i , що відповідають i -й категорії, через абсолютні числа *нульових* (*bad*) клієнтів B_i , що відповідають i -й категорії, ймовірності g_i та b_i виражаються таким чином:

$$g_i = \frac{G_i}{G},$$

$$b_i = \frac{B_i}{B},$$

де введено позначення загальної кількості *одиничних* (G) та загальної кількості *нульових* (B) елементів вибірки:

$$G = \sum_{i=1}^k G_i,$$

$$B = \sum_{i=1}^k B_i.$$

Таким чином даний метод дозволяє ставити у відповідність категоріальним значенням числові значення ваг категорій змінної. При використанні моделей типу логістичної регресії у даному методі для кожної категоріальної змінної використовуються ваги категорій змінної замість відповідних категорій.

Інформаційна статистика (Information Value) – це найбільш популярний у методології побудови скорингових моделей [5, 8, 9, 11, 13, 42–45] показник сили взаємозв'язку вхідної категоріальної змінної з цільовою бінарною змінною:

$$\text{Information statistic} = \text{divergence} = IV = \sum_{i=1}^k (g_i - b_i) \ln \left(\frac{g_i}{b_i} \right) = \sum_{i=1}^k (g_i - b_i) W_o E_i .$$

Область значень інформаційної статистики відповідає інтервалу $[0; +\infty)$, оскільки виконується просте твердження, що також доводить «симетричність» відносно двох розподілів:

$$\forall a \in (0; 1] \forall b \in (0; 1]: (a - b) \ln \left(\frac{a}{b} \right) = (b - a) \ln \left(\frac{b}{a} \right) \geq 0 .$$

Поняття показника інформативності (інформаційної статистики) є класичним поняттям теорії інформації [8, 9]. Судження про прогностичну здатність категоріальної змінної по обчисленому значенню інформаційної статистики, які використовуються при розробці системи підтримки прийняття рішень у рамках дисертаційної роботи, наводяться в таблиці 2.6.

Також альтернативна шкала прогностичної здатності, що часто використовується [8, 9], наведена в таблиці 2.7.

У таблиці 2.8 наводиться суть методу використання ваг категорій змінних та приклад розрахунку інформаційної статистики для змінної з таблиці 2.5.

Таблиця 2.6 – Прогностична сила на основі інформаційної статистики

Інформаційна статистика (Information Value)	Прогностична сила (Predictive Power)
<0,03	Мізерна (Poor)
[0,03; 0,10)	Слабка (Weak)
[0,10; 0,30)	Середня (Average)
[0,30; 0,50)	Сильна (Strong)
$\geq 0,50$	Дуже сильна (Very strong)

Таблиця 2.7 – Приклад іншої шкали прогностичної здатності [8, 9]

Інформаційна статистика (Information Value)	Прогностична сила (Predictive Power)
<0,02	Відсутня (Unpredictive)
[0,02; 0,10)	Слабка (Weak)
[0,10; 0,30)	Середня (Medium)
$\geq 0,30$	Сильна (Strong)

Таблиця 2.8 – Приклад застосування методу WoE для підрахунку IV

Змінна сімейного стану	Аналіз розподілу цільової змінної відносно категорій					
	Bads	Goods	Bads %	Goods %	WoE	delta IV
<i>Неодружений/Незаміжня (включаючи відсутність наявних даних)</i>	3315	32080	35,71	21,46	-0,5093	0,0726
<i>Одружений/Заміжня</i>	4893	99918	52,71	66,84	0,2375	0,0336
<i>Вдівець/Вдова</i>	249	4777	2,68	3,20	0,1751	0,0009
<i>Розвідний / Розвідна</i>	826	12718	8,90	8,51	-0,0449	0,0002
Підсумки (Total):	9283	149493	100	100	0 (ln 1)	IV=0,1072

2.3 Доведення взаємозв'язку відстані Кульбака-Лейблера з показником інформаційної статистики та показником ваги значення змінної

Відстань Кульбака-Лейблера (Kullback–Leibler divergence) є спрямованою дивергенцією (directed divergence) для двох розподілів – наближеного розподілу $Q(i)$, що припускається та перевіряється, відносно «істинного» розподілу $P(i)$, який постульований априорі [42, 46–49]:

$$D_{KL}(P \parallel Q) = \sum_i \ln \left(\frac{P(i)}{Q(i)} \right) P(i),$$

Очевидно, при цьому виконуються такі рівності:

$$\sum_i P(i) = 1,$$

$$\sum_i Q(i) = 1.$$

Розподіл $Q(i)$ зазвичай представляє собою модель або певну апроксимацію «істинного» розподілу даних $P(i)$, де тоді наведена відстань Кульбака-Лейблера інтерпретується як величина втрат інформації при заміні «істинного» розподілу $P(i)$ на модельний розподіл $Q(i)$. Згідно з лемою Шеннона, відстань є різницею фактичної середньої довжини коду та ентропії (мінімальної середньої довжини коду). Для щільностей двох розподілів – «істинної» $p(x)$ та «наближеної» $q(x)$ – відстань Кульбака-Лейблера обчислюється таким чином [42, 47, 48, 50]:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{+\infty} \ln \left(\frac{p(x)}{q(x)} \right) p(x) dx.$$

У більш загальному випадку при виконанні необхідних умов, відстань Кульбака-Лейблера можна переписати з використанням похідної Радона-Нікодима (Radon-Nikodym derivative) [42, 51], використовуючи кумулятивні функції двох розподілів ($F_P(x)$ та $F_Q(x)$):

$$D_{KL}(P \parallel Q) = \int_{x \in X} \ln \left(\frac{dF_P(x)}{dF_Q(x)} \right) dF_P(x) = \int_{x \in X} \ln \left(\frac{dF_P(x)}{dF_Q(x)} \right) \frac{dF_P(x)}{dF_Q(x)} dF_Q(x).$$

Доведення узагальненого виявленого взаємозв'язку та імплементація показників методології скорингу через обчислення функцій від відстаней Кульбака-Лейблера описується наступними формулами:

$$IV = \sum_i \ln \left(\frac{g_i}{b_i} \right) g_i - \sum_i \ln \left(\frac{g_i}{b_i} \right) b_i = \sum_i \ln \left(\frac{g_i}{b_i} \right) g_i + \sum_i \ln \left(\frac{b_i}{g_i} \right) b_i,$$

$$IV = D_{KL}(\{g_i\} \parallel \{b_i\}) + D_{KL}(\{b_i\} \parallel \{g_i\}) = D_{KL}(\mathbf{g} \parallel \mathbf{b}) + D_{KL}(\mathbf{b} \parallel \mathbf{g}),$$

$$WoE_i = \ln \left(\frac{g_i}{b_i} \right) = \ln \left(\frac{g(x_i)}{b(x_i)} \right) = \ln \left(\frac{d^- F_G(x_i)}{d^- F_B(x_i)} \right),$$

$$WoE_i = \frac{d^- \left(\int_{x \in X} \ln \left(\frac{d^- F_G(x_i)}{d^- F_B(x_i)} \right) d^- F_G(x_i) \right)}{d^- F_G(x_i)} = \frac{d^- D_{KL}(\mathbf{g} \parallel \mathbf{b})}{d^- F_G(x_i)},$$

$$WoE_i = - \frac{d^- \left(\int_{x \in X} \ln \left(\frac{d^- F_B(x_i)}{d^- F_G(x_i)} \right) d^- F_B(x_i) \right)}{d^- F_B(x_i)} = - \frac{d^- D_{KL}(\mathbf{b} \parallel \mathbf{g})}{d^- F_B(x_i)}.$$

У доведенні для WoE використовуються штучно означені для дискретного категоріального розподілу щільності $g(\bullet)$, $b(\bullet)$ і функції розподілу $F_G(\bullet)$, $F_B(\bullet)$.

2.4 Категоризація змінних. Розробка методу оптимальної дискретизації неперервних змінних за допомогою динамічного програмування Беллмана

Дискретизація (discretization) або категоризація неперервних змінних (categorization of continuous variables), яку ще часто називають бінаризацією (binning), зазвичай використовується по двох причинах [10]: 1) простіша інтерпретація статистичних показників категоріальних змінних у вигляді таблиць (в т.ч. таблиць частот) у порівнянні з класичними статистичними показниками (в т.ч. описовою статистикою) для неперервних числових змінних; 2) головне: традиційні методи моделювання, такі як лінійна або логістична регресії, не враховують нелінійні взаємозв'язки – нелінійності (nonlinearities) – для даних, якщо ці відомості попередньо не перетворені до включення у лінійну частину моделі (для логістичної регресії це логіт – лінійна функція від множини вхідних змінних, яка надалі підлягає нелінійному логістичному перетворенню – застосуванню сигмоїдальної функції). Після проведення дискретизації, всім створеним інтервалам значень, як і у випадку категоріальних змінних, зазвичай ставиться у відповідність числове WoE-перетворення, тренд якого відносно порядкового номеру категорій (на відміну класичних категоріальних змінних, категорії яких не впорядковані) є нелінійним у загальному випадку. Таким чином бінаризація допомагає аналітикам побудувати скорингову модель для роботи з нелінійними даними [10]. Головними недоліками застосування процесу категоризації є [10]: (1) висока трудоемність процесу; (2) висока ресурсоемність процесу; (3) можлива суб'єктивність налаштувань процесу; (4) у процесі бінаризації може втрачатися інформація у результаті дискретизації; (5) процес вимагає багато часу та засобів для впровадження. Також слід додати, що у результаті дискретизації втрачається властивість неперервності відносно оригінальних неперервних значень числових вхідних змінних.

Методологія кредитного скорингу та побудови скорингових моделей в цілому є досить нечіткою та слабо формалізованою щодо питання наведення конкретних методів дискретизації неперервних змінних.

Багато авторитетних джерел наводять набір основоположних принципів (критеріїв) оптимальності результату дискретизації неперервних змінних [8, 9]:

1) бажано, щоб пропущені значення (пусті значення, порожні значення, NULLs) були виділені в окрему категорію, якщо їх не можна логічно точно поєднати з іншою категорією;

2) для забезпечення значимості категорій повинно застосовуватись правило на мінімальну долю від навчальної вибірки для інтервалів непустих значень типу «мінімум 5% вибірки у кожному інтервалі»;

3) групи з нульовою кількістю одиничних або нульових елементів вибірки не допускаються (така ситуація може виникати у зв'язку з малою долею вибірки в інтервалі, навіть при виконанні принципу 2, коли неможливо обчислити WoE);

4) долі нульового класу (bad rate) у сусідніх групах (інтервалах) i , що еквівалентно (як буде показано в підрозділі 2.6), ваги категорій змінної (WoE) у сусідніх групах суттєво відрізняються, тобто дискретизація виконана з ціллю максимальної диференціації між нульовими та одиничними елементами вибірки. Це i є однією з головних цілей категоризації: виявити та виділити атрибути з високим рівнем диференціації. Частоти створених категорій, абсолютні значення WoE та, відповідно, різниця між вагами категорій змінної є ключовими для встановлення диференціації. Чим більша різниця по долі нульового класу та, відповідно, різниця значень WoE для сусідніх категорій змінної, тим вища прогностична сила (предикативна сила або предикативність) для даної характеристики (див. підрозділ 2.2 та підрозділ 2.6).

Ще одним опціональним критерієм оптимальності процедури розбиття на інтервали є забезпечення строго монотонного тренду WoE, якщо він логічно доречний та можливий для даної вхідної змінної [52, 53].

Недоліком більшості популярних методів категоризації неперервних змінних є розбиття процедури дискретизації змінних на два окремі етапи [52, 53]:

- 1) початкове розбиття на велику кількість інтервалів (fine classing);
- 2) завершальне об'єднання деяких сусідніх інтервалів у більші інтервали (coarse classing).

Стандартний підхід до початкового розбиття на інтервали неперервних змінних (fine classing) полягає у створенні стандартного числа рівнорозподілених груп (груп рівного розміру), зазвичай від 25 до 50 груп [53]. Таким чином, на першому етапі *інтервали по можливості повинні містити приблизно однакову кількість елементів вибірки* [52], що відповідають сумі кількостей нульових та одиничних елементів. При невеликих вибірках початкове розбиття може здійснюватись на 20 початкових груп або менше [52]. Також труднощі при початковому розбитті можуть спричиняти значення неперервних змінних, що часто повторюються, наприклад, значення нуля, порожнє значення, значення 100% або будь-які інші значення, що, часто повторюючись у вибірці, значно зменшують долі інших значень у вибірці [53]. Також при початковому розбитті можуть ставитись інші додаткові вимоги: (а) наявність щонайменше 1% або 2% від представників кожного класу (нульового та одиничного) в кожному інтервалі або [53] (б) наявність мінімум 50-ти елементів вибірки та мінімум 10-ти елементів вибірки кожного класу (нульового та одиничного) у кожному інтервалі [52]. При неможливості забезпечення таких умов кількість бажаних груп може бути зменшена задля виконання вимог наведеного типу. Серед програмних інструкцій, що корисні для автоматизації першого етапу можна виділити процедурний крок (PROC step) мови програмування SAS Base [17, 31], а саме «PROC RANK» [8, 9], задаючи необхідні значення прихованих опцій, а також у якості частини скриптів для дискретизації, що потребує перевірок та додаткових команд у разі необхідності, можна використовувати віконну аналітичну функцію «NTILE(N) OVER(ORDER BY ...)» або функцію «TOP(N)

PERCENT WITH TIES» («TOP N PERCENT WITH TIES») сумісно з оператором «ORDER BY ...» мови програмування T-SQL (Transact SQL) системи керування базами даних Microsoft® SQL Server® [54, 55].

Процедура завершального об'єднання впорядкованих сусідніх інтервалів (coarse classing) у певній мірі втілює в собі суть четвертого основоположного принципу дискретизації, а саме полягає у рекурсивному об'єднанні сусідніх інтервалів, що найменше відрізняються по WoE та долі нульових значень відповідно, таким чином забезпечуючи мінімальні втрати предикативності. Показовим прикладом є метод рекурсивного об'єднання двох сусідніх інтервалів з найменшою втратою по інформаційній статистиці (Information Value) [53]. Для демонстрації суті методу наведемо приклад застосування вказаного методу для множини деяких довільних інтервалів (які необов'язково були попередньо сформовані за допомогою fine classing) – початок першої ітерації рекурсивного методу представлений в таблиці 2.9 [53]. Вхідним налаштуванням даного методу, очевидно, може бути або встановлення обмеження на максимально допустиму втрату інформаційної статистики при об'єднанні двох сусідніх інтервалів або встановлення цільової кількості необхідних інтервалів. Відповідно, умовою завершення методу може бути або досягнення ситуації відсутності інтервалів, які задовольняють заданій умові обмеження на максимально допустиму втрату інформаційної статистики у разі можливого об'єднання, або досягнення бажаної кількості інтервалів. Щоб на кожній ітерації обрати оптимальну сусідню пару інтервалів, які підлягають подальшому об'єднанню, зручно аналізувати лінійну комбінацію приростів інформаційної статистики, обчислюючи приріст для кожного інтервалу (починаючи з другого) об'єданого з попереднім, віднімаючи його від суми приростів для окремих поточного для попереднього інтервалів (таблиця 2.9). Результуюче число є абсолютним значенням втрати інформаційної статистики у разі можливого об'єднання поточного інтервалу з попереднім, тому обирається пара з мінімальним абсолютним значенням втрати при об'єднанні.

Таблиця 2.9 – Метод рекурсивного об'єднання двох сусідніх інтервалів з найменшою втратою по інформаційній статистиці (початок першої ітерації) [53]

Інтервал	Всього	Цільова змінна		Приріст інформаційної статистики * 10E4			
		Одинична (Good)	Нульова (Bad)	dIV[i]	dIV[i]+ +dIV[i-1]	dIV [i, i-1]	(dIV[i]+dIV[i-1])- -dIV[i, i-1]
до 3	517	410	107	30,63			
до 6	469	368	101	38,11	68,74	68,36	0,39
до 100	1151	957	194	2,22	40,33	21,83	18,50
до 106	590	485	105	5,25	7,47	6,50	0,96
до 206	1386	1144	242	8,10	13,34	13,22	0,13
до 300	1334	1086	248	24,67	32,77	30,46	2,31
до 400	1349	1106	243	15,09	39,75	39,17	0,58
до 500	1487	1217	270	19,10	34,18	34,14	0,04
до 700	2090	1736	354	4,89	23,99	20,44	3,55
до 1000	2531	2184	347	44,96	49,86	11,16	38,69
до 1500	2382	2062	320	53,31	98,27	97,92	0,35
до 2000	1624	1399	225	25,29	78,60	77,96	0,64
Вище	1195	1045	150	49,13	74,42	69,87	4,55
Всього:	18105	15199	2906	320,74			

Вказаний метод реалізує один з алгоритмів сусіднього злиття (Adjacent Pooling Algorithms, APA) [53].

Ще одним критерієм для методів завершального об'єднання сусідніх інтервалів у більші інтервали (coarse classing) може бути формування строго монотонного тренду долі нульових значень (bad rate) відносно порядку інтервалів, що також означає формування монотонного тренду WoE. Такий критерій застосовний, якщо монотонний тренд можливий та логічно коректний. Наведемо інший метод, що реалізує алгоритм монотонного сусіднього злиття (Monotone Adjacent Pooling Algorithm, MAPA), тобто вказаний критерій [53]. Суть методу полягає в оптимізації кумулятивної долі нульових елементів (Cumulative Bad Rate, CBR) лише на даних після попередньої точки розбиття [53]:

$$\forall v \in [V_{k-1} + 1; N]: CBR_{k,v} = \frac{\sum_{i=V_{k-1}+1}^v B_i}{\sum_{i=V_{k-1}+1}^v (B_i + G_i)},$$

$$V_k = \max \left\{ v^* \mid \left\{ \begin{array}{l} CBR_{k,v^*} = \max \{ CBR_{k,v} \mid v \in [V_{k-1} + 1; N] \}, \\ CBR_{k,V_{k-1}} > \max \{ CBR_{k,v} \mid v \in [V_{k-1} + 1; N] \} \end{array} \right. \right\},$$

де v – номер інтервалу, N – кількість інтервалів, B_i та G_i кількості нульових та одиничних значень в інтервалі під номером i (як в підрозділі 2.2), V_k – номер інтервалу, що є завершальним у групі об'єднання під номером k (точка розбиття).

Дані формули наведені для кроку алгоритму під номером k (на першому кроці V_{k-1} слід вважати нульовим) та для випадку формування тренду монотонно спадаючого ризику (bad rate). Приклад наведено в таблиці 2.10 [53].

Таблиця 2.10 – Застосування методу точок розбиття з формуванням тренду

Інтервал	G[i]	B[i]	Всього	Точка розбиття / CBR				
				1	2	3	4	-
до 180	361	52	413	12,6%				
до 181	359	55	414	12,9%				
до 183	827	140	967	13,8%				
до 184	437	54	491	13,2%	11,0%			
до 185	988	135	1123	12,8%	11,7%			
до 187	1184	160	1344	12,5%	11,8%			
до 189	1137	115	1252	11,8%	11,0%	9,2%		
до 191	1386	145	1531	11,4%	10,6%	9,3%		
до 193	1790	202	1992	11,1%	10,5%	9,7%		
до 195	1971	179	2150	10,6%	10,0%	9,3%	8,3%	
до 196	1959	185	2144	10,3%	9,8%	9,1%	8,5%	
до 197	1329	110	1439	10,0%	9,5%	8,9%	8,3%	7,6%
до 199	993	83	1076	9,9%	9,4%	8,8%	8,2%	7,7%

Для іншого напрямку тренду у формулах можна CBR замінити на CGR (Cumulative Good Rate).

Ще однією групою методів, які вже не потребують розбиття на етапи (fine classing та coarse classing), є дерева рішень (підрозділ 3.1.3) у рекурсивному застосуванні до однієї змінної.

Пропонований у рамках дисертаційної роботи метод категоризації полягає у використанні ідеї динамічного програмування Беллмана [56] таким чином, щоб максимізувати значення інформаційної статистики (Information Value).

Пропонованими входними параметрами (налаштуваннями) методу є:

- 1) бажана кількість інтервалів для непустих значень k ;
- 2) мінімально допустима доля вибірки для всіх інтервалів s ;
- 3) кратність правих нестрогих границь інтервалу $\{r_i\}_{i=1}^{k-1}$ (числу t).

Праві нестрогі границі інтервалу для наглядності можна доповнити константними значеннями $r_0 = -\infty$ та $r_k = +\infty$, тоді змінна, що дискретизується, розбивається на k інтервалів: $(r_0; r_1], \dots, (r_{i-1}; r_i], \dots, (r_{k-1}; r_k]$, де ознака включення границі для безкінечних границь не є принциповою. Представимо інформаційну статистику для k непустих інтервалів і пуского (при наявності) у такому вигляді:

$$IV_k(r_0, r_1, \dots, r_{i-1}, r_i, \dots, r_{k-1}, r_k) = dIV_{null} + \sum_{i=1}^k dIV(r_{i-1}, r_i),$$

$$dIV_{null} = (g_{null} - b_{null}) \ln \left(\frac{g_{null}}{b_{null}} \right),$$

$$dIV_{null} = (g_{null} - b_{null}) WoE_{null},$$

$$dIV(r_{i-1}, r_i) = (g(r_{i-1}, r_i) - b(r_{i-1}, r_i)) \ln \left(\frac{g(r_{i-1}, r_i)}{b(r_{i-1}, r_i)} \right),$$

$$dIV(r_{i-1}, r_i) = (g(r_{i-1}, r_i) - b(r_{i-1}, r_i)) WoE(r_{i-1}, r_i),$$

де $g(r_{i-1}, r_i)$ – доля одиничних («good») елементів навчальної вибірки, що відповідає інтервалу $(r_{i-1}; r_i]$ неперервної змінної, по відношенню до загальної кількості одиничних («good») елементів навчальної вибірки; $b(r_{i-1}, r_i)$ – доля нульових («bad») елементів навчальної вибірки, що відповідає інтервалу $(r_{i-1}; r_i]$ неперервної змінної, по відношенню до загальної кількості нульових («bad») елементів навчальної вибірки; g_{null} – доля одиничних («good») елементів навчальної вибірки, що відповідає порожньому значенню («null») неперервної змінної, по відношенню до загальної кількості одиничних («good») елементів навчальної вибірки; b_{null} – доля нульових («bad») елементів навчальної вибірки, що відповідає порожньому значенню («null») неперервної змінної, по відношенню до загальної кількості нульових («bad») елементів навчальної вибірки відповідно.

Тоді оптимізаційна задача, що вирішується за допомогою динамічного Беллмана, записується таким чином, використовуючи рівняння Беллмана і умови:

$$\left\{ \begin{array}{l} \forall L \in \{1, \dots, k-1\} \forall R \in \{L+1, \dots, k\} \forall j \in \mathbb{N} : R < j < L : \\ \max_{\{r_i\}_{i=L}^{R-1}} \sum_{i=L}^R dIV(r_{i-1}, r_i) = \max_{r_j} \left(\max_{\{r_i\}_{i=L}^{j-1}} \sum_{i=L}^j dIV(r_{i-1}, r_i) + \max_{\{r_i\}_{i=j+1}^{R-1}} \sum_{i=j+1}^R dIV(r_{i-1}, r_i) \right), \\ \forall j \in \{1, \dots, k\} : r_{j-1} < r_j, \\ r_0 = -\infty, \\ r_k = +\infty, \\ \forall i \in \{1, \dots, k-1\} : \left[\frac{r_i}{t} \right] t = r_i, \\ \forall i \in \{1, \dots, k\} : g(r_{i-1}, r_i) > 0, \\ \forall i \in \{1, \dots, k\} : b(r_{i-1}, r_i) > 0, \\ \forall i \in \{1, \dots, k\} : \frac{b(r_{i-1}, r_i)B + g(r_{i-1}, r_i)G}{B + G} \geq s. \end{array} \right.$$

де G – загальна кількість одиничних («good») елементів, B – загальна кількість нульових («bad») елементів, позначення « \vee » означає максимум двох чисел, а « \wedge » – мінімум двох чисел.

Класичний метод динамічного програмування Беллмана застосовний до вирішення задачі пакування рюкзака (knapsack problem) [57], що відноситься до NP-повних задач Ричарда Карпа – класу задач обчислювальної NP-складності [57]. Дана задача є задачею цілочисельної оптимізації [57] (задачею булевого програмування [56] або задачею цілочисельного програмування [5]) – задачею максимізації сумарної вартості обраних речей у рюкзак при обмеженні на сумарну вагу (при наявності декількох рішень з однаковою сумарною вартістю, перевагу мають рішення з меншою сумарною вагою). Задача може вирішуватись за допомогою динамічного програмування, оскільки має місце рівняння Беллмана та виконується *принцип оптимальності Беллмана: стратегія на певному кроці повинна бути оптимальною лише відносно поточного стану системи та не повинна залежати від попередніх станів* [56].

Грубу оцінку (верхню границю) кількості можливих варіантів розбиття можна отримати у вигляді біноміального коефіцієнту [58], використовуючи бажану кількість інтервалів k , кратність точок розбиття t та спостережувані емпіричні характеристики неперервної змінної:

$$C_K^{k-1},$$

$$K = \frac{K_{\max} - K_{\min}}{t} + 1,$$

$$K_{\min} = \left[\frac{\min x_i}{t} \right] t - I \left(\left[\frac{\min x_i}{t} \right] t = \min_i x_i \right) t + t,$$

$$K_{\max} = \left[\frac{\max x_i}{t} \right] t - I \left(\left[\frac{\max x_i}{t} \right] t = \max_i x_i \right) t,$$

де $I(\bullet)$ – індикаторна функція відносно істинності умови, $\min_i x_i$ та $\max_i x_i$ – мінімальне та максимальне значення неперервної змінної відповідно, $[\bullet]$ – функція цілої частини у сенсі найбільшого числа, що не перевищує заданого числа (тобто найближчого цілого числа зліва по числовій осі або цілого числа рівного заданому числу).

Більш точні оцінки кількості можливих варіантів розбиття можуть враховувати кількість унікальних значень неперервної змінної, яка виводиться у процесі дискретизації за допомогою системи підтримки прийняття рішень, що розробляється у рамках дисертаційної роботи.

Суть методу полягає у пошуці максимального шляху для спеціальним чином побудованому орієнтованому графу. Така оптимізаційна задача досить легко вирішується за допомогою динамічного програмування [56], коли аналіз починається з кінцевої вершини й надалі для кожної вершини розраховується тут оптимальне по максимуму ребро (якщо таких декілька, то обирається довільне, наприклад, перше), де максимум відповідає максимальній сумі власне значення ребра та зафіксованій максимальній кумулятивній сумі дочірньої вершини.

У разі порушення умови на мінімально допустиму долю вибірки s у інтервалі відповідному ребру графа або у разі відсутності щонайменше по одному представнику для кожного значення бінарної цільової змінної у відповідному інтервалі (три останні сформульовані нерівності у рамках системи рівнянь і нерівностей) ребро графа відповідає від’ємній безкінечності ($dIV = -\infty$).

На рисунку 2.4 схематично зображено спеціальний тип графу, який відповідає методу, що пропонується. При цьому кожна з вершин графу з’єднана з вершинами наступного рівня, але при цьому лише з такими, які відповідають строго більшим значенням точок розбиття. Кількість вершин на рівнях (окрім № 0 та № k , які відповідають рівням з однією вершиною – значенням від’ємної та додатної безкінечності) враховує, що $k - 1$ – загальна кількість точок розбиття:

$$|layer| = K - (k - 2) = \left(\frac{K_{\max} - K_{\min}}{t} + 1 \right) - (k - 2).$$

Формула враховує, що кожен середній рівень починається зі значення зміщеного на крок дискретизації t , а середніх рівнів всього має бути $k - 1$, тому має місце $k - 2$ зміщень відносно K_{\min} (рівень № 1), а останній середній рівень (тут № $k - 1$) має містити у якості останньої вершини вершину зі значенням K_{\max} .

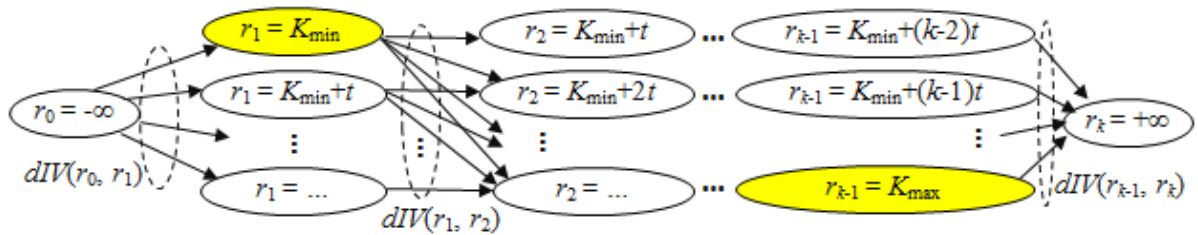


Рисунок 2.4 – Структура графу, яка відповідає пропонованому методу

2.5 Індекс Джині та статистика Колмогорова-Смирнова для аналізу однієї вхідної змінної

Альтернативними показниками предикативної сили вхідних змінних по відношенню до показника інформаційної статистики (Information Value), що по суті є ентропійною метрикою [59], є класичні показники роздільної здатності моделі – індекс Джині та статистика Колмогорова-Смирнова, які детально описані у підрозділах 3.3.1.1 та 3.3.1.2 відповідно. У контексті застосування з метою оцінювання предикативної сили вхідної змінної дані показники обчислюються на навчальній вибірці (аналогічно Information Value), оцінюючи взаємозв'язок WoE-ряду категоріальної або дискретизованої змінної та бінарного ряду цільової змінної. Іноді для неперервних змінних використовується оригінальне неперервне значення, але, як було показано в підрозділі 2.4, з метою урахування нелінійних взаємозв'язків з цільовою змінною, досліджувана

неперервна змінна повинна бути дискретизована на інтервали значень, які надалі підлягають дійснозначному WoE-перетворенню.

Однією з найбільш пов'язаних зі статистичною термінологією формул для обчислення індексу Джині є формула з використанням емпіричних функцій кумулятивного розподілу цільових класів (нульового та одиничного), що також узгоджується з класичною формулою для обчислення статистики Колмогорова-Смирнова [1, 60]:

$$GINI = \left(\int_{x \in X} F_B(x) dF_G(x) - \frac{1}{2} \right) / \left(\frac{1}{2} \right),$$

$$KS = \max_{x \in X} |F_B(x) - F_G(x)|,$$

де $F_B(x)$ та $F_G(x)$ – емпіричні кумулятивні функції розподілів нульових («bad») та одиничних («good») елементів навчальної вибірки (у даному контексті) за бінарною цільовою змінною та відносно дійсних значень досліджуваної змінної (у даному контексті) – числових значень ваг категорій або інтервалів змінної (WoE).

Графічна ілюстрація показників наводиться в підрозділах 3.3.1.1 та 3.3.1.2.

У термінах емпіричних функцій розподілу, індекс Джині обчислюється згідно з формулою Брауна [60, 61] за допомогою методу трапецій:

$$GINI = \frac{\sum_{x_i \in X} \frac{(F_B(x_i) + F_B(x_{i-1})))}{2} (F_G(x_i) - F_G(x_{i-1})) - \frac{1}{2}}{\frac{1}{2}},$$

$$GINI = \sum_{x_i \in X} (F_B(x_i) + F_B(x_{i-1}))(F_G(x_i) - F_G(x_{i-1})) - 1,$$

де множина унікальних значень змінної доповнюється значенням $x_0 = -\infty$.

Показники індексу Джині та статистики Колмогорова-Смирнова як незалежні індикатори відносно параметричних координат-функцій співпадають з власними значеннями при оцінюванні якості прогнозів на навчальній вибірці для однофакторних моделей [44] типу логістичної регресії, які зберігають монотонність відносно входу моделі (у даному разі єдиного). Тому шкала предикативної сили при аналізі окремих змінних повинна бути нижчою, ніж та, що представлена у підрозділі 3.3 для багатфакторних моделей.

2.6 Розробка методу розрахунку статистики Колмогорова-Смирнова, ваги категорії змінної та інформаційної статистики при відомому розподілі категорій та умовному розподілі цільової змінної

Актуальність даного підрозділу полягає у практичній цінності наведення відповідних формул у термінах та поняттях безумовного дискретного розподілу (total distribution) категорій (інтервалів) змінної, що аналізується, та у термінах умовних ймовірностей нульових значень цільової змінної (bad rate) по кожній з категорій (інтервалів) змінної, що аналізується, оскільки два наведені розподіли найбільш ілюстративні при відображенні таблиць та графіків аналізу характеристик (зокрема групи ризику фінального скорингового балу) [10, 44].

У наведеній задачі мають місце вхідні вже агреговані дані без наведення оригінальної множини вибірки – матриця $M = (\mathbf{t} \quad \mathbf{p})$ розмірності $k \times 2$, перший стовпець якої відповідає безумовному розподілу k категорій (інтервалів) вхідної змінної (total distribution), а другий – умовним ймовірностям частот елементів з нульовими значеннями бінарної цільової змінної (bad rate) [44].

Має місце така рівність [44]:

$$\sum_{i=1}^k t_i = 1.$$

Відповідну ймовірнісну вхідну матрицю зручно представити у вигляді графіку аналізу вхідної характеристики відносно бінарної цільової змінної, де гістограмі відповідає безумовний розподіл категорій (інтервалів) вхідної характеристики, а ламаній лінії – відсоток елементів з нульовим цільовим результатом (умовна ймовірність) [44]. Наведемо приклад розподілу клієнтів банку по інтервалах віку клієнта та відсоток випадків некредитоспроможності для кожної вікової категорії (рисунок 2.5) [44].

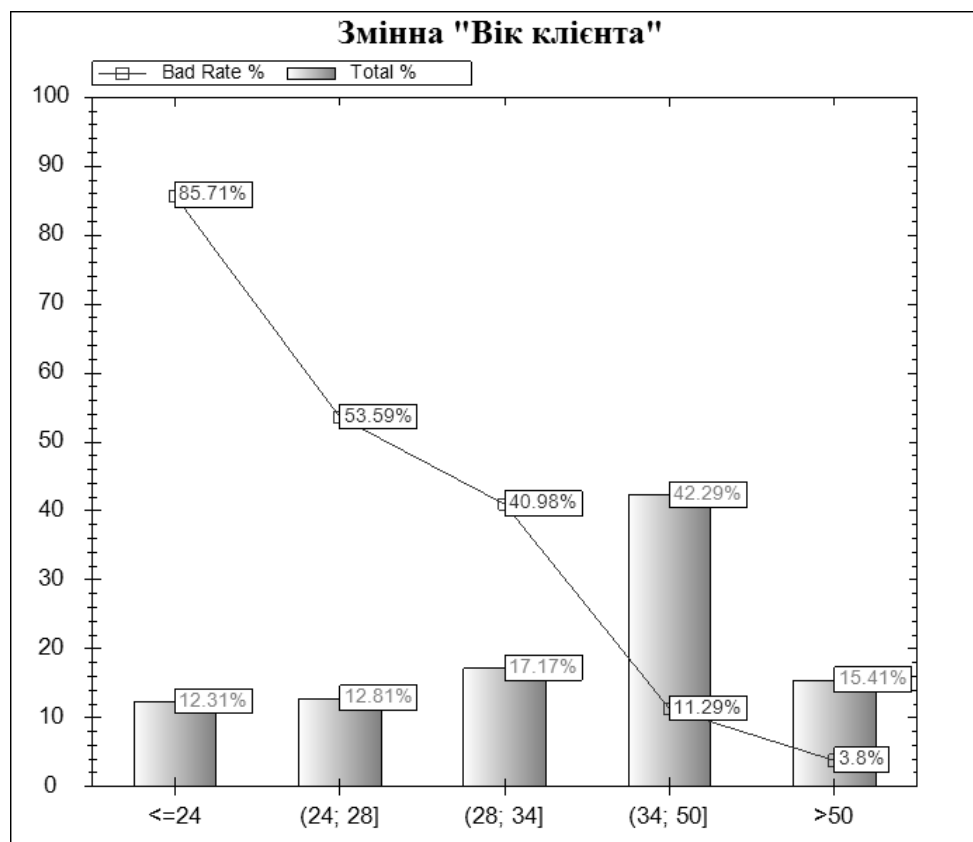


Рисунок 2.5 – Графік аналізу характеристики «вік клієнта» [44]

Щодо аналізу характеристик насамперед переписується формула WoE [44]:

$$\begin{aligned}
WoE_i &= \ln\left(\frac{g_i}{b_i}\right) = \ln\left(\frac{\frac{G_i}{G}}{\frac{B_i}{B}}\right) = \ln\left(\frac{B}{G}\right) - \ln\left(\frac{B_i}{G_i}\right) = \ln\left(\frac{\frac{B}{B+G}}{1 - \frac{B}{B+G}}\right) - \ln\left(\frac{\frac{B_i}{B_i+G_i}}{1 - \frac{B_i}{B_i+G_i}}\right), \\
WoE_i &= \ln\left(\frac{p(y=0)}{1-p(y=0)}\right) - \ln\left(\frac{p(y=0|i)}{1-p(y=0|i)}\right) = \ln(odds_i) - \ln(odds), \\
WoE_i &= \ln\left(\frac{\frac{\sum_{j=1}^c p_j t_j}{\sum_{k=1}^c t_k}}{1 - \frac{\sum_{j=1}^c p_j t_j}{\sum_{k=1}^c t_k}}\right) - \ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{(\mathbf{p}, \mathbf{t})}{1 - (\mathbf{p}, \mathbf{t})}\right) - \ln\left(\frac{p_i}{1-p_i}\right),
\end{aligned}$$

де $p_i = p(y=0|i)$ – умовна ймовірність нульового значення цільової змінної (bad rate) при категорії (інтервалі) під номером i вхідної змінної, тобто координата вектору \mathbf{p} під номером i , $p(y=0)$ – це безумовна ймовірність нульового значення цільової змінної (average bad rate), $odds_i$ та $odds$ – це співвідношення шансів (*odds ratio*; 1 до 0) для категорії i в цілому. Очевидно, що:

$$p(y=0) = (\mathbf{p}, \mathbf{t}) = \mathbf{p}^T \mathbf{t} = \mathbf{t}^T \mathbf{p}.$$

Виведені формули запису WoE доводять, що, якщо умовна ймовірність нульового значення (bad) для категорії (інтервалу) нижча за безумовну (на всій вибірці), то значення WoE є позитивним, аналогічно, якщо «bad rate» вище за «average bad rate», то значення WoE є негативним, якщо умовна ймовірність дорівнює безумовній, то WoE дорівнює нулю. Тобто має місце «обернена» залежність. Аналогічно, якщо співвідношення шансів (*odds ratio*) для категорії є

«кращим» (вищим), то WoE є позитивним, якщо «гіршим» (нижчим) – негативним, якщо однаковим – нульовим. Тут має місце «пряма» залежність.

Аналогічно, формулу WoE можна доповнити до формули Information Value:

$$IV = \sum_{i=1}^k \left(\frac{(1-p_i)t_i}{\sum_{l=1}^k (1-p_l)t_l} - \frac{p_i t_i}{\sum_{r=1}^k p_r t_r} \right) WoE_i = \sum_{i=1}^k \left(\frac{(1-p_i)t_i}{1-\mathbf{p}^T \mathbf{t}} - \frac{p_i t_i}{\mathbf{p}^T \mathbf{t}} \right) WoE_i,$$

$$IV = \sum_{i=1}^k \left(\frac{(1-p_i)t_i}{1-(\mathbf{p}, \mathbf{t})} - \frac{p_i t_i}{(\mathbf{p}, \mathbf{t})} \right) \left(\ln \left(\frac{(\mathbf{p}, \mathbf{t})}{1-(\mathbf{p}, \mathbf{t})} \right) - \ln \left(\frac{p_i}{1-p_i} \right) \right).$$

Щодо статистики Колмогорова-Смирнова, то достатньо означити два оператори: (а) оператор умовної перестановки (сортування відносно спадання координат вектора умови); (б) оператор проектування на підпростір перших координат. Вводиться оператор ранжування (перестановки) одного вектора як перестановка його координат, що відповідає сортуванню другого вектора по спаданню координат $R(\mathbf{x}, \mathbf{y}) : R^k \times R^k \rightarrow R^k$. Суть оператора сортування першого вектора відносно другого вектора по спаданню координат можна представити через обернену функцію рангу $r_{inv}(i, \mathbf{y})$, що визначена на натуральних числах (не більше розмірності власне вектора) та векторі, що сортується. Функція повертає оригінальний номер координати ще не відсортованого вектора \mathbf{y} для заданого як аргумент номера координати вектора \mathbf{y} у відсортованому вигляді [44]:

$$R(\bar{x}, \bar{y}) : \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_k \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_k \end{pmatrix} \rightarrow \begin{pmatrix} x_{r_{inv}(1, \mathbf{y})} \\ x_{r_{inv}(2, \mathbf{y})} \\ \dots \\ x_{r_{inv}(k, \mathbf{y})} \end{pmatrix}.$$

При цьому можна вважати, що сортування відбувається методом вибору (selection sort) [62], при цьому при рівних елементах обирається найперший згідно з початковим порядком координат. Суть оператора дуже проста: запис двох векторів в одну матрицю, подальше сортування рядків матриці відносно спадання другого стовпця (вектора) та відображення лише першого стовпця (вектора) у результаті. Другий стовпець тут завжди «bad rate» (вектор \mathbf{p}).

Другий оператор – оператор проектування перших координат – можна означити як діагональну матрицю, де лише m перших діагональних елементів ненульові та дорівнюють одиниці:

$$\mathbf{P}^{k,m} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \in \text{Mat}(k \times k),$$

$$\mathbf{P}_{ij}^{k,m} = \begin{cases} 1, & \forall i, j \in \{1, \dots, k\} : (i = j) \wedge (i \leq m), \\ 0, & \forall i, j \in \{1, \dots, k\} : (i \neq j) \vee (i > m). \end{cases}$$

Означений таким чином (за допомогою матриці) проектор $\mathbf{P}^{k,m} : R^k \rightarrow R^k$ відповідає основній властивості оператора проектування (проектора), що полягає в рівності проекції від проекції власне проекції [63]:

$$\left(\mathbf{P}^{k,m}\right)^2 = \mathbf{P}^{k,m},$$

$$\forall \mathbf{x} \in R^k : \left(\mathbf{P}^{k,m}\right)^2(\mathbf{x}) = \mathbf{P}^{k,m}(\mathbf{P}^{k,m}(\mathbf{x})) = \mathbf{P}^{k,m}(\mathbf{x}).$$

Пропонований метод обчислення статистики Колмогорова-Смирнова:

$$KS = \max_{m=1, \dots, k} \left| \frac{\left(\mathbf{P}^{k,m}(R(\mathbf{t}, \mathbf{p})), \mathbf{P}^{k,m}(R(\mathbf{p}, \mathbf{p}))\right)}{(\mathbf{t}, \mathbf{p})} - \frac{\left(\mathbf{P}^{k,m}(R(\mathbf{t}, \mathbf{p})), \mathbf{P}^{k,m}(R(\mathbf{e} - \mathbf{p}, \mathbf{p}))\right)}{(\mathbf{t}, \mathbf{e} - \mathbf{p})} \right|,$$

де \mathbf{e} – одиничний вектор з простору R^k .

Введемо позначення для композиції операторів перестановки та проектування:

$$P_{\mathbf{p},k,m}(\bullet) = \mathbf{P}^{k,m}(R(\bullet, \mathbf{p})).$$

Тоді запис формули для статистики Колмогорова-Смирнова набуває такого вигляду:

$$KS = \max_{m=1,\dots,k} \left| \frac{(P_{\mathbf{p},k,m}(\mathbf{t}), P_{\mathbf{p},k,m}(\mathbf{p}))}{(\mathbf{t}, \mathbf{p})} - \frac{(P_{\mathbf{p},k,m}(\mathbf{t}), P_{\mathbf{p},k,m}(\mathbf{e}-\mathbf{p}))}{(\mathbf{t}, \mathbf{e}-\mathbf{p})} \right|.$$

Для прикладу з рисунку 2.5 округлене значення інформаційної статистики дорівнює числу 1,97; аналогічно, округлене значення статистики Колмогорова-Смирнова на навчальній вибірці після дискретизації дорівнює 56,60%.

2.7 Кореляційний аналіз змінних

Одним з найбільш важливих етапів побудови довільних статистичних моделей, особливо регресійного типу [64–68], є кореляційний аналіз вхідних змінних з метою зменшення кількості вхідних змінних. У даному випадку кореляційна матриця $\mathbf{R} \in \text{Mat}(M \times M)$ будується для рядів WoE M предикторів:

$$\mathbf{R}_{ij} = \frac{\frac{1}{N-1} \sum_{n=1}^N \left(\text{WoE}_i(n) - \frac{1}{N} \sum_{r=1}^N \text{WoE}_i(r) \right) \left(\text{WoE}_j(n) - \frac{1}{N} \sum_{l=1}^N \text{WoE}_j(l) \right)}{\sqrt{\frac{1}{N-1} \sum_{n=1}^N \left(\text{WoE}_i(n) - \frac{1}{N} \sum_{r=1}^N \text{WoE}_i(r) \right)^2} \sqrt{\frac{1}{N-1} \sum_{n=1}^N \left(\text{WoE}_j(n) - \frac{1}{N} \sum_{l=1}^N \text{WoE}_j(l) \right)^2}},$$

$$\mathbf{R}_{ij} = \frac{\text{Cov}(WoE_i, WoE_j)}{\sigma(WoE_i)\sigma(WoE_j)},$$

де $WoE_i(n)$ – WoE-перетворення змінної під номером i для спостереження (елемента вибірки) під номером n , аналогічно для змінної під номером j : $WoE_j(n)$ – WoE-перетворення змінної під номером j для спостереження (елемента вибірки) під номером n , N – розмір навчальної вибірки.

Очевидно, кореляційна матриця є симетричною, а її елементи відображають ступені лінійного взаємозв'язку між змінними (тут за посередництвом цільової змінної через WoE-перетворення). Якщо деяка пара вхідних змінних високо корельована між собою по абсолютному значенню (модулю) коефіцієнта кореляції (що є елементом матриці), то змінна з нижчою предикативною силою (зазвичай інформаційною статистикою) відкидається. Іноді ще береться до уваги наявність логічного тренду і т.д.

У рамках дисертаційної роботи використовуються такі границі абсолютного значення кореляції:

- 1) $|\mathbf{R}_{ij}| \in [0\%;25\%]$ – кореляція низька;
- 2) $|\mathbf{R}_{ij}| \in (25\%;50\%]$ – кореляція середня;
- 3) $|\mathbf{R}_{ij}| \in (50\%;75\%]$ – кореляція висока;
- 4) $|\mathbf{R}_{ij}| \in (75\%;100\%]$ – кореляція дуже висока.

2.8 Факторний аналіз. Метод головних компонент

Факторний аналіз [35, 69–72] у сенсі виділення ортогональних (незалежних) прихованих лінійних комбінацій вхідних змінних з різними дисперсіями іноді використовується опціонально при побудові скорингових моделей з метою:

- (а) формування множини взаємно некорельованих вхідних змінних;
- (б) формального зменшення кількості вхідних змінних.

Метод головних компонент (Principal Component(s) Analysis, PCA) [35, 69–72] є одним з методів зменшення розмірності матриці вимірів, що заміняє оригінальний набір змінних на менший набір інших змінних, що відповідають певним взаємно ортогональним лінійним комбінаціям оригінальних змінних (головним компонентам). Суть зменшення розмірності полягає у відкиданні лінійних комбінацій з низьким значенням дисперсії, що забезпечує мінімальні втрати інформації.

Метод головних компонент зводиться до пошуку власних чисел (eigenvalues) та власних векторів (eigenvectors) матриці коваріації (covariance matrix).

Введемо позначення \mathbf{X}_0 – матриця центрованих стовпців (з нульовими середніми значеннями стовпців), що у випадку скорингу відповідає центрованим значенням ваг категорій змінних (Weight of Evidence), де кожній змінній відповідає стовпець. Центрування означає вирахування з кожного стовпця середнього значення ваги категорії змінної по стовпцю змінної. Також матриця вимірів тут не містить службового одиничного стовпця, як у випадку використання в довільній регресії для обчислення вільного члена (зміщення). Тоді матриця коваріації записується таким чином (де N – кількість рядків):

$$\mathbf{C} = \frac{1}{N-1} \mathbf{X}_0^T \mathbf{X}_0.$$

Матриця коваріації є симетричною матрицею. Згідно з властивостями симетричних матриць з дійсними елементами, справедливі такі твердження:

- 1) власні числа симетричної матриці є дійсними;

2) власні вектори, що відповідають різним власним числам є ортогональними:

$$\begin{aligned} \mathbf{C}\mathbf{v}_i &= \lambda_i \mathbf{v}_i, \\ \mathbf{v}_j^T \mathbf{C}\mathbf{v}_i &= \mathbf{v}_j^T \lambda_i \mathbf{v}_i, \\ (\mathbf{C}\mathbf{v}_j)^T \mathbf{v}_i &= \lambda_j \mathbf{v}_j^T \mathbf{v}_i, \\ \lambda_j \mathbf{v}_j^T \mathbf{v}_i &= \lambda_i \mathbf{v}_j^T \mathbf{v}_i, \\ (\lambda_j - \lambda_i) \mathbf{v}_j^T \mathbf{v}_i &= 0, \\ \mathbf{v}_j^T \mathbf{v}_i &= 0. \end{aligned}$$

3) з власних векторів завжди можна утворити ортонормований базис;

4) матрицю можна привести до діагонального вигляду, тобто вона представлена поворотом діагональної матриці, де матрицею повороту є ортогональна матриця стовпців, що відповідають власним векторам симетричної матриці, тобто: $\mathbf{C} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$, де \mathbf{Q} – ортогональна матриця, стовпці якої утворюють базис власних векторів симетричної матриці, а \mathbf{D} – діагональна матриця з власними значеннями симетричної матриці на діагоналі;

5) якщо у симетричної матриці єдине власне значення, то вона діагональна.

Легко довести для часткового випадку симетричної матриці – матриці коваріації, використовуючи центровану матрицю вимірів, що її власні числа є невід'ємними:

$$\begin{aligned} \mathbf{C}\mathbf{v}_i &= \frac{1}{N-1} \mathbf{X}_0^T \mathbf{X}_0 \mathbf{v}_i = \lambda_i \mathbf{v}_i, \\ \mathbf{v}_i^T \frac{1}{N-1} \mathbf{X}_0^T \mathbf{X}_0 \mathbf{v}_i &= \mathbf{v}_i^T \lambda_i \mathbf{v}_i, \end{aligned}$$

$$\frac{1}{N-1} \|\mathbf{X}_0 \mathbf{v}_i\|^2 = \lambda_i \|\mathbf{v}_i\|^2,$$

$$\lambda_i = \frac{1}{N-1} \frac{\|\mathbf{X}_0 \mathbf{v}_i\|^2}{\|\mathbf{v}_i\|^2} \geq 0.$$

Даний факт буде важливий надалі при доведенні, що кожне власне число відповідає дисперсії деякої головної компоненти:

$$\lambda_i = \sigma_i^2.$$

При розкладі коваріаційної матриці з використанням діагональної матриці та матриці повороту, обидві сортуються по спаданню власних значень: стовпці матриць \mathbf{D} та \mathbf{Q} впорядковуються по спаданню λ_i , тобто по спаданню дисперсій, як буде доведено далі. *Матриця вимірів головних компонент набуває такого вигляду:*

$$\hat{\mathbf{X}}_0 = \mathbf{X}_0 \mathbf{Q}.$$

Отримана матриця також центрована, бо кожен її елемент отримується як лінійна комбінація без вільного члену.

Матриця коваріації для матриці головних компонент:

$$\frac{1}{N-1} \hat{\mathbf{X}}_0^T \hat{\mathbf{X}}_0 = \frac{1}{N-1} (\mathbf{X}_0 \mathbf{Q})^T \mathbf{X}_0 \mathbf{Q} = \frac{1}{N-1} \mathbf{Q}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{Q} = \mathbf{Q}^T \mathbf{C} \mathbf{Q} = \mathbf{D}.$$

Використано властивість ортогональної матриці: $\mathbf{Q}^{-1} = \mathbf{Q}^T$.

Даний факт підтверджує, що головні компоненти не корелюють між собою, бо матриця коваріації для головних компонент містить лише дисперсії на діагоналі.

Найпростішим методом скорочення розмірності матриці вимірів головних компонент $\hat{\mathbf{X}}_0$ по кількості стовпців є правило Кайзера (Kaiser's rule) на власні числа матриці \mathbf{C} (де $M \times M$ – розмірність матриці \mathbf{C}):

$$\lambda_i > \frac{\text{tr}(\mathbf{C})}{M}.$$

Після скорочення розмірності матриці вимірів головних компонент отримуємо матрицю $\tilde{\mathbf{X}}_0 \in \text{Mat}(n \times \tilde{M})$, де $\tilde{M} < M$.

Метод головних компонент можна описати як послідовний процес виділення лінійних комбінацій з найбільшою дисперсією при умові вирахування проєкцій на попередні головні компоненти. Зокрема, на першому етапі, при виділенні першої головної компоненти (лінійної комбінації) виконується задача пошуку вектора-аргумента максимізації (argmax) дисперсії першої головної компоненти, що відповідає невідомому вектору $\mathbf{w}_{(1)}$:

$$\mathbf{w}_{(1)} = \arg \max_{\|\mathbf{w}\|=1} \sigma_1^2(\mathbf{w}) = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{1}{N-1} \|\mathbf{X}_0 \mathbf{w}\|^2 \right\} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{1}{N-1} \mathbf{w}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w} \right\},$$

$$\mathbf{w}_{(1)} = \arg \max \left\{ \frac{1}{N-1} \frac{\|\mathbf{X}_0 \mathbf{w}\|^2}{\|\mathbf{w}\|^2} \right\} = \arg \max \left\{ \frac{1}{N-1} \frac{\mathbf{w}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}.$$

Останню форму виразу можна переписати через відношення Релея (Rayleigh quotient) [73] від вектору \mathbf{w} та матриці $\mathbf{X}_0^T \mathbf{X}_0$:

$$\mathbf{w}_{(1)} = \arg \max \left\{ \frac{1}{N-1} R(\mathbf{X}_0^T \mathbf{X}_0, \mathbf{w}) \right\}.$$

Перша форма виразу пошуку аргументу максимуму записується у вигляді системи:

$$\left\{ \begin{array}{l} \frac{1}{N-1} \mathbf{w}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w} \rightarrow \max, \\ \|\mathbf{w}\| = 1; \end{array} \right.$$

$$\left\{ \begin{array}{l} \frac{1}{N-1} \mathbf{w}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w} \rightarrow \max, \\ \sqrt{\mathbf{w}^T \mathbf{w}} = 1; \end{array} \right.$$

$$\left\{ \begin{array}{l} \frac{1}{N-1} \mathbf{w}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w} \rightarrow \max, \\ \mathbf{w}^T \mathbf{w} = 1. \end{array} \right.$$

Функція Лагранжа для системи:

$$L(\mathbf{w}, \lambda_1) = \frac{1}{N-1} \mathbf{w}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w} - \lambda_1 (\mathbf{w}^T \mathbf{w} - 1).$$

Необхідна умова точки екстремуму системи похідних функції Лагранжа:

$$\left\{ \begin{array}{l} \frac{\partial L(\mathbf{w}, \lambda_1)}{\partial \mathbf{w}} = 0, \\ \frac{\partial L(\mathbf{w}, \lambda_1)}{\partial \lambda_1} = 0; \end{array} \right.$$

$$\left\{ \begin{array}{l} \frac{2}{N-1} \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w} - 2\lambda_1 \mathbf{w} = 0, \\ -(\mathbf{w}^T \mathbf{w} - 1) = 0; \end{array} \right.$$

$$\begin{cases} \frac{1}{N-1} \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w} = \lambda_1 \mathbf{w}, \\ \mathbf{w}^T \mathbf{w} = 1; \end{cases}$$

$$\begin{cases} \mathbf{C} \mathbf{w} = \lambda_1 \mathbf{w}, \\ \|\mathbf{w}\| = 1. \end{cases}$$

Дисперсія, що максимізується, має вигляд:

$$\sigma_1^2 = \frac{1}{N-1} \frac{\mathbf{w}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \frac{\mathbf{w}^T \mathbf{C} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \lambda_1,$$

тобто дорівнює коефіцієнту Лагранжа, що є власним числом матриці коваріації.

Остаточно, у термінах матриці коваріації, де $\mathbf{q}_{\lambda_{\max}}$ – власний вектор, що відповідає максимальному власному числу λ_{\max} :

$$\sigma_1^2 \rightarrow \max \Leftrightarrow \lambda_1 \rightarrow \max \Leftrightarrow \begin{cases} \lambda_1 = \lambda_{\max}, \\ \mathbf{w}_{(1)} = \mathbf{q}_{\lambda_{\max}} \end{cases}.$$

Тобто першій головній компоненті відповідає власний вектор матриці коваріації, що відповідає найбільшому власному числу коваріаційної матриці.

Наступні головні компоненти (які надалі у даному підрозділі пронумеровані за допомогою індексу $k = 2, \dots, M$) вираховуються аналогічно, але після вирахування проєкцій на всі попередні головні компоненти з матриці вимірів:

$$\mathbf{X}_{0,k} = \mathbf{X}_0 - \sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{w}_{(i)} \mathbf{w}_{(i)}^T.$$

Наведена формула означає вирахування з кожного рядка (транспонованого вектору спостережень) рядків транспонованих проєкцій (вектору скалярних добуток помноженого на транспонований вектор напрямку). У даній формулі можна виділити добуток Кронекера $\mathbf{w}_{(i)} \mathbf{w}_{(i)}^T$, а саме «квадрат» Кронекера.

Далі виконується аналогічна задача оптимізації:

$$\mathbf{w}_{(k)} = \arg \max_{\|\mathbf{w}\|=1} \sigma_k^2(\mathbf{w}) = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{1}{N-1} \|\mathbf{X}_{0,k} \mathbf{w}\|^2 \right\} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{1}{N-1} \mathbf{w}^T \mathbf{X}_{0,k}^T \mathbf{X}_{0,k} \mathbf{w} \right\}.$$

Підставимо у вираз власний вектор, що відповідає k -тому власному числу оригінальної коваріаційної матриці $\mathbf{C} = \frac{1}{N-1} \mathbf{X}_0^T \mathbf{X}_0$ по спаданню значень її власних чисел:

$$\mathbf{w}_{(k)} = \mathbf{q}_{\lambda_k}.$$

Використаємо властивість ортогональності власних векторів симетричної матриці при підрахунку дисперсії, враховуючи, що попередні оптимальні вектори також відповідають власним векторам:

$$\sigma_k^2 = \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \mathbf{X}_{0,k}^T \mathbf{X}_{0,k} \mathbf{q}_{\lambda_k},$$

$$\sigma_k^2 = \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \left(\mathbf{X}_0 - \sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right)^T \left(\mathbf{X}_0 - \sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right) \mathbf{q}_{\lambda_k},$$

$$\begin{aligned}
\sigma_k^2 &= \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T (\mathbf{X}_0^T \mathbf{X}_0 - \mathbf{X}_0^T \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right) - \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right)^T \mathbf{X}_0 + \\
&\quad + \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right)^T \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right) \mathbf{q}_{\lambda_k}, \\
\sigma_k^2 &= \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{q}_{\lambda_k} - \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \mathbf{X}_0^T \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right) \mathbf{q}_{\lambda_k} - \\
&\quad - \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right)^T \mathbf{X}_0 \mathbf{q}_{\lambda_k} + \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right)^T \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right) \mathbf{q}_{\lambda_k}, \\
\sigma_k^2 &= \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{q}_{\lambda_k} - 0 - \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \left(\sum_{i=1}^{k-1} \mathbf{X}_0 \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right)^T \mathbf{X}_0 \mathbf{q}_{\lambda_k} + 0, \\
\sigma_k^2 &= \mathbf{q}_{\lambda_k}^T \lambda_k \mathbf{q}_{\lambda_k} - \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \left(\sum_{i=1}^{k-1} (\mathbf{q}_{\lambda_i}^T)^T \mathbf{q}_{\lambda_i}^T \mathbf{X}_0^T \right) \mathbf{X}_0 \mathbf{q}_{\lambda_k}, \\
\sigma_k^2 &= \lambda_k - \mathbf{q}_{\lambda_k}^T \left(\sum_{i=1}^{k-1} (\mathbf{q}_{\lambda_i}^T)^T \mathbf{q}_{\lambda_i}^T \right) \frac{1}{N-1} \mathbf{X}_0^T \mathbf{X}_0 \mathbf{q}_{\lambda_k} = \lambda_k - \mathbf{q}_{\lambda_k}^T \left(\sum_{i=1}^{k-1} \mathbf{q}_{\lambda_i} \mathbf{q}_{\lambda_i}^T \right) \lambda_k \mathbf{q}_{\lambda_k} = \lambda_k - 0, \\
\sigma_k^2 &= \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{q}_{\lambda_k} = \frac{1}{N-1} \mathbf{q}_{\lambda_k}^T \mathbf{X}_0^T \mathbf{X}_0 \mathbf{q}_{\lambda_k} = \lambda_k.
\end{aligned}$$

Для пошуку власних векторів та відповідних власних чисел, які на кожній ітерації є максимальними, ідеально підходить саме степеневий метод [74]:

$$\mathbf{x}_{n+1} = \rho_n \mathbf{C} \mathbf{x}_n,$$

$$\rho_n = \frac{1}{\|\mathbf{C} \mathbf{x}_n\|}.$$

Застосування методу аналізу головних компонент у задачах моделювання (зокрема, у задачах кредитного скорингу) може полягати в ортогоналізації вхідних змінних шляхом використання описаних лінійних комбінацій з метою повного виключення кореляцій для перетворених вхідних параметрів. Також

даний метод дозволяє зменшити кількість входів основної моделі (наприклад, логістичної регресії), однак, у загальному випадку це не зменшує кількість задіяних оригінальних входних змінних, оскільки вони зазвичай входять у кожен лінійну комбінацію (головну компоненту або фактор), але з різним значенням коефіцієнтів (наприклад, з близьким до нуля).

2.9 Виключення змінних за допомогою мультиваріаційного аналізу коефіцієнтів побудованої моделі

Мультиваріаційний аналіз коефіцієнтів регресійних моделей полягає в оцінці статистичної значимості (p -value) коефіцієнтів [8–10, 32–35], виходячи з отриманих значень коефіцієнтів (математичних сподівань) та дисперсій коефіцієнтів. Таким чином, оцінки коефіцієнтів регресії розглядаються як випадкові величини, що мають математичне сподівання та варіацію.

На статистичну значимість (p -value) накладається обмеження по виключенню параметра з моделі (наприклад, якщо p -value дорівнює або більше 0,1% – відповідна входна змінна відкидається). Після виключення (при необхідності) змінних з високим значенням p -value, їх статистична значимість переходить, оскільки коефіцієнти переоцінюються й матимуть інші математичне сподівання та дисперсію.

Підрахунок статистичної значимості (p -value) полягає у спростуванні нуль-гіпотези (null hypothesis) на рівність коефіцієнта моделі нулю [10, 32–34].

На основі ітеративного виключення та (або) включення змінних по рівню статистичної значимості ґрунтуються ітеративні методи автоматичного вибору множини змінних: прямий вибір (forward selection), зворотнє виключення (backward elimination), покрокова (stepwise) регресія [8, 9].

Часто замість умови 0,1% для коефіцієнтів логістичної регресії використовують перевірку p -value < 5% (p -value < 0,05) [35].

2.9.1 Оцінювання математичного сподівання та дисперсії коефіцієнтів моделі на прикладі логістичної регресії

Математичне сподівання параметра регресійної моделі дорівнює фактичному оціненому значенню параметру регресії [10, 33].

Для обчислення дисперсії коефіцієнта моделі логістичної регресії спочатку розраховується матриця:

$$\mathbf{I} = \mathbf{X}^T \mathbf{V} \mathbf{X},$$

де \mathbf{X} – матриця вимірів-рядків (доповнена одиничним вектор-стовпцем, що відповідає першому параметру – коефіцієнту зміщення), \mathbf{V} – діагональна матриця, на діагоналі якої розташовані добутки прогнозів ймовірностей двох взаємно виключних класів:

$$\mathbf{V}_{ii} = p_i(1 - p_i),$$

де $p_i \in (0; 1)$ – прогнозне значення моделі логістичної регресії, ймовірність одиничного класу.

Надалі оцінюється коваріаційна матриця, що дорівнює оберненій матриці від отриманої матриці \mathbf{I} :

$$\mathbf{C} = \mathbf{I}^{-1} = (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1}.$$

Діагональ коваріаційної матриці відповідає дисперсії коефіцієнтів моделі:

$$\text{Var}_i = \sigma_i^2 = \mathbf{C}_{ii}.$$

2.9.2 Обчислення рівня статистичної значимості p -value за допомогою функції розподілу Хі-квадрату Вальда

Значення Хі-квадрату Вальда (Wald chi-square statistic) для спростування нуль-гіпотези обчислюється як відношення квадрату значення коефіцієнту моделі до його дисперсії [10, 33]:

$$\chi_i^2 = \frac{c_i^2}{\sigma_i^2}.$$

Значення p -value розраховується виходячи з функції кумулятивного розподілу Хі-квадрату Пірсона (Pearson's chi-square) з одним степенем свободи:

$$F(x) = \int_0^x \frac{e^{-\frac{t}{2}}}{\sqrt{2\pi t}} dt,$$

$$pvalue_i = P(t > \chi_i^2) = 1 - F(\chi_i^2) = 1 - \int_0^{\chi_i^2} \frac{e^{-\frac{t}{2}}}{\sqrt{2\pi t}} dt.$$

Дана функція розподілу є частковим випадком функції розподілу – інтегралу функції щільності розподілу Хі-квадрату Пірсона з n степенями свободи, визначеної на правій дійсній півосі [67]:

$$f(x, n) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}},$$

$$F(x, n) = \int_0^x f(t, n) dt,$$

де гамма-функція $\Gamma(t)$ означена таким чином:

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx.$$

Можна провести взаємозв'язок методу оцінювання статистичної значимості з декомпозицією на зміщення та варіацію (bias-variance decomposition) [47, 48]. Нехай коефіцієнт c_i описується випадковою величиною $\hat{\theta}$ його оцінки. Тоді:

$$E(\hat{\theta}) = c_i.$$

Позначення θ відповідає справжньому значенню коефіцієнта (константі).

Середньоквадратична помилка (Mean Squared Error, MSE) при розгляді коефіцієнта як випадкової величини розписується таким чином [47, 48, 75, 76]:

$$\begin{aligned} E((\hat{\theta} - \theta)^2) &= E((\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2), \\ E((\hat{\theta} - \theta)^2) &= E((\hat{\theta} - E(\hat{\theta}))^2) + E((E(\hat{\theta}) - \theta)^2) + 2E((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)), \\ E((\hat{\theta} - \theta)^2) &= E((\hat{\theta} - E(\hat{\theta}))^2) + (E(\hat{\theta}) - \theta)^2 + 2E((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)), \\ E((\hat{\theta} - \theta)^2) &= \text{Var}(\hat{\theta}) + \text{Bias}^2(E(\hat{\theta}), \theta) + 2E(\hat{\theta}E(\hat{\theta}) - \hat{\theta}\theta - (E(\hat{\theta}))^2 + \theta E(\hat{\theta})), \\ E((\hat{\theta} - \theta)^2) &= \Delta^2(E(\hat{\theta}), \theta) + \text{Var}(\hat{\theta}) + 2((E(\hat{\theta}))^2 - \theta E(\hat{\theta}) - (E(\hat{\theta}))^2 + \theta E(\hat{\theta})), \\ E((\hat{\theta} - \theta)^2) &= \Delta^2(E(\hat{\theta}), \theta) + \text{Var}(\hat{\theta}) + 2 \cdot 0, \\ E((\hat{\theta} - \theta)^2) &= \Delta^2(E(\hat{\theta}), \theta) + \text{Var}(\hat{\theta}) = (c_i - \theta)^2 + \text{Var}(\hat{\theta}). \end{aligned}$$

При спростуванні нуль-гіпотези маємо:

$$\theta = 0,$$

$$E(\hat{\theta}^2) = c_i^2 + \text{Var}(\hat{\theta}) = (E(\hat{\theta}))^2 + \text{Var}(\hat{\theta}) = c_i^2 + \sigma_i^2.$$

У даній дисертаційній роботі пропонується обчислювати p -value шляхом пошуку первісної функції у вигляді ряду шляхом інтегрування ряду Тейлора:

$$pvalue_i = 1 - \int_0^{x_i^2} \frac{e^{-\frac{t}{2}}}{\sqrt{2\pi t}} dt = 1 - \int_0^{x_i^2} \frac{\sqrt{2} e^{-\frac{(\sqrt{t})^2}{2}}}{\sqrt{\pi}} d\sqrt{t} = 1 - \int_0^{x_i^2} \frac{2e^{-\left(\frac{\sqrt{t}}{2}\right)^2}}{\sqrt{\pi}} d\sqrt{\frac{t}{2}},$$

$$pvalue_i = 1 - \frac{2}{\sqrt{\pi}} \int_0^{\frac{x_i^2}{2}} e^{-\left(\frac{\sqrt{t}}{2}\right)^2} d\sqrt{\frac{t}{2}} = 1 - \frac{2}{\sqrt{\pi}} \int_0^{\frac{x_i}{\sqrt{2}}} e^{-y^2} dy = 1 - \text{erf}\left(\frac{x_i}{\sqrt{2}}\right),$$

де функція помилки (error function) $\text{erf}(z)$ означена таким чином:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-y^2} dy,$$

після урахування розкладу функції експоненти у ряд Тейлора:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots + \frac{x^n}{n!} + \dots = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

маємо:

$$\int_0^z e^{-y^2} dy = \int_0^z \sum_{n=0}^{\infty} \frac{(-1)^n y^{2n}}{n!} dy = \sum_{n=0}^{\infty} \frac{(-1)^n y^{2n+1}}{n!(2n+1)} \Big|_0^z = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!(2n+1)},$$

остаточно:

$$pvalue_i = 1 - \frac{1}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n \chi_i^{2n+1}}{n!(2n+1)2^{\frac{n-1}{2}}} = 1 - \frac{1}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n |c_i|^{2n+1}}{n!(2n+1)\sigma_i^{2n+1}2^{\frac{n-1}{2}}},$$

що при програмній реалізації мовами програмування третього покоління може бути імплементовано, опираючись на особливості арифметичного оперування типами даних типу IEEE 754 [77, 78] для збереження дійсних чисел, а саме може бути запрограмовано алгоритм віднімання члену ряду до тих пір, поки значення результату змінюється, оскільки арифметичні операції додавання та віднімання не змінюють мантису (у сенсі без урахування знаку) більшого по модулю числа, якщо порядки (експоненти) чисел дуже відрізняються.

Необхідно також наголосити на очевидній рівності:

$$pvalue_i = P(t > \chi_i^2) = P(\sqrt{t} > |\chi_i|) = P\left(\sqrt{t} > \frac{|c_i|}{\sigma_i}\right).$$

2.9.3 Розробка алгоритму обчислення нормованих ваг змінних моделі

Виходячи з критерію досягнення максимального значення функції правдоподібності (див. підрозділ 3.1.1.2), при заміні деякої змінної на точно пропорційну змінну (помножену на константу) відповідний коефіцієнт змінної змінюється обернено пропорційно, тому для того, щоб оцінити ступінь впливу змінної в моделі логістичної регресії, виключаючи фактор масштабу змінної, недостатньо лише значень коефіцієнтів логістичної регресії.

Одним з незалежних від масштабу змінних пропонованих алгоритмів для оцінки ступеню впливу змінних у складі моделі може бути обчислення

відношення добутку абсолютного значення коефіцієнта і середньоквадратичного відхилення змінної до суми добутків абсолютних значень коефіцієнтів та середньоквадратичних відхилень змінних:

$$w_i = \frac{|c_i| \delta_i}{\sum_{i=1}^M |c_i| \delta_i}.$$

Коефіцієнт зміщення (intercept) не використовується при оцінці нормованих ваг змінних, однак при його включенні ваги не змінюються, оскільки середньоквадратичне відхилення одиничної константи як штучної вхідної змінної дорівнює нулю.

У випадку використання логістичної регресії таким чином оцінюється вплив змінних на рівні логіта (лінійної частини формули логістичної моделі).

Як правило, у скорингових моделях при використанні перетворень типу WoE в моделі логістичної регресії значення коефіцієнтів співпадають з їх абсолютними значеннями (тобто вони невід'ємні, а саме позитивні), якщо в моделі не використовуються пари змінних з високою взаємною кореляцією, не використовуються змінні з високим рівнем статистичної значимості (*p*-value) спростування нуль-гіпотези (null hypothesis) для відповідних коефіцієнтів моделі, та не використовуються змінні з низьким значенням предикативності, наприклад, з низьким значенням інформаційної статистики (Information Value). Якщо деякі коефіцієнти виявляються від'ємними, то в моделі інвертується тренд WoE відповідних змінних, тому множину змінних моделі необхідно змінити.

Виходячи з зазначеного вище, при розробці програмного продукту операція взяття модуля не використовувалася.

2.10 Висновки до другого розділу

У даному розділі детально проаналізовано існуючі способи побудови вибірок на прикладі кредитного скорингу, методи використання категоріальних змінних та методи дискретизації неперервних змінних, включаючи методи попередньої обробки даних, що відображають суттєву частину процесу побудови моделей (а саме перший рівень ієрархії диференціації моделі) шляхом урахування нелінійних взаємозв'язків з цільовою змінною за допомогою WoE-перетворення, також розглянуто методи оцінювання предикативної сили вхідних змінних, методи виявлення та усунення взаємозв'язків вхідних змінних, що включаються до моделі, метод оцінювання статистичної значимості коефіцієнтів на прикладі логістичної моделі. Також запропоновано наступні методи та алгоритми згідно з постановкою задач для дисертаційного дослідження:

(1) оригінальний метод оптимальної дискретизації неперервних змінних за допомогою методу динамічного програмування Беллмана, де критерієм оптимальності є максимізація інформаційної статистики, використовуючи три основні умови: (а) кратність границь інтервалів; (б) мінімально допустима доля вибірки у найменшому інтервалі непустих значень; (в) бажана кількість інтервалів непустих значень (задача № 1 постановки задач);

(2) метод обчислення статистики Колмогорова-Смирнова, ваги категорій змінної, інформаційної статистики при відомому розподілі категорій і умовному розподілі цільової змінної для кожної з категорій (задача № 2 постановки задач);

(3) алгоритм безпосереднього обчислення рівня статистичної значимості (p -value) для коефіцієнтів логістичної моделі шляхом інтегрування розкладу в ряд Тейлора, пошуку первісної функції у вигляді ряду (задача № 3 постановки задач);

(4) алгоритм нормування ваг вхідних змінних з урахуванням їх варіації та коефіцієнтів моделі (перша частина завдання № 5 постановки задач).

Додатково доведено зв'язок відстані Кульбака-Лейблера з IV і навіть WoE.

РОЗДІЛ 3

МЕТОДИ ПОБУДОВИ ПОПЕРЕДНІХ МОДЕЛЕЙ, МЕТОДИ ВКЛЮЧЕННЯ ВІДХИЛЕНИХ ЗАЯВОК, МЕТОДИ ОЦІНЮВАННЯ ЯКОСТІ ПРОГНОЗІВ ТА АНАЛІЗ СТІЙКОСТІ МОДЕЛЕЙ

3.1 Методи побудови попередніх скорингових моделей

Сучасний стрімкий прогрес в області сучасного ризик-менеджменту [4, 11], зокрема в галузі кредитного скорингу [5, 8–11, 13, 42, 52, 53, 60], забезпечується швидким розвитком методів кількісного аналізу [4, 11], розвитком інформаційних технологій [8, 9, 11, 60], розвитком методів інтелектуального аналізу даних (data mining) [5, 8–11, 13, 16–18, 40], зокрема статистичних та нестатистичних методів побудови скорингових моделей [5, 11].

Основні методи побудови моделей можна розділити таким чином [5, 11]:

1) статистичні методи побудови скорингових карт:

1.1) лінійна регресія;

1.2) логістична регресія (*нелінійна*) [5, 8–11, 13, 32–35, 43, 52, 53];

1.3) пробіт-регресія (*нелінійна*) [5, 52, 53];

1.4) дерева рішень (рекурсивний підхід розбиття);

1.5) методи найближчих сусідів:

1.5.1) метод найближчого сусіда;

1.5.2) метод k -найближчих сусідів;

2) нестатистичні методи побудови скорингових карт:

2.1) лінійне програмування;

2.2) цілочисельне програмування;

2.3) нейронні мережі;

2.4) генетичні алгоритми;

2.5) експертні системи;

3) альтернативні змішані методи побудови скорингових карт:

3.1) байєсівські мережі та графічні моделі [16, 17, 79];

3.2) моделі аналізу виживання.

3.1.1 Регресійні методи моделювання

Регресійні методи моделювання у скорингу призначені для побудови статистичних моделей множинної регресії такого типу:

$$E(y | \mathbf{x}) = f(\mathbf{x}, \mathbf{c}),$$

де y – цільова змінна (target variable), \mathbf{x} – вектор вхідних змінних (input variables), \mathbf{c} – вектор оптимальних оцінок (estimates) коефіцієнтів (coefficients) моделі, f – функція регресії, $E(y | \mathbf{x})$ – умовне сподівання цільової змінної y відносно вектору вхідних змінних \mathbf{x} . Тут вхідні змінні – значення WoE змінних.

3.1.1.1 Лінійна регресія. Метод найменших квадратів. Основний недолік лінійної моделі

Модель множинної лінійної регресії записується таким чином [52, 66, 80]:

$$E(y | \mathbf{x}) = c_0 + \sum_{i=1}^M c_i x_i,$$

де M – кількість вхідних змінних, а також $\mathbf{x}^T = (x_1 \quad x_2 \quad \dots \quad x_i \quad \dots \quad x_M)$,
 $\mathbf{c}^T = (c_0 \quad c_1 \quad c_2 \quad \dots \quad c_i \quad \dots \quad c_M)$.

Якщо доповнити вектор змінних одиничним входом (нульова координата):

$$\hat{y} = E(y | \mathbf{x}_{obs}) = c_0 x_0 + \sum_{i=1}^M c_i x_i = \sum_{i=0}^M c_i x_i = \mathbf{c}^T \mathbf{x}_{obs} = \mathbf{x}_{obs}^T \mathbf{c} = (\mathbf{x}_{obs}, \mathbf{c}),$$

де для вектору спостережень \mathbf{x}_{obs} розмірності вже $1+M$ маємо $x_0 \equiv 1$.

У термінах рядків матриці спостережень \mathbf{X} , де кожен *рядок* $\mathbf{X}(i)$ – це *транспонований вектор спостережень* (тому у матриці \mathbf{X} перший стовпець одиничний), та у термінах координат-прогнозів $\hat{y}(i)$ (математичних сподівань) прогнозного вектору вимірів $\hat{\mathbf{y}}$ (для фактичного вектору вимірів \mathbf{y} з координатами $y(i)$):

$$\hat{y}(i) = E(y(i) | \mathbf{X}(i)) = \mathbf{X}(i) \mathbf{c}_{LS},$$

$$\hat{\mathbf{y}} = \mathbf{X} \mathbf{c}_{LS},$$

де \mathbf{c}_{LS} – оптимальний вектор \mathbf{c} обчислений за допомогою методу найменших квадратів (МНК; method of Least Squares, *LS*), де критерій мінімізації:

$$\|\mathbf{y} - \mathbf{X} \mathbf{c}_{LS}\|^2 \rightarrow \min ,$$

формула обчислення оптимального вектору за допомогою МНК [47, 48, 66]:

$$\mathbf{c}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{X}^+ \mathbf{y} .$$

де \mathbf{X}^+ – псевдообернена матриця матриці \mathbf{X} [47, 48].

Основний недолік з точки зору застосування у задачах кредитного скорингу полягає у необмеженості прогнозних значень, тоді як описані далі нелінійні типи регресії забезпечують обмеженість в рамках інтервалу (0; 1).

3.1.1.2 Узагальнена ймовірнісна нелінійна регресія. Метод максимальної правдоподібності

У термінах підрозділу 3.1.1 (зокрема підрозділу 3.1.1.1) функція регресії f є неперервним диференційовним монотонно зростаючим відображенням F від лінійної функції змінних, тобто $f(\mathbf{x}, \mathbf{c}) = F(\mathbf{x}_{obs}^T \mathbf{c})$, де $F: R \rightarrow (0; 1)$. Це означає, що функція F є деякою кумулятивною функцією розподілу $F(z) = P(\xi \leq z)$. Введемо позначення $P(\mathbf{c}, \mathbf{x}_{obs}) = f(\mathbf{x}, \mathbf{c}) = F(\mathbf{x}_{obs}^T \mathbf{c})$, тоді функція правдоподібності:

$$L(\mathbf{c}) = \prod_{i: y_i=1} P(\mathbf{c}, \mathbf{X}(i)) \prod_{i: y_i=0} (1 - P(\mathbf{c}, \mathbf{X}(i))) = \prod_{i=1}^N (P(\mathbf{c}, \mathbf{X}(i)))^{y_i} (1 - P(\mathbf{c}, \mathbf{X}(i)))^{1-y_i},$$

де $y_i = y(i)$, N – кількість елементів навчальної вибірки.

Метод максимальної правдоподібності (ММП; method of Maximum Likelihood Estimation, *MLE*) передбачає максимізацію $L(\mathbf{c})$. При цьому зазвичай простіше максимізувати логарифм функції правдоподібності ($\ln L(\mathbf{c}) \rightarrow \max$):

$$\ln L(\mathbf{c}) = \sum_{i=1}^N (I_{\{1\}}(y_i) \ln P(\mathbf{c}, \mathbf{X}(i)) + I_{\{0\}}(y_i) \ln(1 - P(\mathbf{c}, \mathbf{X}(i)))) ,$$

$$\ln L(\mathbf{c}) = \sum_{i=1}^N (y_i \ln P(\mathbf{c}, \mathbf{X}(i)) + (1 - y_i) \ln(1 - P(\mathbf{c}, \mathbf{X}(i)))) .$$

Для пошуку \mathbf{c}_{MLE} зазвичай використовується метод Ньютона:

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \left(\frac{d^2 \ln L(\mathbf{c})}{d\mathbf{c}^2} \right)^{-1} \frac{d \ln L(\mathbf{c})}{d\mathbf{c}} = \mathbf{c}_n - \mathbf{H}^{-1}(\mathbf{c}_n) \mathbf{g}(\mathbf{c}_n).$$

3.1.1.2.1 Пробіт-регресія на основі нормального розподілу

Суть пробіт-регресії полягає у використанні нормального розподілу [5]:

$$P(\mathbf{c}, \mathbf{X}(i)) = P(y_i = 1 | \mathbf{X}(i)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathbf{X}(i)\mathbf{c}} e^{-\frac{t^2}{2}} dt.$$

3.1.1.2.2 Логістична регресія. Аналітичні формули для градієнта та матриці Гессе

Суть логістичної регресії полягає у використанні логістичного перетворення (сигмоїдальної функції), що відповідає логістичному розподілу з нульовим середнім та середньоквадратичним відхиленням, що дорівнює числу

$$\frac{\pi}{\sqrt{3}} \text{ [32, 43]:}$$

$$\tilde{y}(i) = P(\mathbf{c}_{MLE}, \mathbf{X}(i)) = P(y_i = 1 | \mathbf{X}(i)) = \frac{1}{1 + e^{-\mathbf{X}(i)\mathbf{c}_{MLE}}}.$$

Це означає, що логарифм відношення шансів для конкретного прогнозу:

$$\ln(odds(\tilde{y}(i))) = \ln\left(\frac{P(y_i = 1 | \mathbf{X}(i))}{P(y_i = 0 | \mathbf{X}(i))}\right) = \ln\left(\frac{\tilde{y}(i)}{1 - \tilde{y}(i)}\right) = \mathbf{X}(i)\mathbf{c}_{MLE}.$$

Оскільки кожне спостереження є рядком значень WoE, доповненим одиницею на початку, а WoE кожної змінної виражається як різниця логарифмів співвідношення шансів в межах категорії та співвідношення шансів для вибірки

в цілому (див. підрозділ 2.6), то суть моделі логістичної регресії при використанні WoE-перетворень концептуально можна записати таким чином:

$$\ln\left(\frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})}\right) = c_0 - \frac{P(y=1)}{1-P(y=1)} \sum_{i=1}^M c_i + \sum_{i=1}^M c_i \ln\left(\frac{P(y=1|x_i)}{1-P(y=1|x_i)}\right).$$

Логістичне перетворення $\varphi(t) = \frac{1}{1+e^{-t}}$ є розв'язком диференціального рівняння:

$$\frac{d\varphi(t)}{dt} = \varphi(t)(1-\varphi(t)).$$

Зважаючи на це, легко довести, що аналітичні формули вектора градієнта (перша похідна) та матриці Гессе (друга похідна) для використання в методі Ньютона для оцінювання вектора коефіцієнтів логістичної регресії мають такий вигляд [32, 33, 43]:

$$\mathbf{g}(\mathbf{c}) = \frac{d \ln L(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^N (y_i - P(\mathbf{c}, \mathbf{X}(i))) \mathbf{X}^T(i),$$

$$\mathbf{H}(\mathbf{c}) = \frac{d^2 \ln L(\mathbf{c})}{d\mathbf{c}^2} = -\sum_{i=1}^N P(\mathbf{c}, \mathbf{X}(i))(1 - P(\mathbf{c}, \mathbf{X}(i))) \mathbf{X}^T(i) \mathbf{X}(i).$$

У другій формулі має місце «квадрат Кронекера» $\mathbf{X}^T(i) \mathbf{X}(i) = \mathbf{X}^T(i) \otimes \mathbf{X}(i)$.

3.1.1.3 Формула Фробеніуса в лінійній та нелінійній регресії

У регресійному аналізі як правило обернення потребують лише симетричні матриці, а саме зазвичай відбувається пошук обернених матриць типу $(\mathbf{X}^T \mathbf{X})^{-1}$ або відносно комбінацій таких добутоків або матриць Гессе. Тому найзручніший спосіб – формула Фробеніуса для рекурсивного обернення симетричних матриць:

$$\begin{bmatrix} \mathbf{H} & \mathbf{h} \\ \mathbf{h}^T & \eta \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{H}^{-1} + \frac{1}{\beta} \mathbf{H}^{-1} \mathbf{h} \mathbf{h}^T \mathbf{H}^{-1} & -\frac{1}{\beta} \mathbf{H}^{-1} \mathbf{h} \\ -\frac{1}{\beta} \mathbf{h}^T \mathbf{H}^{-1} & \frac{1}{\beta} \end{bmatrix},$$

$$\beta = \eta - \mathbf{h}^T \mathbf{H}^{-1} \mathbf{h}.$$

3.1.2 Кластерний аналіз та машинне навчання на основі пам'яті

Кластерний аналіз (cluster analysis) та машинне навчання на основі пам'яті (memory-based learning) ґрунтуються на метричному просторі лише вхідних змінних (input variables), а цільова змінна (target variable) може слугувати тільки для подальшої класифікації нового вхідного елемента, який потребує прогнозування. Часто методи даної області відносять до «навчання без вчителя» (unsupervised learning) [47, 48]. Виділяють такі основні методи [47, 48]:

1) методи кластеризації (data clustering):

1.1) метод k -середніх (k -means clustering) [47, 48];

2) методи найближчих сусідів (nearest neighbor methods) [5, 11, 47, 48, 81–83]:

2.1) метод найближчого сусіда (nearest neighbor search, NNS);

2.2) метод k -найближчих сусідів (k -nearest neighbors algorithm, k -NN).

3.1.2.1 Метод k -середніх

Метод складається з таких кроків:

- 1) ініціалізація: вибір початкового положення «центрів» $\{\mathbf{C}_i\}_{i=1}^k$;
- 2) цикл поки центри не стабілізуються (ознака збіжності методу):
 - 2.1) віднесення кожного елемента вибірки до найближчого центру;
 - 2.2) перерахування центрів $\{\mathbf{C}_i\}_{i=1}^k$ як середнього згрупованих елементів.

У результаті методу мінімізується критерій:

$$V = \sum_{i=1}^k \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \mathbf{C}_i\|^2,$$

де S_i – множина згрупованих елементів з кроку 2.2 – кластер, який відповідає центру \mathbf{C}_i . Ініціалізація у класичному методі відбувається шляхом випадкового вибору початкового положення центрів. Вдосконалення даного методу передбачають, що кожен елемент навчальної вибірки відноситься до кожного центру (тобто кластеру) з різними ймовірностями, залежними від відстаней. Ймовірнісна класифікація нового елемента у класичному методі може відбуватися шляхом віднесення до найближчого по центру кластеру і присвоєння ймовірностей, що відповідають частотам бінарних значень у кластері.

3.1.2.2 Метод найближчого сусіда

Метод найближчого сусіда полягає у простій класифікації нового елемента шляхом присвоєння йому прогнозу класу рівного класу найближчого по метриці елемента у навчальній вибірці. Даний метод є частковим випадком методу k -найближчих сусідів при $k = 1$.

3.1.2.3 Метод k -найближчих сусідів

Приклад застосування класичного методу k -найближчих сусідів (k -nearest neighbor method) як машинного навчання на основі пам'яті (memory-based learning) у двовимірному просторі змінних при $k = 3$ (результат прогнозу: $y^* = 1$) зображено на рисунку 3.1 [11, 47, 48].

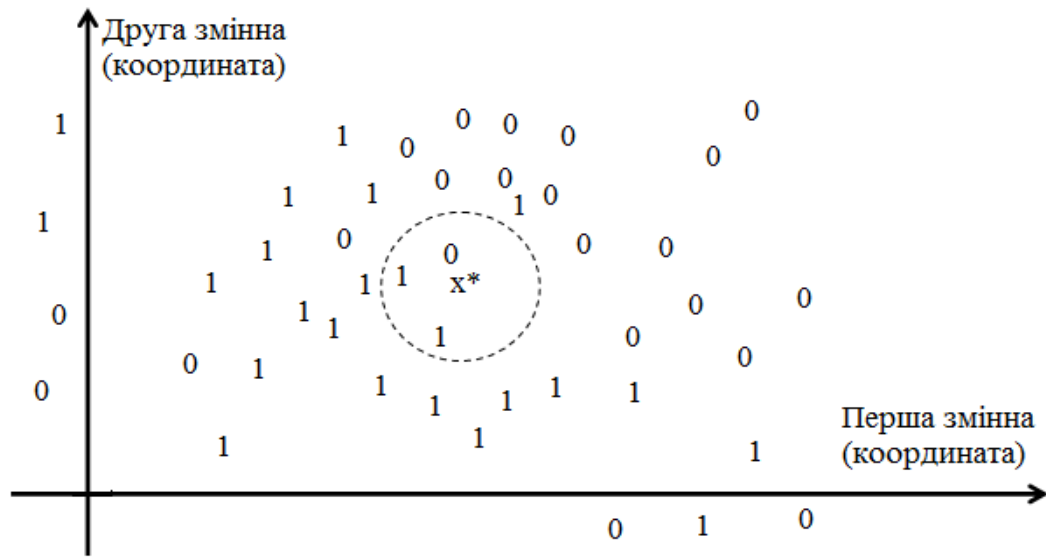


Рисунок 3.1 – Приклад для методу трьох найближчих сусідів [11, 47, 48]

Формалізована суть методу описується такими формулами [11]:

$$\begin{cases} \mathbf{x}_1 = \arg \min_{\mathbf{x} \in X} d(\mathbf{x}, \mathbf{x}^*); \\ \forall i \in \{2, \dots, k\} : \mathbf{x}_i = \arg \min_{\mathbf{x} \in X \setminus \bigcup_{j=1}^{i-1} \mathbf{x}_j} d(\mathbf{x}, \mathbf{x}^*), \end{cases}$$

$$y^* = \begin{cases} 1, & \sum_{i=1}^k \frac{y_i}{k} > \frac{1}{2}; \\ 0, & \sum_{i=1}^k \frac{y_i}{k} \leq \frac{1}{2}. \end{cases}$$

3.1.3 Древа рішень

Основним альтернативним методом класифікації по відношенню до регресійних методів моделювання (в т.ч. таких вдосконалень як наполеглива регресія [84]) є рекурсивно-партиційні алгоритми (recursive partitioning algorithms, RPA) у вигляді дерев класифікації (classification trees) – дерев рішень (*decision trees*) [5, 8, 9, 19, 52, 53, 80, 85]. Основу дерев рішень та алгоритмів типу ID3 становить індекс ентропії інформації [86]. Рисунок 3.2 – приклад дерева рішень.

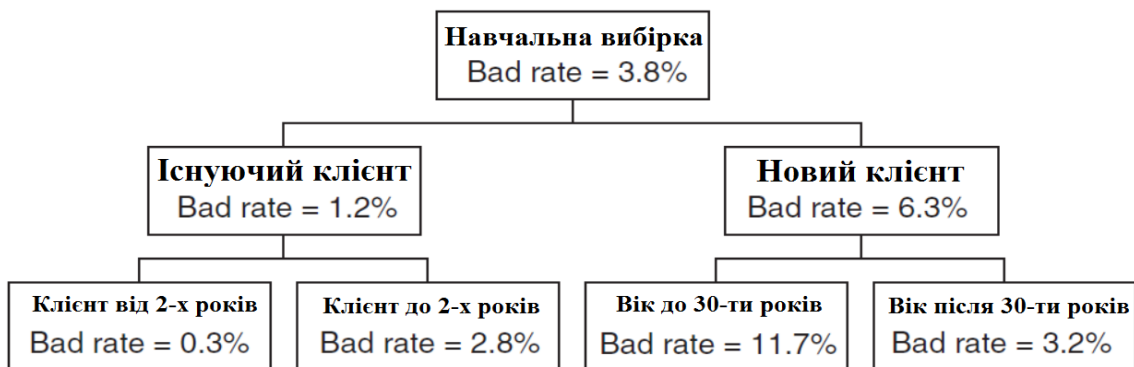


Рисунок 3.2 – Приклад дерева рішень (decision tree) [8, 9]

Аналогічно інформаційній статистиці вводиться поняття приросту інформації (*information gain*) при кожному розбитті підмножини (множини) A на менші підмножини A_i , що відповідають категоріям або інтервалам змінної Q :

$$Gain(A, Q) = H(A, Good) - \sum_{i=1}^q \frac{|A_i|}{|A|} H(A_i, Good),$$

де ентропія [59, 86] розподілу «один \ нуль» («*Good* \ *Bad*») означена як:

$$H(A, Good) = - \sum_{j=0}^1 p(y = j | A) \log_2 p(y = j | A),$$

$$H(A_i, Good) = - \sum_{j=0}^1 p(y = j | A_i) \log_2 p(y = j | A_i).$$

При кожному розгалуженні дерева обирається змінна з максимальним *Gain*. Для неперервних змінних має місце описана раніше задача оптимального розбиття, але тут вже по відношенню до приросту інформації, що в силу скінченності навчальної вибірки може бути зведена до задачі дискретного програмування [87], яка може бути спрощена описаними раніше способами.

Дерева рішень – простий інструмент включення корельованих змінних.

3.1.4 Нейронні мережі. Алгоритм оберненого розповсюдження помилки

Нейронні мережі – ефективний інструмент для ймовірнісної класифікації, що має недолік високої схильності до перенавчання (*overfitting*) [5, 19, 47, 48]. Основою машинного навчання мереж є рекурсивний алгоритм оберненого розповсюдження помилки (*error back-propagation algorithm*) – див. рисунок 3.3:

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_i(n),$$

$$\delta_k(n) = \varphi'_k(v_k(n)) e_k(n) = \varphi'_k(v_k(n)) (d_k(n) - y_k(n)),$$

$$\delta_j(n) = \varphi'_j(v_j(n)) \sum_{k \in \text{Child}(j)} w_{kj}(n) \delta_k(n).$$

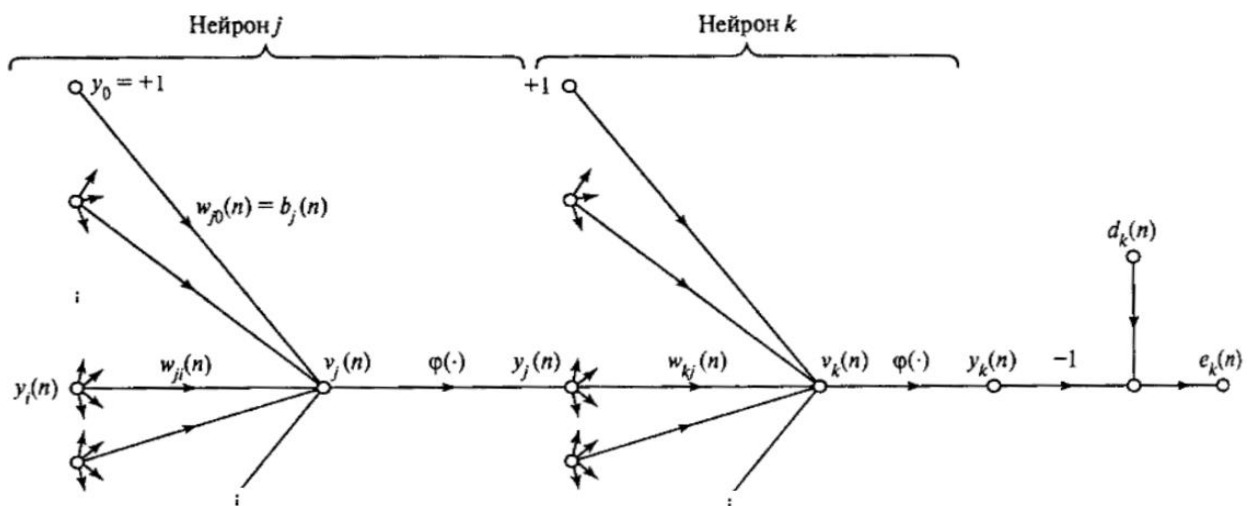


Рисунок 3.3 – Ілюстрація двох нейронів сусідніх шарів мережі [47, 48]

3.1.5 Генетичні та еволюційні алгоритми

Суть класичного генетичного алгоритму полягає у представленні вектора коефіцієнтів деякої абстрактної апроксимуючої моделі у вигляді бінарного рядка. Кожне «покоління» (generation) закодованих у бінарному вигляді варіантів векторів коефіцієнтів (наприклад, $n_{pop} = 100$ рішень) підлягає «схрещуванню» (crossover) з ймовірністю, наприклад, $p_c = 0,75$ для кожної «хромосоми» (chromosome), тобто бінарної послідовності числових представлень коефіцієнтів (наприклад, типу IEEE 754 [77, 78]) [5]. Операція означає обмін випадковими n останніми бітами. Також кожен біт кожного бінарного рядка підлягає мутаціям (інвертуванню) з ймовірністю мутації, наприклад, $p_m = 0,05$ [5]. Кожне наступне покоління формується виходячи з ймовірності включення рядка з попереднього (доля в нових n_{pop}), що дорівнює (де f_j – цільовий критерій оптимальності) [5]:

$$p_j = \frac{f_j}{\sum_{j=1}^{n_{pop}} f_j}.$$

Суть класичного методу диференціальної еволюції полягає у оперуванні поколіннями звичайних векторів коефіцієнтів розміру n_{pop} (параметр методу), коли для кожного у поколінні вектора \mathbf{v}_i обираються три інші випадкові вектори з покоління та обчислюється «мутантний вектор» $\mathbf{v}_i^m = \mathbf{v}_i^1 + F(\mathbf{v}_i^2 - \mathbf{v}_i^3)$, де F – параметр методу з інтервалу $[0; 2]$. Далі координати вектора \mathbf{v}_i^m з ймовірністю p_c (параметр методу) замінюються на відповідні координати \mathbf{v}_i (crossover), потім «пробний вектор» \mathbf{v}_i^{m*} порівнюється з \mathbf{v}_i по критерію мінімізації на можливість його заміни у наступному поколінні [88]. Ці методи популярні в скорингу [89].

3.1.6 Наївний байєсівський класифікатор. Інші методи моделювання

Наївний байєсівський класифікатор (Naive Bayes classifier) [16, 17, 19], що відноситься до *байєсівських мереж*, оснований на припущенні незалежності вхідних змінних та не потребує системного критерію у вигляді «згортки» [90, 91]:

$$p(y = 1 | \mathbf{x}) = \frac{p(y = 1) \prod_{i=1}^M p(x_i | y = 1)}{\sum_{j=0}^1 p(y = j) \prod_{i=1}^M p(x_i | y = j)}.$$

До інших методів відносяться експертні системи, лінійне програмування, цілочисельне програмування, моделі аналізу виживання, машини опорних векторів і т.д. Генетичні і еволюційні алгоритми застосовні в нейронних мережах.

3.1.7 Розробка алгоритму прискорення збіжності вектору коефіцієнтів логістичної регресії

Суть пропонованого алгоритму: (1) аналітичним методом обчислюється вектор коефіцієнтів \mathbf{c}_{LS} лінійної регресії $\hat{y}(i) = \mathbf{X}(i)\mathbf{c}_{LS}$; (2) чисельним методом обчислюється вектор коефіцієнтів \mathbf{c}_{MLE} логістичної регресії $\tilde{y}(i) = \frac{1}{1 + e^{-\mathbf{X}(i)\mathbf{c}_{MLE}}}$, беручи за початкову точку вектор лінійної регресії: $\mathbf{c}_{initial} = \mathbf{c}_{LS}$. Висувається гіпотеза: якщо не брати до уваги зміщення моделей (*intercept*, вільний член), то вектори коефіцієнтів мають дуже близький напрямок. Відносно зміщення, то для моделей без параметрів, очевидно, що $intercept_{LS} = \frac{1}{1 + e^{-intercept_{MLE}}} = p(y = 1)$, тому можливе вдосконалення: $intercept_{initial} = \ln(intercept_{LS} / (1 - intercept_{LS}))$.

3.1.8 Вдосконалення, а саме комплексне узагальнення логістичної регресії та ваги категорії змінної для ймовірнісної цільової змінної

Для запропонованого узагальнення ваги категорії змінної (WoE) та інформаційної статистики (IV) вводиться така подвійна нумерація – (i : номер категорії (інтервалу) змінної, j : внутрішній номер в межах категорії), тобто тоді y_{ij} – j -те значення цільової змінної в межах i -тої категорії вхідної змінної, а n_i – кількість елементів навчальної вибірки для i -тої категорії. Суть узагальнення:

$$\begin{aligned}
 WoE_i &= \ln\left(\frac{g_i}{b_i}\right) = \ln\left(\frac{\frac{G_i}{B_i}}{\frac{G}{B}}\right) = \ln\left(\frac{\frac{\frac{G_i}{\sum_{t=1}^k G_t}}{\frac{B_i}{\sum_{t=1}^k B_t}}}{\frac{\sum_{t=1}^k G_t}{\sum_{t=1}^k B_t}}}\right) = \ln\left(\frac{G_i}{\sum_{t=1}^k G_t}\right) - \ln\left(\frac{B_i}{\sum_{t=1}^k B_t}\right), \\
 WoE_i^* &= \ln\left(\frac{\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{p=1}^k \sum_{q=1}^{n_p} y_{pq}}}{\frac{\sum_{j=1}^{n_i} (1-y_{ij})}{\sum_{p=1}^k \sum_{q=1}^{n_p} (1-y_{pq})}}}\right) = \ln\left(\frac{G_i^*}{\sum_{p=1}^k G_p^*}\right) - \ln\left(\frac{B_i^*}{\sum_{p=1}^k B_p^*}\right), \\
 IV &= \sum_{i=1}^k (g_i - b_i) WoE_i, \\
 IV^* &= \sum_{i=1}^k \left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{p=1}^k \sum_{q=1}^{n_p} y_{pq}} - \frac{\sum_{j=1}^{n_i} (1-y_{ij})}{\sum_{p=1}^k \sum_{q=1}^{n_p} (1-y_{pq})} \right) WoE_i^* = \sum_{i=1}^k \left(\frac{G_i^*}{\sum_{p=1}^k G_p^*} - \frac{B_i^*}{\sum_{p=1}^k B_p^*} \right) WoE_i^*.
 \end{aligned}$$

Очевидно, що G_i^* , B_i^* , WoE_i^* , IV^* – відповідно узагальнені значення [13, 43].

Узагальнення логістичної регресії (логарифма функції правдоподібності), де $P(\mathbf{c}, \mathbf{X}(i)) = \frac{1}{1 + e^{-\mathbf{X}(i)\mathbf{c}}}$, що приводиться автором дисертації детальніше в [43]:

$$\ln L(\mathbf{c}) = \sum_{i=1}^N (I_{\{1\}}(y_i) \ln P(\mathbf{c}, \mathbf{X}(i)) + I_{\{0\}}(y_i) \ln(1 - P(\mathbf{c}, \mathbf{X}(i)))) ,$$

$$\ln L(\mathbf{c}) = \sum_{i=1}^N (y_i \ln P(\mathbf{c}, \mathbf{X}(i)) + (1 - y_i) \ln(1 - P(\mathbf{c}, \mathbf{X}(i)))) ,$$

$$\ln L^*(\mathbf{c}) = \sum_{i=1}^N (w(y_i) \ln P(\mathbf{c}, \mathbf{X}(i)) + (1 - w(y_i)) \ln(1 - P(\mathbf{c}, \mathbf{X}(i)))) ,$$

$$L^*(\mathbf{c}) = \prod_{i=1}^N (P(\mathbf{c}, \mathbf{X}(i))^{w(y_i)} (1 - P(\mathbf{c}, \mathbf{X}(i)))^{1-w(y_i)}) ,$$

$$\left\{ \begin{array}{l} w(1) = 1; \\ w(0) = 0; \\ \forall y \in [0;1]: w(y) \in [0;1]; \\ \forall y_i \in [0;1] \forall y_j \in [0;1]: y_i > y_j \Rightarrow w(y_i) > w(y_j). \end{array} \right.$$

У роботі [43] автор дисертації доводить, що для $\ln L^*(\mathbf{c})$:

$$\mathbf{g}(\mathbf{c}) = \frac{d \ln L^*(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^N (w(y_i) - P(\mathbf{c}, \mathbf{X}(i))) \mathbf{X}^T(i),$$

тобто при $w(y) = y$ отримуємо класичну формулу градієнта, яка застосовна як для бінарного, так і для ймовірнісного випадку цільової змінної. Формула матриці Гессе не залежить від вагової функції (наприклад, $w_\alpha(y) = y^\alpha$):

$$\mathbf{H}(\mathbf{c}) = - \sum_{i=1}^N P(\mathbf{c}, \mathbf{X}(i)) (1 - P(\mathbf{c}, \mathbf{X}(i))) \mathbf{X}^T(i) \mathbf{X}(i).$$

3.1.9 Вдосконалення методу k -найближчих сусідів

Перший етап – «базовий метод k -plus-найближчих сусідів» [11]:

а) відносно метрики (де $WoE(i, p)$ – вага категорії i -ї змінної для p -того вектору спостережень):

$$\mathbf{x}_p^T = (WoE(1, p) \quad WoE(2, p) \quad \dots \quad WoE(i, p) \quad \dots \quad WoE(M, p)),$$

$$d(\mathbf{x}_p, \mathbf{x}_r) = \|\mathbf{x}_p - \mathbf{x}_r\| = \sqrt{\sum_{i=1}^M (WoE(i, p) - WoE(i, r))^2};$$

б) відносно використання параметру k та щодо ймовірнісного прогнозу: $k^+(\mathbf{x}^*) \geq k$ – фактична кількість найближчих сусідів (не менше k) для вектору \mathbf{x}^* (використовуючи правило включення «хвоста» – всіх елементів, які мають відстань від \mathbf{x}^* точно рівну відстані від \mathbf{x}^* до останнього, тобто k -того, сусіда):

$$y^* = p(y=1 | \mathbf{x}^*) = \sum_{i=1}^{k^+(\mathbf{x}^*)} \frac{y_i}{k^+(\mathbf{x}^*)};$$

в) критерій якості апроксимації навчальної вибірки (при цьому тестову можна не створювати) – індекс Джині, отриманий за допомогою перехресного тестування – перехресної валідації (*cross-validation*) – на навчальній вибірці методом «*leave-one-out*», коли для кожного елемента ставиться прогноз за допомогою моделі перерахованої на інших $N - 1$ елементах:

$$GINI = \left(\int_{y^* \in Y^*} F_B(y^*) dF_G(y^*) - \frac{1}{2} \right) / \left(\frac{1}{2} \right).$$

Включення «хвоста» в MS SQL (T-SQL): SELECT TOP(@k) WITH TIES ...

Недоліки класичного методу k -найближчих сусідів, що вирішуються:

1) відсутність конкретики та деталізації особливостей застосування при використанні категоріальних та дискретизованих змінних, тому часто використовується метод створення бінарних фіктивних змінних (*dummy variables*), що призводить до збільшення сукупної кількості змінних в процесі моделювання за рахунок фіктивних;

2) відсутність конкретики та деталізації особливостей застосування в умовах можливості виникнення ситуацій з рівновіддаленими групами найближчих сусідів відносно елемента, що класифікується, тобто якщо останній найближчий сусід входить у рівновіддалену групу елементів вибірки;

3) класичне правило класифікації по принципу значення більшості (*majority*) дає детермінований, а не ймовірнісний прогноз;

4) проблема вибору метричного простору та власне метрики;

5) відсутність чіткого критерію оптимальності моделі класифікації та методу його застосування для вибору оптимального параметру – оптимальної кількості числа найближчих сусідів k (проблема вибору k);

б) питання рівноцінності округлення при парних k .

Другий етап вдосконалення – «повний метод k -plus-найближчих сусідів» (обчислювальна складність: $O(N^2 \dim(\mathbf{x}))$, тобто $O(N^2M)$) [11] – зображено на рисунку 3.4 у вигляді блок-схеми. Для $k = N - 1$ перехресна валідація тривіальна («антикласифікатор») [11].

Результати подолання недоліків класичного методу: (1) для оперування категоріальними змінними використовуються WoE-перетворення; (2) запропоновано урахування групи рівновіддалених елементів, у яку входить останній найближчий сусід; (3) запропоновано просту частотну оцінку по всіх врахованих елементах у якості ймовірнісного прогнозу; (4) запропоновано

використання евклідової метрики у просторі WoE-перетворень категоріальних та дискретизованих змінних; (5) критерій оптимальності для вибору k – індекс Джині для перехресного тестування; (6) ймовірнісна оцінка не потребує округлень.

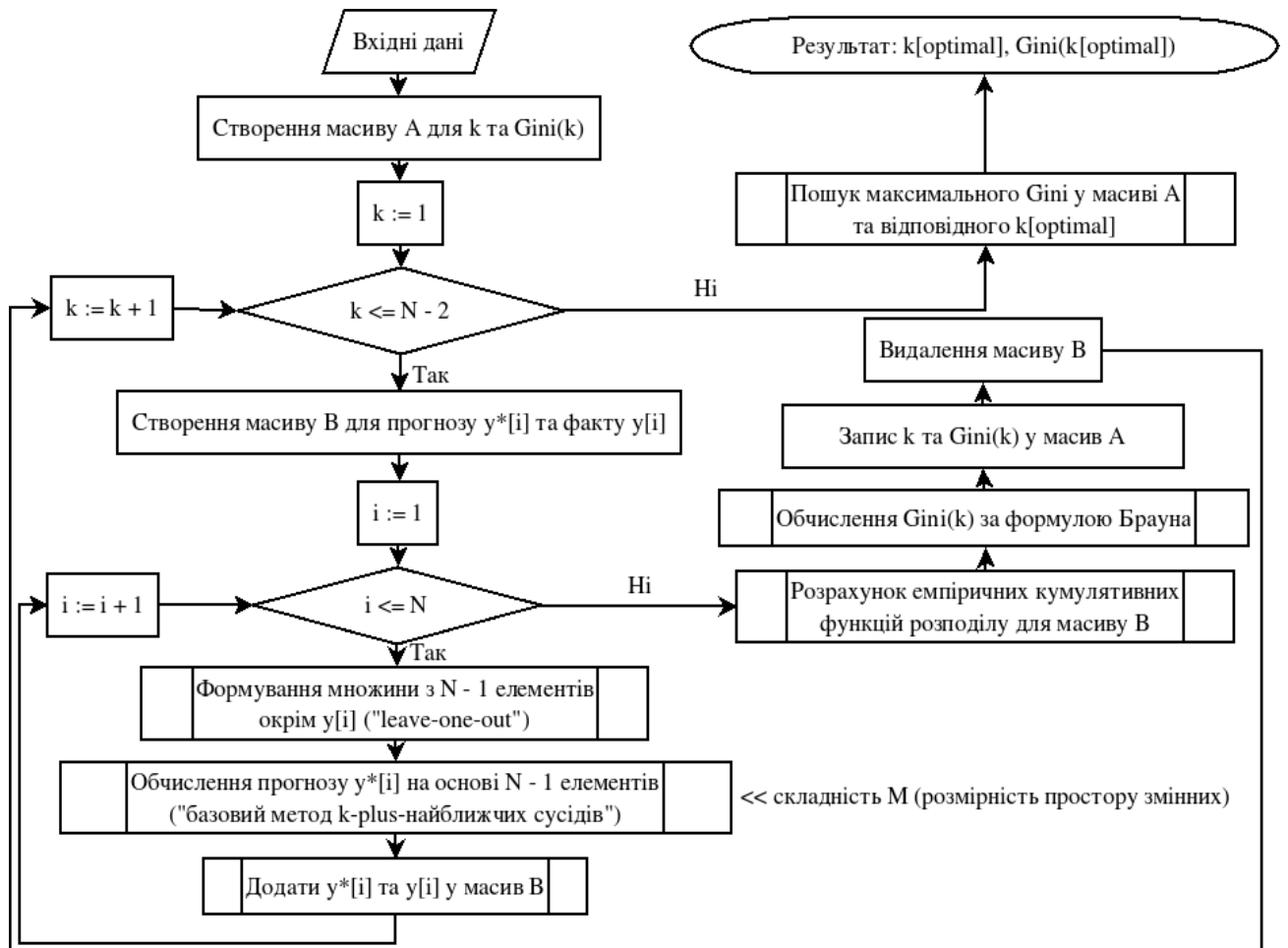


Рисунок 3.4 – Блок-схема другого вдосконалення методу k -найближчих сусідів

3.1.10 Калібрування карти балів відносно логарифму відношення шансів

Приклад калібрування шкали $score(\mathbf{x}) = A \ln(odds(\mathbf{x})) + B$ [8, 9] з вимогами збільшення шансів в K разів при прирості на кожні D балів та значенні $score_{fixed}$ при $odds_{fixed}$:

$$\begin{cases} score(\mathbf{x}) = A \ln(odds(\mathbf{x})) + B, \\ score(\mathbf{x}) + D = A \ln(K \cdot odds(\mathbf{x})) + B, \\ score_{fixed} = A \ln(odds_{fixed}) + B; \end{cases}$$

$$\begin{cases} A = \frac{D}{\ln(K)}, \\ B = score_{fixed} - A \ln(odds_{fixed}) = score_{fixed} - \frac{D}{\ln(K)} \ln(odds_{fixed}). \end{cases}$$

При подвоєнні шансів кожені 20 балів та відношенні 3,5 для 200 балів $A = \frac{20}{\ln(2)} \approx 28,8539$ та $B = 200 - \frac{20}{\ln(2)} \ln(3,5) \approx 163,8529$. Тут $odds(\mathbf{x}) = \mathbf{c}_{MLE}^T \mathbf{x}_{obs}$.

3.2 Методи включення та аналізу відхилених заявок

У підрозділі 1.4 детально проаналізовано основні поняття, суть проблеми та мету аналізу відхилених заявок. Зокрема, вказано, що аналіз відхилених заявок (reject inference) – це процес присвоєння елементам вибірки, по яким невідоме бінарне значення цільової змінної, деякого значення цільової змінної та подальше включення таких елементів вибірки у навчальну вибірку моделі з метою забезпечення стійкості фінальної моделі відносно генеральної сукупності.

У даному підрозділі розглядаються конкретні методи включення та аналізу відхилених заявок.

3.2.1 Метод присвоєння нульового результату відхиленим заявкам

Метод присвоєння нульового результату (негативного класу) відхиленим заявкам (assigning all rejects to bads) [8, 9] є примітивним методом, що іноді використовується, якщо твірні для вибірки рішення були точними або бажаними.

3.2.2 Метод присвоєння аналогічної пропорції

Метод присвоєння аналогічної пропорції (assigning rejects in the same proportion of goods to bads as reflected in the acceptees) [8, 9], який полягає в присвоєнні відхиленим заявкам співвідношення класів рівного співвідношенню для прийнятих заявок, використовується при непослідовній системі відбору, коли рішення приймалися випадковим чином при формуванні вибірки.

3.2.3 Метод повного ігнорування відхилених заявок

Метод повного ігнорування відхилених заявок (ignoring the rejects altogether) [8, 9] полягає у побудові моделі лише на прийнятих заявках. При цьому коректним впровадженням такої моделі є [8, 9]: 1) застосування спочатку поточної системи прийняття рішень для первинного відбору, який відповідає навчальній вибірці; 2) подальше застосування побудованої моделі на підмножині.

3.2.4 Метод тимчасового погодження всіх заявок для збору даних

Метод тимчасового погодження всіх заявок для накопичення даних (approving all applications) [8, 9] – єдиний метод фіксування фактичної поведінки клієнтів за заявками, що мають бути відхиленими. Даний метод ще називається «купівлею даних» («buying data»), оскільки передбачає одобрення заявок з високим ризиком. Стратегії тимчасового накопичення даних повинні рівноцінно враховувати всі річні періоди у випадку, коли має місце сезонність.

3.2.5 Метод використання аналогічних даних банку або кредитних бюро

Метод використання аналогічних даних банку або кредитних бюро (similar in-house or bureau data based method) [8, 9] полягає у використанні поведінкових даних про клієнтів у ситуаціях, коли заявникові було відмовлено в отриманні кредиту досліджуваного продукту, але було видано кредит по аналогічному продукту того ж кредитора, або, коли заявник отримав відмову у даного кредитора, але невдовзі отримав аналогічний кредит у іншого, поведінка за яким фіксується у кредитних бюро. Тобто цільова змінна для відхилених заявок обчислюється виходячи з поведінкових даних по аналогічному кредиту (similar credit), отриманого в аналогічні строки (similar time frame) невдовзі після відмови.

3.2.6 Метод доповнення для експертного процесу прийняття рішень

Метод доповнення вибірки в умовах експертного процесу прийняття рішень або «м'якого відсікання» (augmentation in historically judgmental decision-making environment (soft cutoff)) [8, 9] полягає власне у попередній побудові *скорингової моделі* з цільовою змінною прийняття/відхилення (accept/reject), яка надалі застосовується на всю множину прийнятих та відхилених заявок, де потім встановлюються інтервали скорингу прийняття/відхилення, а для кожного такого інтервалу балів розраховується коефіцієнт доповнення (augmentation factor) для прийнятих заявок, що дорівнює відношенню кількості всіх заявок в інтервалі до кількості прийнятих заявок в інтервалі: $F_k = \frac{A_k + R_k}{A_k}$. Надалі метод передбачає формування нової доповненої вибірки, що формується шляхом включення з оригінальної вибірки лише прийнятих заявок кожної прийнятої заявки з її цільовим поведінковим результатом (good/bad) n_i разів, де кількість входжень дорівнює або пропорційна коефіцієнту доповнення. Тоді залежність відношення шансів доповненої вибірки є випуклою комбінацією оригінального відношення

шансів прийнятих заявок і отриманого відношення відхилених заявок ($\alpha \in [0;1]$):

$$\frac{AG}{AB} = \alpha \frac{G}{B} + (1-\alpha) \frac{AG-G}{AB-B}. \text{ Зазвичай } \frac{odds_{accepted}}{odds_{rejected}^*} = \frac{AG(AB-B)}{AB(AG-G)} \in [1,5; 4,0].$$

У даному методі може також використовуватися зважена логістична регресія, що описується в підрозділі 3.2.10, для *оригінальної вибірки прийнятих заявок* з вагами рівними коефіцієнтам доповнення.

3.2.7 Метод простого доповнення («жорстке відсікання»)

Метод простого доповнення або «жорстке відсікання» (simple augmentation (hard cutoff)) [8, 9] полягає у включенні відхилених заявок класифікованих за допомогою моделі на прийнятих заявках і певного порогу відсікання (cut-off rate).

3.2.8 Метод доповнення на основі ймовірностей погодження заявок

Даний метод передбачає попередню побудову допоміжної скорингової моделі з цільовою змінною прийняття/відхилення (accept/decline model) подібно як у підрозділі 3.2.6, але далі допоміжна модель застосовується на лише множині прийнятих заявок для визначення $p(\text{accepted} | \mathbf{x})$ для кожної прийнятої заявки. Доповнена вибірка складається з прийнятих заявок включених пропорційно ненормованій вазі $w(\mathbf{x}) = p^{-1}(\text{accepted} | \mathbf{x})$ обчисленої як w_i окремо для кожної прийнятої заявки, що включається з даною вагою у нову доповнену вибірку [8, 9].

У даному методі може також використовуватися зважена логістична регресія, що описується в підрозділі 3.2.10, для *оригінальної вибірки прийнятих заявок* з вагами $w(\mathbf{x}) = p^{-1}(\text{accepted} | \mathbf{x})$ обернено ймовірностям прийняття заявок.

3.2.9 Метод розбиття груп ризику відхилених заявок

Метод розбиття (parceling) груп ризику відхилених заявок полягає у дослідженні розподілу скорингових груп моделі побудованої лише на прийнятих заявках (*known good/bad*) при застосуванні на множині відхилених заявок. Далі, виходячи з фактичної долі нульових елементів (bad rate) для кожної групи балів на прийнятих заявках, серед відхилених заявок у скоринговій групі випадковим чином обираються $B_{rejected}(j) = p_{accepted}(y = 0 | j)N_{rejected}(j)$ «поганих» (нульових) елементів, інші вважаються «хорошими» (одиничними). Приклад застосування наведено в таблиці 3.1 (група 190–199: **733** виводяться як **10%** від **7334** rej.) [8, 9].

Таблиця 3.1 – Приклад методу розбиття для аналізу відхилених заявок [8, 9]

Бал	Bad	Good	% Bad	% Good	Rejects	Rej. Bad	Rej. Good
0–169	290	971	23,0%	77,0%	1646	379	1267
170–179	530	2414	18,0%	82,0%	1732	312	1420
180–189	365	2242	14,0%	86,0%	3719	521	3198
190–199	131	1179	10,0%	90,0%	7334	733	6601
200–209	211	2427	8,0%	92,0%	1176	94	1082
210–219	213	4047	5,0%	95,0%	3518	176	3342
220–229	122	2928	4,0%	96,0%	7211	288	6923
230–239	139	6811	2,0%	98,0%	3871	77	3794
240–249	88	10912	0,8%	99,2%	4773	38	4735
250+	94	18706	0,5%	99,5%	8982	45	8937

Вдосконаленням даного методу є завищення очікуваного значення долі нульових елементів (exprected bad rate) для відхилених заявок відносно долі на прийнятих заявках таким чином, щоб, наприклад, загальна доля нульових елементів (overall bad rate) на відхилених заявках була в 2–4 рази вищою ніж на

прийнятих заявках [8, 9]. Метою такого вдосконалення є уникнення недооцінки реальної долі нульових елементів на відхилених заявках [8, 9].

3.2.10 Метод нечіткого доповнення

Метод нечіткого доповнення (fuzzy augmentation) [8, 9, 92–94] власне у класичному формулюванні полягає у використанні зваженої логістичної регресії для множини, в яку входять прийняті заявки з одиничним ваговим коефіцієнтом, а кожна відхилена заявка входить у вибірку *2 рази*: з вагою рівною ймовірності «хорошого» (одиничного) цільового індикатора $p_{i1} = p(y=1 | \mathbf{x})$ з *одиничним цільовим результатом*, а також з вагою рівною ймовірності «поганого» (нульового) цільового індикатора $p_{i0} = p(y=0 | \mathbf{x}) = 1 - p(y=1 | \mathbf{x}) = 1 - p_{i1}$ з *нульовим цільовим результатом*. Прогнозні ймовірності (ваги) отримуються за допомогою моделі побудованої лише на прийнятих заявках (known good/bad).

Бінарна зважена логістична регресія (*weighted logistic regression*) описується зваженою функцією правдоподібності [92]:

$$L(\mathbf{c}) = \prod_{i=1}^{N^*} (P(\mathbf{c}, \mathbf{X}(i)))^{w_i y_i} (1 - P(\mathbf{c}, \mathbf{X}(i)))^{w_i (1 - y_i)} .$$

Для часткового випадку, методу нечіткого доповнення, зважену логістичну регресію можна замінити на узагальнену в підрозділі 3.1.8 без потреби розбиття кожної відхиленої заявки на 2 заявки з цільовими результатами 0 та 1, оскільки:

$$\begin{aligned} & \left. (P(\mathbf{c}, \mathbf{X}(i)))^{w_i y_i} (1 - P(\mathbf{c}, \mathbf{X}(i)))^{w_i (1 - y_i)} \right|_{\substack{y_i=1, \\ w_i=p_{i1}}} \cdot \left. (P(\mathbf{c}, \mathbf{X}(i)))^{w_i y_i} (1 - P(\mathbf{c}, \mathbf{X}(i)))^{w_i (1 - y_i)} \right|_{\substack{y_i=0, \\ w_i=p_{i0}}} \\ & \Rightarrow (P(\mathbf{c}, \mathbf{X}(i)))^{p_{i1}} \cdot 1 \cdot 1 \cdot (1 - P(\mathbf{c}, \mathbf{X}(i)))^{1 - p_{i1}} \end{aligned}$$

$$\Rightarrow (P(\mathbf{c}, \mathbf{X}(i)))^{y_i} (1 - P(\mathbf{c}, \mathbf{X}(i)))^{1-y_i} \Big|_{y_i = \tilde{y}_i = p_{i1}} .$$

Модифікацією методу є такий перерахунок ваг з дільником для odds [94]:

$$w_{i1}^* = p_{i1} \cdot W \cdot \text{EventRateIncrease} ,$$

$$w_{i0}^* = p_{i0} \cdot W ,$$

$$W = \frac{r}{1-r} \frac{N_A}{N_R^*} = \frac{N_R / (N_A + N_R)}{1 - (N_R / (N_A + N_R))} \frac{N_A}{\sum_{i: \text{Rejects}} (p_{i0} + p_{i1})} .$$

Суть класичного методу нечіткого доповнення з використанням зваженої логістичної регресії представлено на рисунку 3.5. Суть вдосконаленого методу нечіткого доповнення, де власне вдосконаленням є використання узагальненої логістичної регресії, зображено на рисунку 3.6.



Рисунок 3.5 – Суть класичного методу нечіткого доповнення

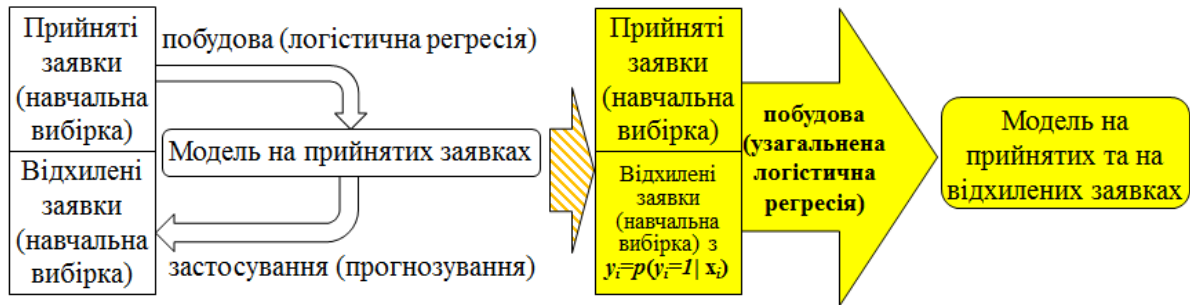


Рисунок 3.6 – Суть вдосконалення методу нечіткого доповнення

При використанні у класичному методі та його вдосконаленні значень узагальнених WoE та їх переоцінці на етапі включення відхилених заявок, отримується однакова фінальна модель, але у другому випадку без необхідності дублювання відхилених заявок з різними цільовими результатами. Зображене вдосконалення відповідає нульовій та першій ітерації вдосконаленого методу ітеративної класифікації, описаного далі у підрозділі 3.2.15, що також додатково підтверджує актуальність вдосконалення методу ітеративної класифікації.

3.2.11 Метод ітеративної класифікації

Метод ітеративної класифікації (iterative reclassification) [8, 9] полягає у побудові спочатку моделі тільки на прийнятих заявках (known good/bad) з подальшою бінарною класифікацією відхилених заявок, включенням відхилених заявок та перебудовою моделі, повторною бінарною класифікацією відхилених заявок, повторним включенням оновлених відхилених заявок перебудовою і т.д. Класифікація відхилених заявок та перебудова з включенням *класифікованих відхилених* та прийнятих заявок виконуються ітеративно до збіжності класів [95].

Суть типового методу ітеративної класифікації зображена на рисунку 3.7. При фіксованому тут порозі відсікання «cut-off rate № 0» (тобто обчисленому на нульовому кроці, наприклад) прогноз на відхилених заявках зазвичай стає

«сингулярним», тобто відхилені заявки відносяться до одного класу (як правило, bad). Тому доцільним є нефіксований поріг відсікання або очікуваний «bad rate».

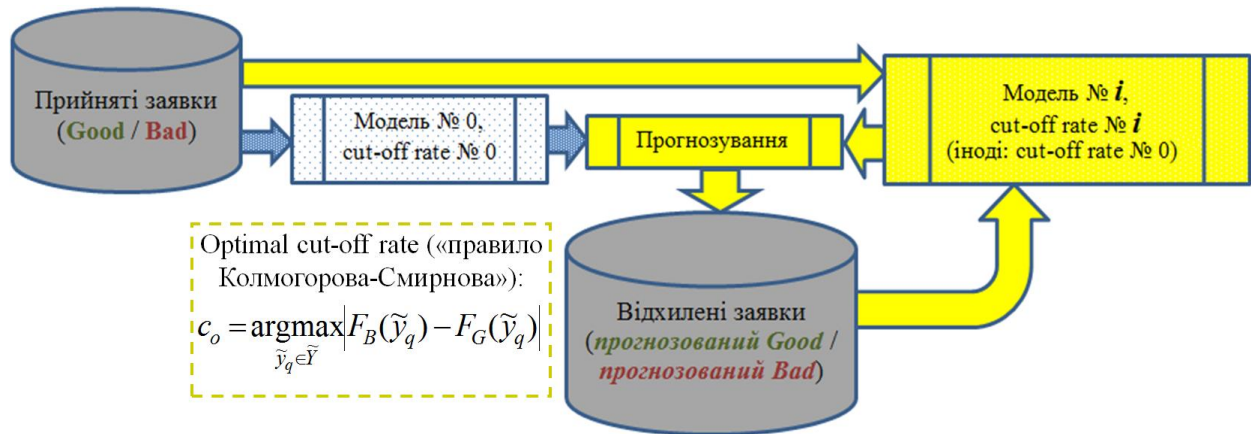


Рисунок 3.7 – Суть типового методу ітеративної класифікації

3.2.12 Метод найближчого сусіда (кластеризація)

Метод (nearest neighbor (clustering)) полягає у включенні відхилених заявок з результатом згідно з методом k -найближчих сусідів при $k = 1$ (див. підрозділ 3.1.2.2).

3.2.13 Метод виводу на основі пам'яті

Метод виводу на основі пам'яті (memory-based reasoning або case-based reasoning) полягає у використанні методу k -найближчих сусідів (підрозділ 3.1.2.3) з модифікацією впровадження ймовірнісної класифікації [8, 9]:

$$\begin{cases} \mathbf{x}_1 = \operatorname{arg min}_{\mathbf{x} \in X} d(\mathbf{x}, \mathbf{x}^*); \\ \forall i \in \{2, \dots, k\} : \mathbf{x}_i = \operatorname{arg min}_{\mathbf{x} \in X \setminus \bigcup_{j=1}^{i-1} \mathbf{x}_j} d(\mathbf{x}, \mathbf{x}^*), \end{cases}$$

$$\begin{cases} p(y^* = 1 | \mathbf{x}^*) = \sum_{i=1}^k \frac{y_i}{k}; \\ p(y^* = 0 | \mathbf{x}^*) = 1 - p(y^* = 1 | \mathbf{x}^*). \end{cases}$$

Далі на множині всіх заявок може використовуватися зважена логістична регресія з розбиттям кожної відхиленої заявки на дві заявки з ймовірнісними вагами, аналогічно нечіткому доповненню (див. підрозділ 3.2.10). Іншим варіантом даного методу є попередній перехід від ймовірнісної класифікації до детермінованої, але вже не по правилу більшості (див. підрозділ 3.1.2.3), а, наприклад, з використанням порогу для ймовірності $p(y^* = 0 | \mathbf{x}^*)$:

$$y^* = \begin{cases} 1, & p(y^* = 0 | \mathbf{x}^*) = 1 - \sum_{i=1}^k \frac{y_i}{k} > p_{\text{bad cut-off}}; \\ 0, & p(y^* = 0 | \mathbf{x}^*) = 1 - \sum_{i=1}^k \frac{y_i}{k} \leq p_{\text{bad cut-off}}. \end{cases}$$

3.2.14 Інші методи аналізу відхилених заявок

До інших методів аналізу відхилених заявок відносяться зокрема використання алгоритму максимізації математичного сподівання (expectation-maximization algorithm) – EM-алгоритму – для оцінювання змішаного розподілу (mixture distributions) двох окремих бінарних класів [14, 15] та модель Хекмана (Heckman's model) для біваріаційного розподілу [14]. При цьому EM-алгоритм для оцінювання змішаного (mixture) розподілу використовується для вибірки, що сформована по принципу відносно випадкового відхилення (Missing At Random, MAR), коли $p(a=1 | \mathbf{x}, y) = p(a=1 | \mathbf{x})$ або $p(y=1 | \mathbf{x}, a=1) = p(y=1 | \mathbf{x}, a=0) = p(y=1 | \mathbf{x})$ [13–15], а модель Хекмана використовується для вибірки, що сформована по принципу невідповідного відхилення (Missing Not At Random, MNAR), тобто

коли $p(a = 1 | \mathbf{x}, y) \neq p(a = 1 | \mathbf{x})$ або $p(y = 1 | \mathbf{x}, a = 1) \neq p(y = 1 | \mathbf{x}, a = 0)$ [13–15]. Також ситуація, що відповідає абсолютно випадковому формуванню вибірки (Missing Completely At Random, MCAR), коли $p(a = 1 | \mathbf{x}, y) = p(a = 1)$ [13–15], не потребує аналізу відхилених заявок, модель на прийнятих заявках є незміщеною (*unbiased*).

3.2.15 Вдосконалення методу ітеративної класифікації

Пропонований «метод ітеративного обчислення ймовірностей» (рисунок 3.8 та рисунок 3.9) як *вдосконалення методу ітеративної класифікації* [13] полягає у використанні комплексних узагальнень логістичної регресії (на вищому рівні моделі) та ваг категорій змінних (Weight of Evidence, WoE) (на нижчому рівні моделі) для неперервної цільової змінної, що набуває ймовірнісних значень з неперервного інтервалу [0%; 100%], що описано в підрозділі 3.1.8. Описані раніше узагальнення дозволяють перейти від моделі типу «бінарний факт, ймовірнісний прогноз» до моделі типу «ймовірнісний факт, ймовірнісний прогноз». Узагальнення інформаційної статистики (Information Value) дозволяє переоцінити предикативність (див. підрозділ 3.1.8). Пропонований критерій збіжності ітерацій – це відстань Чебишева між послідовністю цільових ймовірностей $p(y = 1 | \mathbf{x})$ для відхилених заявок при навчанні моделі (на вході моделі) і послідовністю прогнозованих ймовірностей $p'(y = 1 | \mathbf{x})$ для відхилених заявок при зворотному прогнозуванні (на виході моделі), тобто максимальна абсолютна різниця між прогнозами послідовних моделей на відхилених заявках: $l_n = \max_{i \in \text{Rejects}} |\hat{p}_n(y_i = 1) - \hat{p}_{n-1}(y_i = 1)|$. Одна з переваг при використанні узагальненої моделі та ймовірнісного прогнозування – це відсутність необхідності обчислень “cut-off rate” та дуальність відхилених заявок: кожна відхилена заявка є одночасно нульовою та одиничною з різними ймовірностями при навчанні моделі [13].

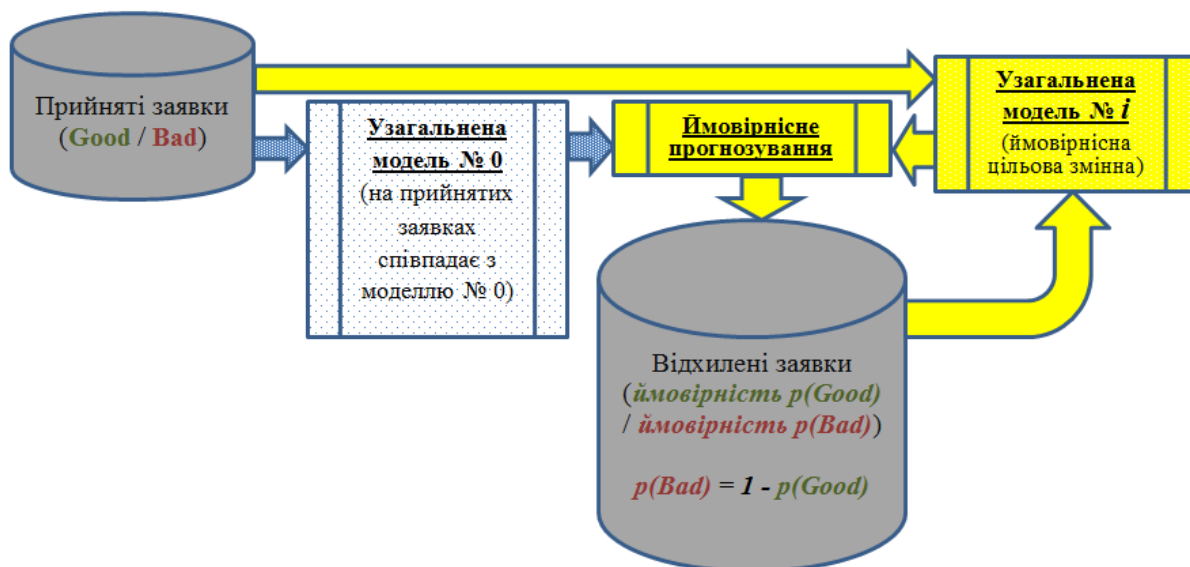


Рисунок 3.8 – Вдосконалення методу ітеративної класифікації

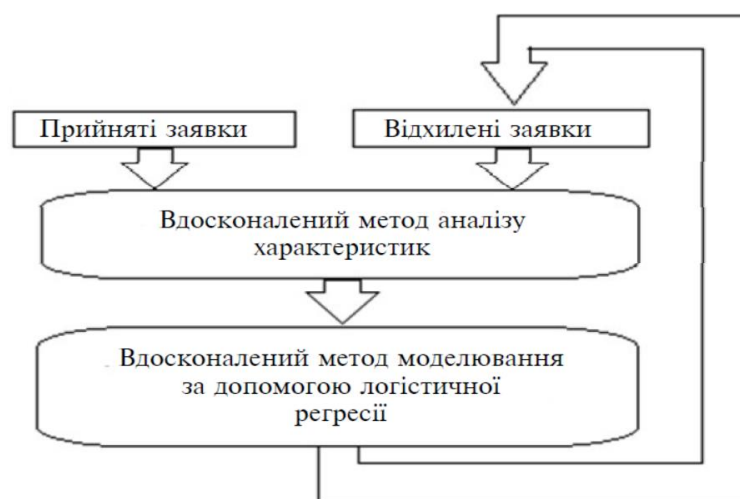


Рисунок 3.9 – Суть «методу ітеративного обчислення ймовірностей» [13]

Вдосконалення вирішує такі недоліки методу ітеративної класифікації:

- 1) бінарність та однозначність проміжних та остаточно виведених значень цільової змінної для відхилених заявок;
- 2) необхідність обчислення та застосування порогу відсікання (cut-off rate) в процесі аналізу відхилених заявок;
- 3) завелике зміщення співвідношення класів (зазвичай, завищення доли нульового класу – bad rate) залежне від методу вибору cut-off rate [3].

У розділі 4 експериментально встановлено, що вдосконалений метод збігається за меншу кількість ітерацій та спричиняє менші відхилення по Джині, статистиці Колмогорова-Смирнова від моделі без урахування відхилених заявок.

Таким чином, вдосконалена процедура подолання **неповноти** має вигляд:

$$L_n^*(\mathbf{c}) = \prod_{i \in \text{Accepts}} (P(\mathbf{c}, \mathbf{X}_{n-1}(i)))^{y_i} (1 - P(\mathbf{c}, \mathbf{X}_{n-1}(i)))^{1-y_i} \cdot \prod_{i \in \text{Rejects}} (P(\mathbf{c}, \mathbf{X}_{n-1}(i)))^{\hat{p}_{n-1}(y_i=1)} (1 - P(\mathbf{c}, \mathbf{X}_{n-1}(i)))^{1-\hat{p}_{n-1}(y_i=1)}.$$

3.3 Методи оцінювання якості прогнозів та аналіз стійкості моделей

У даному підрозділі описуються методи аналізу придатності скорингових моделей щодо їх впровадження, що включають: 1) методи аналізу предикативної сили моделей (за допомогою класичних показників: індексу Джині, статистики Колмогорова-Смирнова, відстані Махаланобіса, статистики Хосмера-Лемешоу); 2) методи аналізу незміщеності розподілів області застосування моделі відносно розподілів навчальної вибірки (індекс стійкості розподілу, звіти зміщення балу, можливості застосування статистики Колмогорова-Смирнова). Також запропоновано ефективні програмні реалізації та узагальнення показників.

3.3.1 Оцінка якості прогнозів скорингових моделей

Оцінка якості прогнозів скорингових моделей по суті є оцінюванням предикативної (прогностичної) сили фінальних моделей, що впроваджуються. Даний етап передбачає підрахунок показників якості бінарних класифікаторів [1]: індексу Джині, статистики Колмогорова-Смирнова, відстані Махаланобіса, іноді статистики Хосмера-Лемешоу, логарифма функції правдоподібності (підрозділ

3.1.1.2), коефіцієнта детермінації, визначеного не лише для бінарних моделей і т.д.

3.3.1.1 Методи розрахунку показника Джині

Для індексу Джині бінарного класифікатора існує щонайменше три методи обчислення, результати яких співпадають:

- 1) класичний аналіз кривої операційної характеристики приймача (Receiver Operating Characteristic, ROC);
- 2) аналіз кривої Лоренца щодо накопичення класу відсортованої вибірки;
- 3) аналіз кривої Лоренца функцій кумулятивного розподілу двох класів.

3.3.1.1.1 Операційна характеристика приймача (ROC-крива)

Крива операційної характеристики приймача будується як параметричний графік по точках порогу відсікання – унікальних значеннях прогнозованого балу або ймовірності. По осі ординат розміщується чутливість (*Sensitivity, Se*) як доля правильно класифікованих *одиночних* елементів від усіх *одиночних* елементів вибірки, по осі абсцис – 100% за вирахуванням долі правильно класифікованих *нульових* елементів від усіх *нульових* елементів вибірки (*Specificity, Sp*), тобто доля неправильно класифікованих *нульових* елементів від усіх *нульових* елементів вибірки (*False Positives Rate, FPR*) [8, 9]. Матриця спряженості для деякого зафіксованого порогу відсікання c (*cut-off rate*) зображена у вигляді таблиці 3.2. При цьому означена помилка I-го роду відповідає прямим втратам при кредитуванні некредитоспроможних позичальників, а помилка II-го роду відповідає недоотриманому прибутку у процесі кредитування, коли відмовлено кредитоспроможним позичальникам у наданні кредиту.

Таблиця 3.2 – Матриця спряженості (*confusion matrix* або *contingency table*)

Матриця спряженості для помилок I-го та II-го роду		Фактичне значення (тестова вибірка)	
		Одиничне	Нульове
Прогнозне значення (модель)	Одиничне	<i>TP (True Positives)</i>	<i>FP (False Positives)</i> помилка I-го роду: «помилкове виявлення»
	Нульове	<i>FN (False Negatives)</i> помилка II-го роду: «помилковий пропуск»	<i>TN (True Negatives)</i>

Очевидно, формули обчислення вищезазначених показників мають вигляд:

$$Se(c) = TPR(c) = \frac{TP(c)}{FN(c) + TP(c)} = \frac{\sum_{i=1}^N y_i I(\tilde{y}_i > c)}{\sum_{i=1}^N y_i},$$

$$Sp(c) = TNR(c) = \frac{TN(c)}{FP(c) + TN(c)} = \frac{\sum_{i=1}^N (1 - y_i) I(\tilde{y}_i \leq c)}{\sum_{i=1}^N (1 - y_i)},$$

$$FPR(c) = \frac{FP(c)}{FP(c) + TN(c)} = \frac{\sum_{i=1}^N (1 - y_i) I(\tilde{y}_i > c)}{\sum_{i=1}^N (1 - y_i)} = 1 - Sp(c).$$

Приклад графіку ROC-кривої на тестовій вибірці наведено на рисунку 3.10.

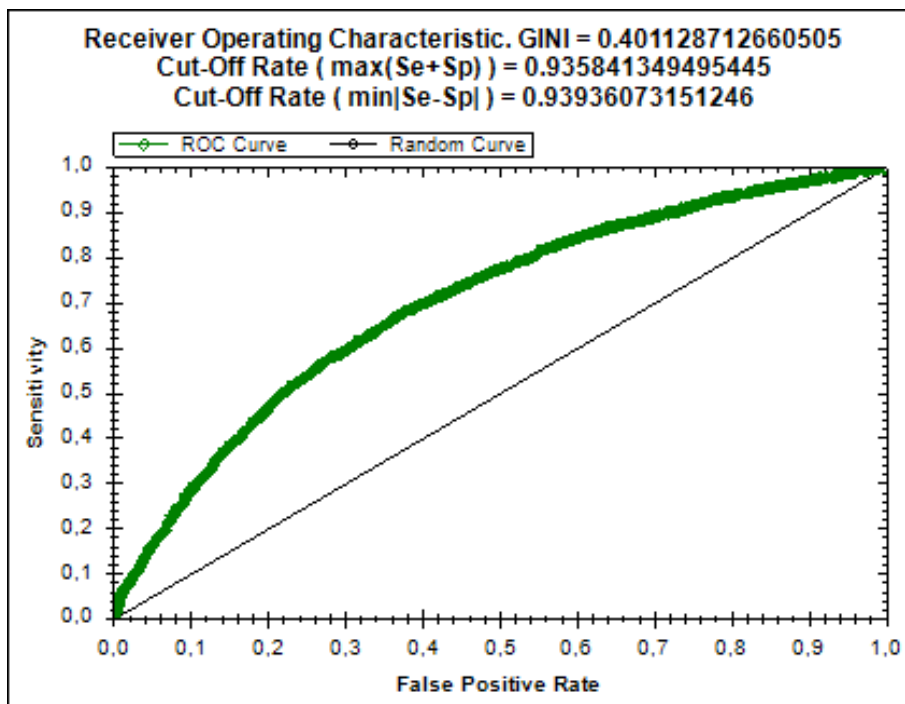


Рисунок 3.10 – Приклад ROC-кривої на тестовій вибірці

Індекс Джині обчислюється як *відношення площі між кривою (зверху) та діагоналлю (знизу) до площі (0,5) трикутника над діагоналлю*, тому індекс Джині можна також виразити через площу під кривою (Area Under the Curve, AUC) [96]:

$$GINI = \frac{AUC - \frac{1}{2}}{\frac{1}{2}} = 2AUC - 1 = 2 \int_{\tilde{y} \in \tilde{Y}} Se(\tilde{y}) dFPR(\tilde{y}) - 1,$$

$$GINI = \left(\sum_{\tilde{y}_q \in \{-\infty\} \cup \tilde{Y}_q} (Se(\tilde{y}_q) + Se(\tilde{y}_{q-1})) (FPR(\tilde{y}_q) - FPR(\tilde{y}_{q-1})) \right) - 1,$$

де множина унікальних прогнозів \tilde{Y}_q доповнюється елементом від'ємної безкінечності, щоб отримати одиничну точку графіку: $Se(-\infty) = FPR(-\infty) = 1$.

Остання форма запису індексу Джині відповідає формулі Брауна [60, 61], використовуючи формулу обчислення інтегралу за допомогою методу трапецій.

3.3.1.1.2 Крива Лоренца для накопичення відсортованої вибірки

Крива Лоренца *функції кумулятивного розподілу нульових (bad) елементів* демонструє швидкість накопичення нульових елементів вибірки *відносно зростання частки всіх елементів вибірки, які впорядковані по рейтингу*. Інакше кажучи, при ідеальній класифікації кожен негативний елемент має рейтинг строго менший від кожного позитивного елемента вибірки, і таким чином накопичення є найшвидшим, тобто лінійне відносно частки всіх елементів й досягає свого максимуму 100% при частці рівній відсотку всіх негативних елементів у вибірці. Індекс Джині є відношенням площі фігури над діагоналлю до площі вже непрямокутного трикутника над діагоналлю – див. рисунок 3.11. На рисунку 3.12 зображено конкретний приклад, що відповідає даним рисунку 3.10.

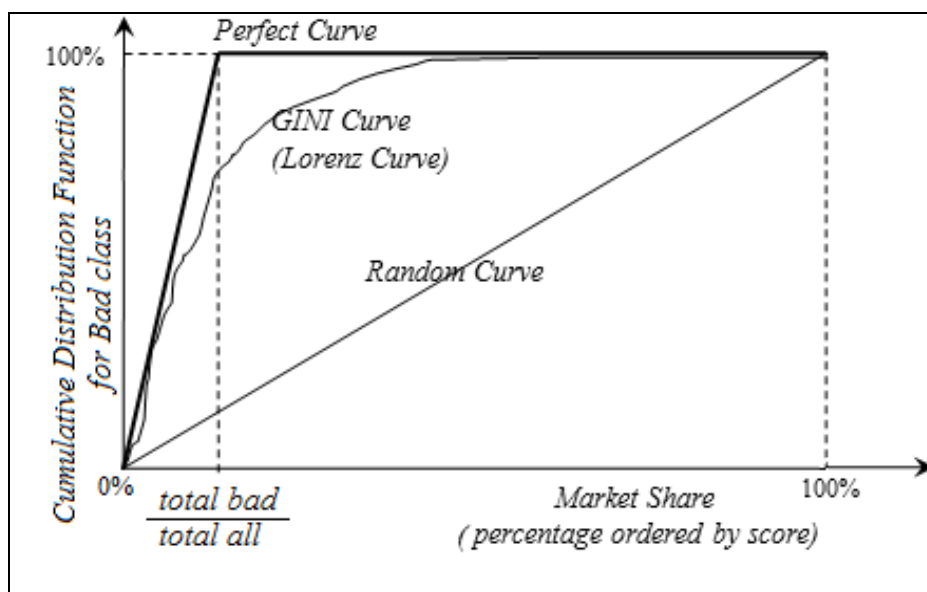


Рисунок 3.11 – Крива Лоренца функції кумулятивного розподілу для класу

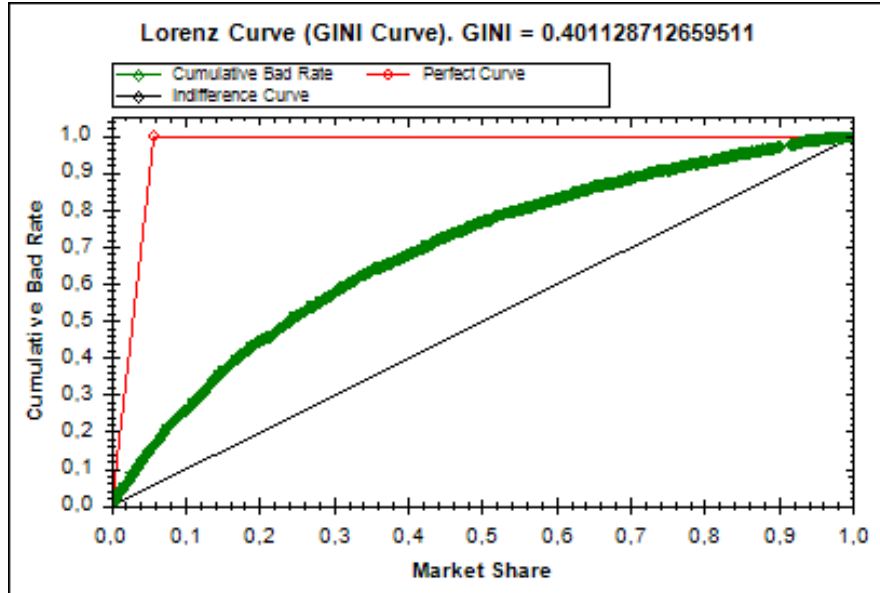


Рисунок 3.12 – Крива Лоренца та приклад розрахунку індексу Джині

Формула розрахунку індексу Джині, де s – доля відсортованої вибірки:

$$GINI = \frac{S_{figure}}{S_{triangle}} = \frac{\int_{\tilde{y} \in \tilde{Y}} F_B(\tilde{y}) ds(\tilde{y}) - \frac{1}{2}}{S_{trapezoid} - \frac{1}{2}},$$

$$GINI = \frac{\sum_{\tilde{y}_q \in \{-\infty\} \cup \tilde{Y}_q} \frac{1}{2} (F_B(\tilde{y}_q) + F_B(\tilde{y}_{q-1})) (s(\tilde{y}_q) - s(\tilde{y}_{q-1})) - \frac{1}{2}}{\frac{1}{2} \left(1 + \left(1 - \frac{\sum_{i=1}^N (1 - y_i)}{N} \right) \right) - \frac{1}{2}},$$

$$GINI = \frac{\sum_{\tilde{y}_q \in \{-\infty\} \cup \tilde{Y}_q} (F_B(\tilde{y}_q) + F_B(\tilde{y}_{q-1})) (s(\tilde{y}_q) - s(\tilde{y}_{q-1})) - 1}{1 - \frac{\sum_{i=1}^N (1 - y_i)}{N}}.$$

3.3.1.1.3 Крива Лоренца функцій кумулятивного розподілу класів

Ще одним (третім) альтернативним способом візуалізації кривої Лоренца та підрахунку показника Джині є крива Лоренца, що відображає залежність кумулятивного розподілу нульових випадків від кумулятивного розподілу одиничних випадків. Показник Джині обчислюється аналогічно до ROC-кривої, тобто відношення площі фігури над діагоналлю до прямокутного трикутника. Власне графік є симетричним до ROC-кривої, яка по суті є теж кривою Лоренца, відносно другої діагоналі (див. рисунок 3.13) типу $Y = 1 - X$. Формула обчислення індексу Джині вже наводилася у підрозділах 2.5 та 3.1.9:

$$GINI = \left(\int_{\tilde{y} \in \tilde{Y}} F_B(\tilde{y}) dF_G(\tilde{y}) - \frac{1}{2} \right) / \left(\frac{1}{2} \right),$$

$$GINI = \left(\sum_{\tilde{y}_q \in \{-\infty\} \cup \tilde{Y}_q} (F_B(\tilde{y}_q) + F_B(\tilde{y}_{q-1})) (F_G(\tilde{y}_q) - F_G(\tilde{y}_{q-1})) \right) - 1,$$

де остання форма запису є формулою Брауна [60, 61]; $F_B(-\infty) = F_G(-\infty) = 0$.

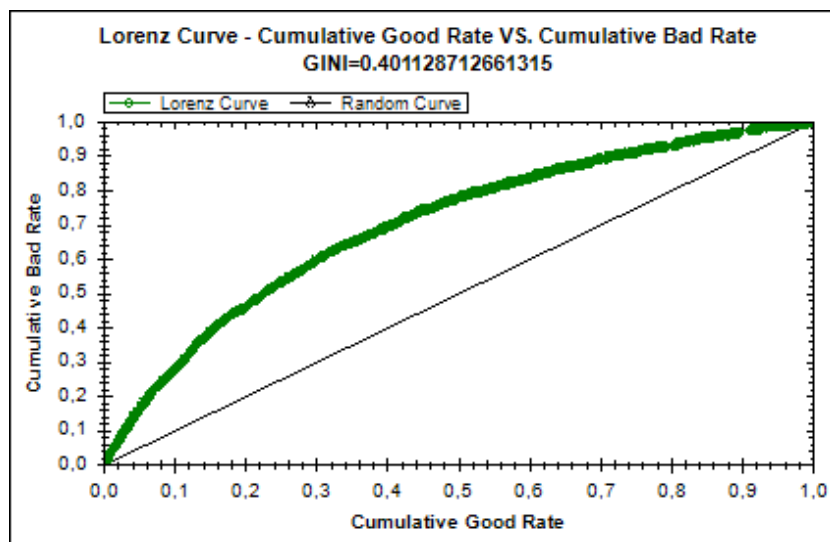


Рисунок 3.13 – Крива порівняння функцій кумулятивних розподілів

Типові твердження про предикативну силу моделі скорингу на тестовій вибірці, використовуючи значення індексу Джині, представлені у таблиці 3.3.

Таблиця 3.3 – Типові судження про предикативну силу на тестовій вибірці

Предикативна сила (прогностична сила)	Тип моделі скорингу (скорингової карти)	
	Аплікаційна модель	Поведінкова модель
Низька	$\text{GINI} < 35\%$	$\text{GINI} < 55\%$
Середня	$35\% \leq \text{GINI} < 45\%$	$55\% \leq \text{GINI} < 65\%$
Висока	$\text{GINI} \geq 45\%$	$\text{GINI} \geq 65\%$

3.3.1.2 Статистика Колмогорова-Смирнова

Як було описано в підрозділах 2.5 і 2.6, статистика Колмогорова-Смирнова – це максимальна абсолютна різниця між двома емпіричними функціями кумулятивного розподілу прогнозів для двох класів (при цьому множину унікальних елементів доповнювати від’ємною безкінечністю немає потреби):

$$KS = \max_{\tilde{y} \in \tilde{Y}} |F_B(\tilde{y}) - F_G(\tilde{y})| = \max_{\tilde{y}_q \in \tilde{Y}_q} |F_B(\tilde{y}_q) - F_G(\tilde{y}_q)|.$$

Тест Колмогорова-Смирнова для двох вибірок (two-sample Kolmogorov-Smirnov test) полягає у спростуванні нуль-гіпотези щодо рівності двох розподілів. Згідно з процедурою даного тесту, обчислюється значення випадкової величини

$$\rho(G, B) = \sqrt{\frac{N_G N_B}{N_G + N_B}} KS, \text{ що має розподіл Колмогорова [97]:}$$

$$K(y) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 y^2} = 1 + 2 \sum_{j=1}^{\infty} (-1)^j e^{-2j^2 y^2},$$

$$pvalue_{KS} = 1 - K \left(\sqrt{\frac{N_G N_B}{N_G + N_B}} KS \right).$$

Розподіл Колмогорова відповідає розподілу випадкової величини супремуму (*sup*) випадкового процесу броунівського мосту (*Brownian bridge*) [98] на неперервному інтервалі $[0; 1]$: $\xi = \sup_{t \in [0; 1]} |B(t)|$.

Броунівський міст є випадковим процесом означеним властивостями [98]:

1) $B(t)$ – гауссівський процес визначений на $t \in [0; 1]$;

2) $E(B(t)) = \bar{B}(t) = 0$;

3) коваріаційна функція (*covariance function*) даного випадкового процесу:

$$\dot{K}(s, t) = E((B(s) - \bar{B}(s))(B(t) - \bar{B}(t))) = E(B(s)B(t)) - E(B(s))E(B(t)) = \min\{s, t\} - st.$$

Як наслідок, очевидно, що $B(0) = B(1) = 0$ і $\delta^2(t) = t \wedge t - t^2 = t - t^2 = t(1 - t)$.

При $KS \geq 30\%$ скорингова модель вважається сильно предикативною.

На рисунку 3.14 зображено приклад обчислення статистики Колмогорова-Смирнова, що не залежить від шкали або монотонного перетворення осі абсцис.

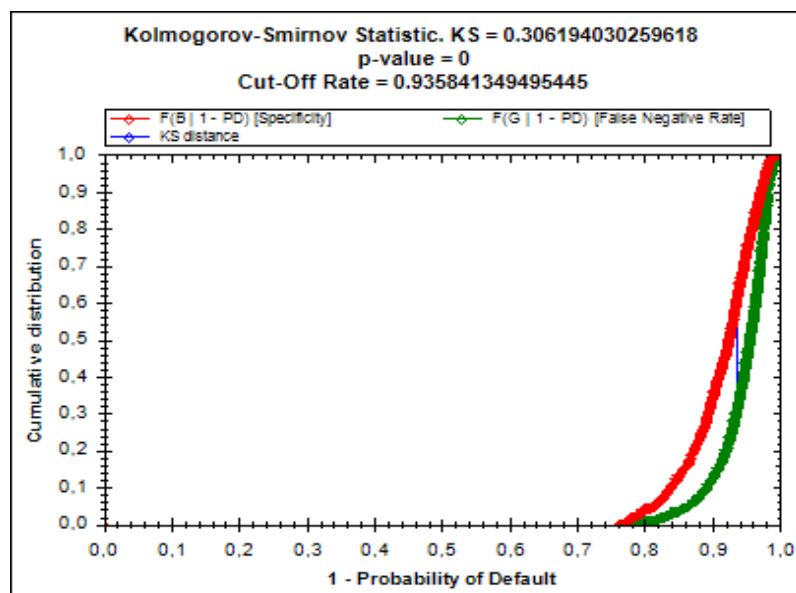


Рисунок 3.14 – Приклад обчислення статистики Колмогорова-Смирнова

3.3.1.3 Відстань Махаланобіса

Відстань Махаланобіса – проста міра віддаленості двох розподілів [5]:

$$M = \frac{|m_G - m_B|}{\sqrt{\frac{N_G \sigma_G^2 + N_B \sigma_B^2}{N_G + N_B}}},$$

$$m_G = E(p(y=1|\mathbf{x}) | \mathbf{x} \in X_{Good}) = E(\tilde{y}(\mathbf{x}) | \mathbf{x} \in X_{Good}) = N_G^{-1} \sum_{i: y_i=1} \tilde{y}_i,$$

$$m_B = E(p(y=1|\mathbf{x}) | \mathbf{x} \in X_{Bad}) = E(\tilde{y}(\mathbf{x}) | \mathbf{x} \in X_{Bad}) = N_B^{-1} \sum_{i: y_i=0} \tilde{y}_i,$$

$$\sigma_G^2 = \frac{1}{N_G - 1} \cdot \sum_{i: y_i=1} (\tilde{y}_i - m_G)^2,$$

$$\sigma_B^2 = \frac{1}{N_B - 1} \cdot \sum_{i: y_i=0} (\tilde{y}_i - m_B)^2,$$

тобто відстань Махаланобіса враховує власне відстань між середніми значеннями двох розподілів прогнозованих ймовірностей одиничного класу та їх дисперсії (див. рисунок 3.15). Приклад зображено на рисунку 3.16. Емпірична формула кількості інтервалів гістограм (через коефіцієнти ексцесу Пірсона) [99]:

$$I = \min \left\{ \frac{\varepsilon_B + 1,5}{6} N_B^{0,4}, \frac{\varepsilon_G + 1,5}{6} N_G^{0,4} \right\}.$$

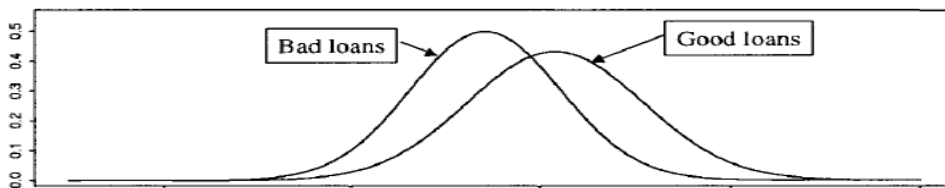


Рисунок 3.15 – Ілюстрація щодо суті відстані Махаланобіса

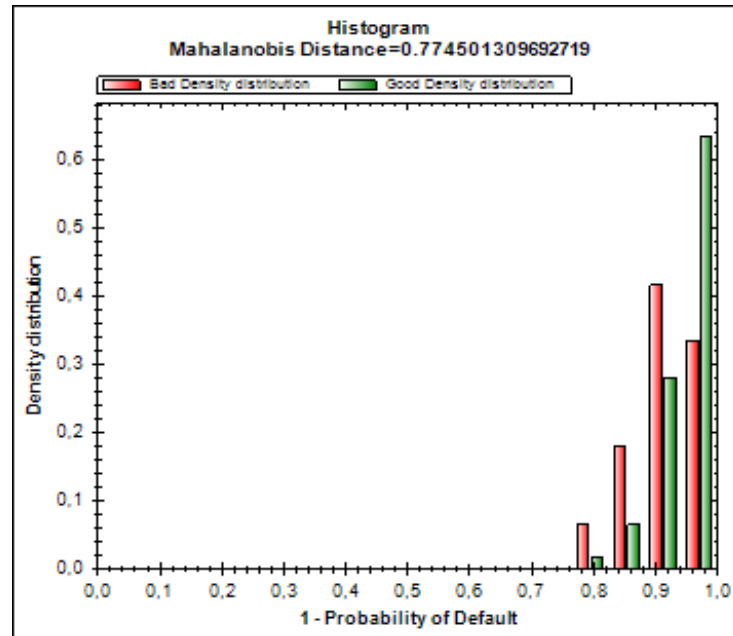


Рисунок 3.16 – Приклад ілюстрації до відстані Махаланобіса

При $M > 0,75$ вважається, що скорингова модель високої прогностичності.

3.3.1.4 Статистика Хосмера-Лемешоу

Статистика Хосмера-Лемешоу є показником предикативності моделі після розбиття значень прогнозів на тестовій вибірці на J інтервалів розмірів n_j [33]:

$$HL = \sum_{j=1}^J \frac{n_j (p_j - \bar{p}_j)^2}{\bar{p}_j (1 - \bar{p}_j)} = \sum_{j=1}^J \frac{(o_j - n_j \bar{p}_j)^2}{n_j \bar{p}_j (1 - \bar{p}_j)},$$

де мають місце середні *прогнознi* ймовірності $\bar{p}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \tilde{y}_{ji}$ та *фактичні*

ймовірності $p_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ji}$ для інтервалів, де останні ймовірності відповідають

фактичним кількостям спостережень одиничного класу $o_j = n_j p_j = \sum_{i=1}^{n_j} y_{ji}$ для

інтервалів. Методом симуляції встановлено, що статистика HL має розподіл $\chi^2(J-2)$ [33]. Тобто більше значення $pvalue_{HL} = 1 - F_{\chi^2(J-2)}(HL)$ є кращим (гіпотеза рівності розподілів прогнозів і фактів) – менше значення HL є кращим.

3.3.1.5 Розробка алгоритму розрахунку показника Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса засобами мови SQL

Завданням даного підрозділу є розробка алгоритму розрахунку індексу Джині, статистики Колмогорова-Смирнова і відстані Махаланобіса засобами мов програмування четвертого покоління (4GL) на прикладі мови SQL. Ключовими особливостями реалізації є операції над множинами, застосування агрегатних та аналітичних віконних функцій, використання узагальнених табличних виразів. Пропонований алгоритм наведено на рисунку 3.17 разом з елементами реалізації мовою T-SQL системи керування базами даних (СКБД) MS SQL Server 2014 [100].

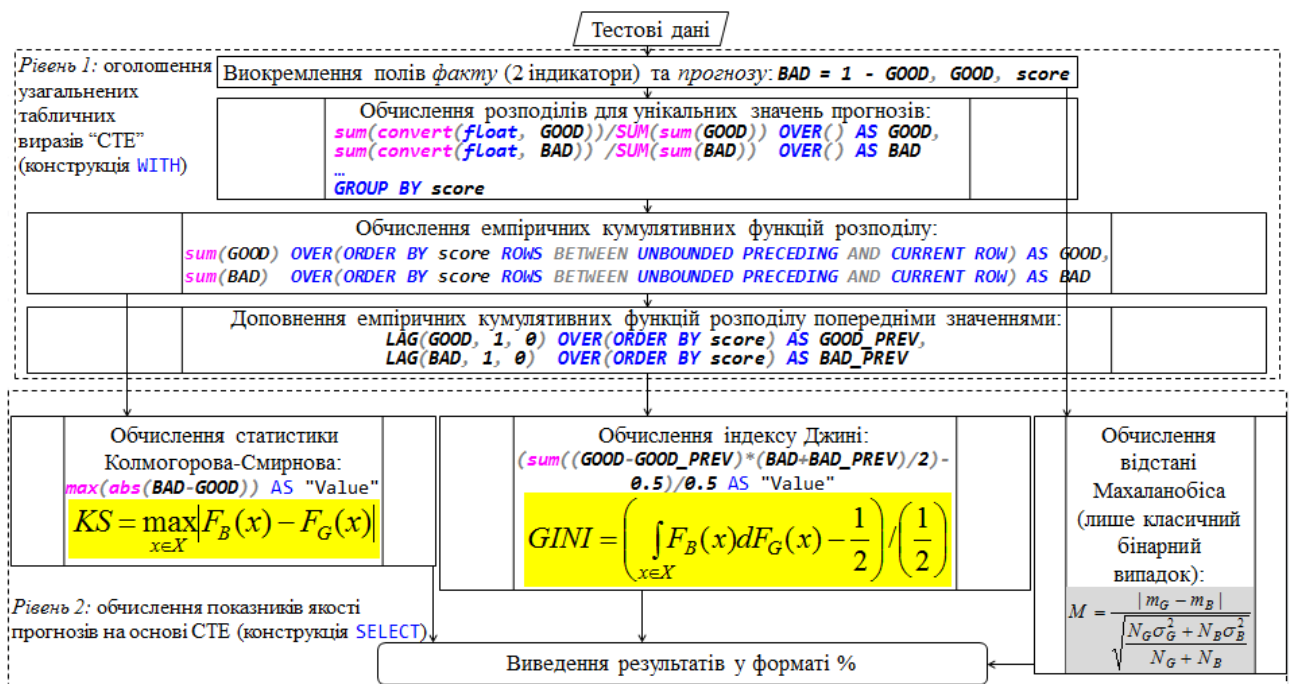


Рисунок 3.17 – Алгоритм розрахунку показників якості прогнозів засобами SQL

Приклад виводу результатів зображено на рисунку 3.18.

Results		Messages	
	Indicator	Value	Value %
1	K-S.	0,306194030259574	30.62%
2	GINI	0,401128712659868	40.11%
3	Mahalanobis distance	0,774501309698799	77.45%

Рисунок 3.18 – Приклад виводу результатів у середовищі MS SQL Server 2014: якість прогнозів моделі споживчого кредитування з урахуванням відхилених заявок на тестовій вибірці прийнятих заявок

Слід зауважити, що запропонований алгоритм застосовний для ймовірнісної цільової змінної, але у такому випадку відстань Махаланобіса не розраховується, оскільки вона не узагальнюється у рамках роботи, на відміну від індексу Джині та статистики Колмогорова-Смирнова. Ключовими перевагами даного алгоритму є зокрема наочність і простота з максимальним використанням можливостей мов програмування четвертого покоління на прикладі мови структурованих запитів.

3.3.1.6 Вдосконалення, а саме узагальнення індексу Джині та статистики Колмогорова-Смирнова для ймовірнісної цільової змінної

У підрозділах 2.5, 3.1.9, 3.3.1.1.3 наведено таку формулу для індексу Джині:

$$GINI = \left(\int_{\tilde{y} \in \tilde{Y}} F_B(\tilde{y}) dF_G(\tilde{y}) - \frac{1}{2} \right) / \left(\frac{1}{2} \right),$$

$$GINI = \left(\sum_{\tilde{y}_q \in \{-\infty\} \cup \tilde{Y}} (F_B(\tilde{y}_q) + F_B(\tilde{y}_{q-1})) (F_G(\tilde{y}_q) - F_G(\tilde{y}_{q-1})) \right) - 1.$$

У підрозділах 2.5, 2.6, 3.3.1.2 наведено формулу обчислення статистики Колмогорова-Смирнова:

$$KS = \max_{\tilde{y} \in \tilde{Y}} |F_B(\tilde{y}) - F_G(\tilde{y})|,$$

$$KS = \max_{\tilde{y}_q \in \tilde{Y}} |F_B(\tilde{y}_q) - F_G(\tilde{y}_q)|.$$

Суть вдосконалення – заміна аналізу вибірки як двох підмножин на обчислення узагальнених емпіричних функцій кумулятивного розподілу:

$$F_G(\tilde{y}_q) \sim F_G^*(\tilde{y}_q) = \frac{\sum_{i: \tilde{y}_i \leq \tilde{y}_q} y_i}{\sum_{i=1}^N y_i},$$

$$F_B(\tilde{y}_q) \sim F_B^*(\tilde{y}_q) = \frac{\sum_{i: \tilde{y}_i \leq \tilde{y}_q} (1 - y_i)}{\sum_{i=1}^N (1 - y_i)}.$$

Аналогічно, у підрозділі 3.3.1.1.1 для оцінювання індексу Джині з використанням ROC-кривої, *вигляд формул* для чутливості (*Se*) та доли неправильно класифікованих нульових елементів відносно всіх нульових елементів вибірки (*FPR*) *уже представлено у формі придатній для використання фактичної ймовірнісної цільової змінної*.

Запропоновані ймовірнісні вдосконалення не досліджуються у межах теорії дисертаційної роботи *лише для методу* розрахунку індексу Джині за допомогою кривої Лоренца для накопичення нульового класу на відсортованій вибірці по зростанню прогнозованої ймовірності одиничного класу (підрозділ 3.3.1.1.2).

3.3.1.7 Cross-валідація скорингової карти. Метод «leave-one-out»

Перехресне тестування (*cross-validation*) – група методів тестування статистичних моделей, що передбачає перебір різних однорідних варіантів розбиття на навчальну та тестову вибірки з подальшою побудовою множини статистичних моделей реалізованих на різних підмножинах з метою усереднення отриманих результатів по всіх тестових вибірках, що сприяє послабленню залежності отриманих результатів валідації від способу розбиття вибірки на дві підвибірки. Метод «*leave-one-out*» – метод, який повністю елімінує проблему розбиття вибірки на навчальну та тестову шляхом побудови N різних статистичних моделей на $N - 1$ навчальних елементах з метою прогнозування на 1 тестовий елемент. У результаті отримується ряд фактичних та прогнозованих значень $\{y_i, \tilde{y}_i\}_{i=1}^N$, що дозволяє однозначно оцінити якість прогнозу [5]. Недолік методу: кількість моделей дорівнює розміру вибірки. Цікавим з практичної точки зору є дослідження ряду логарифма правдоподібності (*Log Likelihood, LogL*) на $N - 1$ елементах – позначимо його як «XLogL» (*Exclusive LogL*), який отримується при виключенні кожного елемента з вибірки на інших елементах.

Метод перехресного тестування «*leave-one-out*» також описано раніше у підрозділі 3.1.9.

3.3.2 Аналіз стійкості скорингових моделей

Аналіз стійкості скорингової моделі являє собою порівняння розподілів поточного загального потоку (множини, на яку застосовується модель; *recent sample*) з розподілами навчальної вибірки (множини, на якій будувалася модель; *development sample*) відносно окремих змінних моделі і прогнозу моделі в цілому.

3.3.2.1 Індекс стійкості розподілу

Індекс стійкості розподілу (Population Stability Index, PSI) є найбільш популярною мірою розбіжності двох дискретних розподілів, що набувають однакових значень [8, 9], і має формулу аналогічну інформаційній статистиці:

$$PSI = \sum_i (r_i - d_i) \ln \left(\frac{r_i}{d_i} \right),$$

де r_i – розподіл загального потоку, d_i – розподіл навчальної вибірки.

Даний показник розраховується як у розрізі фінальних скорингових груп, так і в розрізі розподілів категорій (інтервалів) для кожної з вхідних змінних.

Типові судження про стійкість моделі або окремої змінної [8, 9] представлені у вигляді таблиці 3.4.

Таблиця 3.4 – Типові судження щодо стійкості [8, 9]

Значення індексу стійкості	Рівень стійкості
$< 0,10$	Хороша стійкість
$[0,10; 0,25)$	Середня стійкість
$\geq 0,25$	Погана стійкість

Індекс стійкості розподілу виражається через відстань Кульбака-Лейблера аналогічно інформаційній статистиці (див. підрозділ 2.3):

$$PSI = D_{KL}(\mathbf{r} \parallel \mathbf{d}) + D_{KL}(\mathbf{d} \parallel \mathbf{r}).$$

3.3.2.2 Звіти розрахунку зміщення скорингового балу

Аналогічно обчислюються зміщення середнього балу змінних/моделі [8, 9]:

$$Shift = \sum_i (r_i - d_i) \cdot points_i .$$

3.3.2.3 Статистика Колмогорова-Смирнова та стійкість розподілів

Як альтернатива індексу стійкості розподілу, може використовуватись статистика Колмогорова-Смирнова, яка застосовна й до неперервних змінних:

$$KS = \max_{x \in X} |F_r(x) - F_d(x)| .$$

Статистику Колмогорова-Смирнова зручно використовувати для перевірки гіпотези рівності розподілу фінального скорингового балу на загальному потоці відносно навчальної вибірки, оскільки вже при невеликій кількості вхідних категоріальних або дискретизованих змінних, що мають по декілька категорій (інтервалів), кількість можливих значень прогнозованої ймовірності або відповідного скорингового балу є достатньо великою, що вважати скоринговий бал «неперервним». Менше значення статистики Колмогорова-Смирнова означає кращу стійкість (гіпотеза про рівність розподілів). Також можна розрахувати відповідне значення статистичної значимості гіпотези (див. підрозділ 3.3.1.2):

$$pvalue_{KS} = 1 - K \left(\sqrt{\frac{N_r N_d}{N_r + N_d}} KS \right),$$

яке повинно бути достатньо великим (як вірогідність гіпотези рівності).

3.4 Висновки до третього розділу

У підрозділі 3.1 проаналізовано основні методи побудови скорингових моделей та вдосконалено два методи моделювання – побудова логістичної регресії та метод *k*-найближчих сусідів. «Метод *k-plus*-найближчих сусідів» призначений для подолання головних недоліків класичного методу та може узагальнено застосовуватися при ймовірнісній цільовій змінній. Узагальнення логістичної регресії для неперервної ймовірнісної цільової змінної в комплексі з узагальненням ваги категорії змінної (WoE) та інформаційної статистики (IV) дозволяє на двох рівнях перейти від моделі «бінарний факт, ймовірнісний прогноз» до моделі «ймовірнісний факт, ймовірнісний прогноз». Цей факт цінний при проведенні аналізу відхилених заявок, як надалі показано у підрозділі 3.2. Доведено, що формули вектора градієнта та матриці Гессе класичної логістичної регресії застосовні для ймовірнісної цільової змінної. Критерії якості моделі узагальнені у підрозділі 3.3. Розроблено алгоритм прискорення збіжності логістичної регресії. Отже, у підрозділі 3.1 повністю вирішується завдання № 6, друга частина завдання № 5, частина завдання № 4 загальної постановки задач дисертаційного дослідження.

У підрозділі 3.2 проаналізовано методи аналізу відхилених заявок. Вдосконалено метод ітеративної класифікації щодо включення відхилених заявок у кредитному скорингу, запропоновано критерій асимптотичної (рівномірної) збіжності методу за допомогою відстані Чебишева, наведено переваги перед класичним методом (частина завдання № 4 постановки задач). Для методу нечіткого доповнення представлено можливість заміни зваженої логістичної регресії, яка потребує розбиття кожної відхиленої заявки на дві, на запропоноване ще у підрозділі 3.1.8 узагальнення логістичної регресії для ймовірнісної цільової змінної, яке вже не потребує розбиття відхилених заявок,

які частково (ймовірно) класифіковані. Дане застосування підтверджує актуальність узагальнення логістичної регресії.

У підрозділі 3.3 представлено методи оцінювання якості прогнозів моделі скорингу та методи аналізу стійкості. Розроблено вдосконалення індексу Джині і статистики Колмогорова-Смирнова для неперервної цільової змінної, що набуває ймовірнісних значень (тобто завершено завдання № 4 постановки задач). Також розроблено алгоритм розрахунку індексу Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса засобами мов програмування четвертого покоління (4GL) – тут SQL (відповідає першій частині завдання № 7 постановки задач), при цьому GINI та K-S. застосовні для ймовірнісного випадку. Представлено PSI через відстані Кульбака-Лейблера.

РОЗДІЛ 4

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ПОБУДОВИ
ДОВІЛЬНИХ МОДЕЛЕЙ ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ТА
РЕЗУЛЬТАТИ ПРОВЕДЕНИХ ЕКСПЕРИМЕНТІВ

Метою даного розділу є проектування та реалізація архітектури системи підтримки прийняття рішень для побудови довільних скорингових моделей, а також розробка конкретних скорингових моделей та порівняння деяких методів.

4.1 Архітектура системи підтримки прийняття рішень

Архітектуру оригінальної системи підтримки прийняття рішень (СППР) для побудови довільних скорингових моделей зображено на рисунку 4.1.

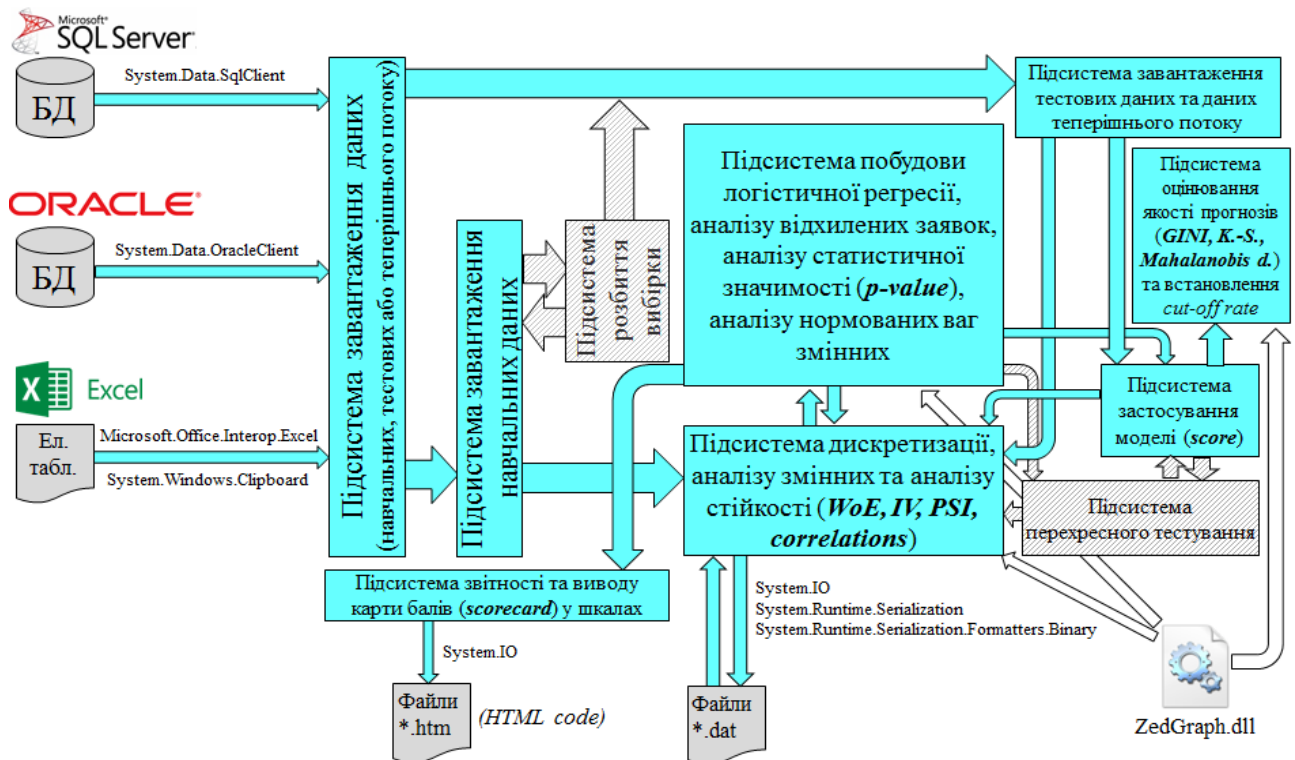


Рисунок 4.1 – Функціональна архітектура (Visual C# / .NET) СППР для скорингу

4.2 Середовище розробки системи та формат вхідних даних

Для розробки СППР з наведеною вище архітектурою використано IDE MS Visual Studio 2008+ (.NET Framework 3.5+) [101–103]. Інтерфейс користувача у вигляді Windows Forms інтуїтивно зрозумілий та створений таким чином, щоб провести аналітика від моменту завантаження даних і до виведення результатів. У програмній реалізації використовується LINQ як компонент .NET [101–103].

Як зображено на рисунку 4.1, вхідні дані можуть бути у трьох форматах: (а) таблиця MS SQL Server; (б) таблиця Oracle Database; (в) електронна таблиця MS Excel формату «*.xls» (MS Excel 2003) або «*.xlsx» (MS Excel 2007+). Головною вимогою до формату таблиці є те, що останньому стовпцеві повинна відповідати цільова змінна, яка може набувати значень: (1) 1 – «good»; (2) 0 – «bad»; (3) NULL – відхилена заявка («rejected application»); (4) дійсне число як ймовірність одиничного («good») значення класу. Оскільки весь вхідний масив спочатку перетворюється до текстового формату, то фізичний формат цільової змінної та інших змінних не важливий. При такому конвертуванні усіх даних у текстовий формат порожні значення, значення NULL з СКБД та значення полів, які у текстовому форматі дорівнюють рядку, що у верхньому регістрі (upper case) дорівнює «NULL», перетворюються у текстові «NULL»-значення, які надалі вважаються порожніми та завжди розміщуються в окрему категорію або інтервал. Щодо дійсних чисел, то при подальшому виборі числового (numerical) формату змінної символи крапки та коми вважаються десятковими роздільниками. Також при завантаженні даних з MS Excel можна спочатку завантажити лише структуру: заголовки стовпців (headers) з кількістю рядків таблиці, а надалі скористатися комбінаціями клавіш копіювання та вставки власне лише даних з буферу обміну.

4.3 Впровадження та технічні вимоги для коректної роботи програми

Для нормальної роботи системи необхідна ЕОМ з характеристиками:

- операційна система: Microsoft Windows 2000/XP/Vista/7/8/8.1;
- встановлена платформа Microsoft .NET Framework не нижче версії 3.5;
- вільний дисковий простір об'ємом не менше 4 Мб для розміщення виконавчого файлу, динамічних бібліотек та супроводжуваних файлів.

4.4 Результати обчислювальних експериментів для споживчого кредитування з наведенням етапів моделювання у рамках графічного інтерфейсу системи

Даний підрозділ розкриває дві підзадачі: (а) побудова моделі кредитного скорингу для споживчого кредитування; (б) демонстрація прикладу експлуатації розробленої системи підтримки прийняття рішень з наведенням інструкцій щодо особливостей графічного інтерфейсу користувача (Graphical User Interface, GUI).

4.4.1 Завантаження навчальних даних на прикладі MS SQL Server

Після натискання кнопки «Load Development Sample from RDBMS» у головному вікні програми заповнене вікно завантаження показано на рисунку 4.2.

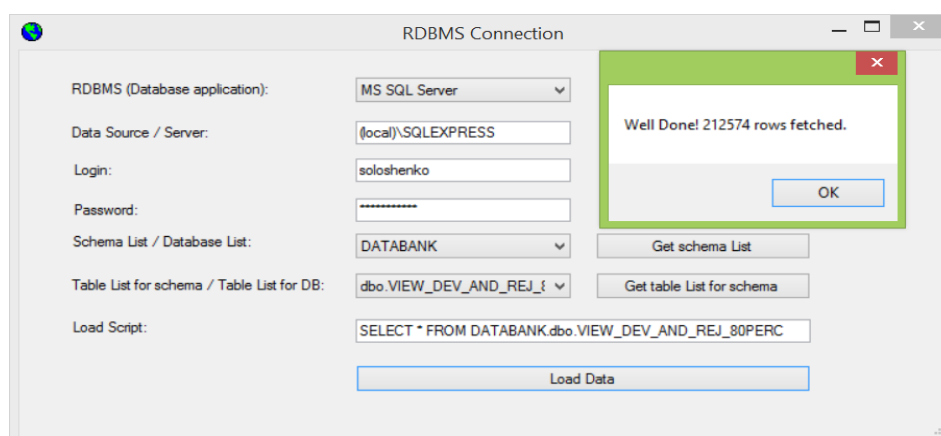
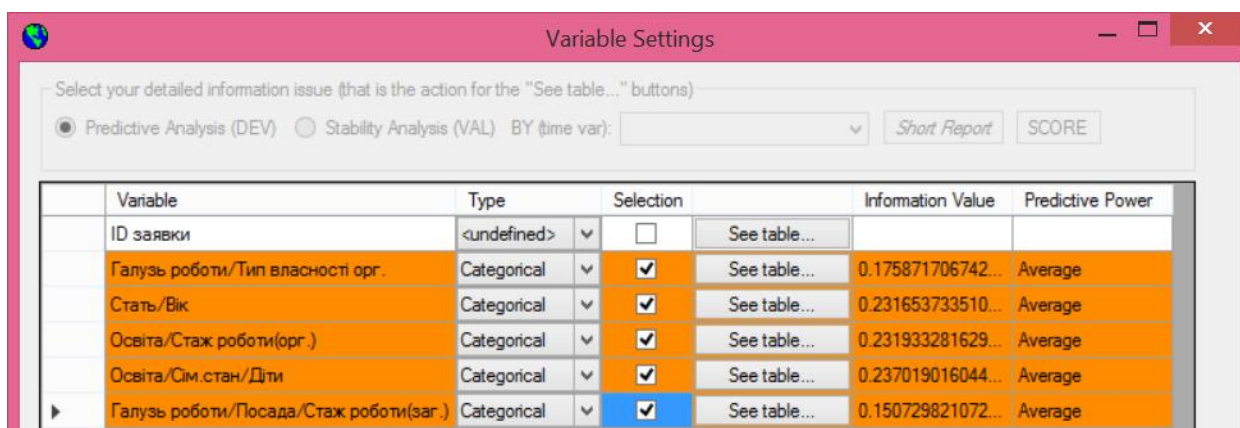


Рисунок 4.2 – Завантаження 212574 заявок (з них 53798 відхиленних)

4.4.2 Аналіз предикативної сили вхідних змінних і кореляційний аналіз

Початковий аналіз предикативної сили відразу після завантаження даних відбувається лише для прийнятих заявок (підмножина завантажених даних з не пустою цільовою змінною). Після натискання кнопки «Variable Settings» у головному вікні та налаштування вхідних змінних у наступному вікні програми обчислюється інформаційна статистика для кожної вхідної змінної завантаженої у передньому підрозділі відносно цільової змінної, що описана у підрозділі 2.2 та зображена на рисунку 2.3. Після натискання кнопки «Commit Variable Selection» вікно налаштування вхідних змінних має вигляд зображений на рисунку 4.3.



Variable	Type	Selection	Information Value	Predictive Power
ID заявки	<undefined>	<input type="checkbox"/>		
Галузь роботи/Тип власності орг.	Categorical	<input checked="" type="checkbox"/>	0.175871706742...	Average
Стать/Вік	Categorical	<input checked="" type="checkbox"/>	0.231653733510...	Average
Освіта/Стаж роботи(орг.)	Categorical	<input checked="" type="checkbox"/>	0.231933281629...	Average
Освіта/Сім. стан./Діти	Categorical	<input checked="" type="checkbox"/>	0.237019016044...	Average
Галузь роботи/Посада/Стаж роботи(sar.)	Categorical	<input checked="" type="checkbox"/>	0.150729821072...	Average

Рисунок 4.3 – Аналіз предикативної сили комбінованих вхідних змінних

Подальший перехід, що описується послідовністю натиснення кнопок «See table...» та «Visual View», дозволяє отримати графік аналізу характеристик, суть якого описано в підрозділі 2.6. Приклад для конкретної змінної зображено на рисунку 4.4. Кореляційну матрицю для WoE-рядів вхідних змінних зображено на рисунку 4.5. Кореляційна матриця зображується після натискання кнопки «WoE Correlation Table» у вікні налаштування множини вхідних змінних.

Варто окремо зауважити, що завантажені відхилені заявки відповідають 80% всіх відхилених заявок, аналогічно прийнятним заявкам у навчальній вибірці.

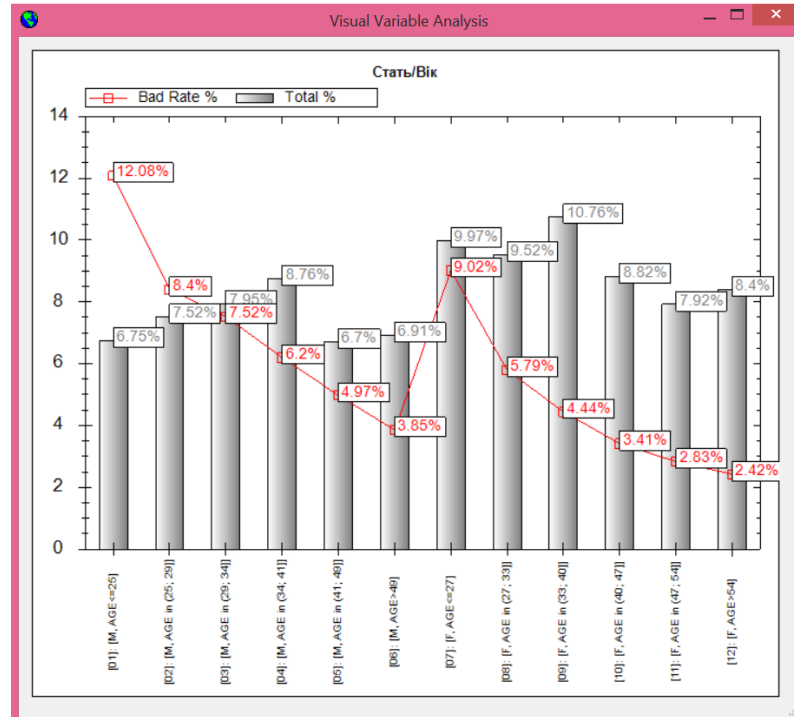


Рисунок 4.4 – Графік аналізу комбінованої вхідної змінної «Стать/Вік»

	Галузь роботи/Тип власності орг.	Стать/Вік	Освіта/Стаж роботи(орг.)	Освіта/Сім.стан./Діти	Галузь роботи/Посада/Стаж роботи(sag.)
▶ Галузь роботи/Тип власності орг.	1	0.3033	0.1787	0.1686	0.4211
Стать/Вік	0.3033	1	0.117	0.1347	0.3038
Освіта/Стаж роботи(орг.)	0.1787	0.117	1	0.5896	0.3503
Освіта/Сім.стан./Діти	0.1686	0.1347	0.5896	1	0.171
* Галузь роботи/Посада/Стаж роботи(sag.)	0.4211	0.3038	0.3503	0.171	1

Рисунок 4.5 – Кореляційна матриця WoE-рядів вхідних змінних

4.4.3 Побудова моделі логістичної регресії. Аналіз відхилених заявок

Після проведеного вище обрання множини вхідних змінних у головному вікні програми необхідно натиснути кнопку «PROC LOGISTIC» для побудови моделі логістичної регресії. При цьому автоматично запускається включення та аналіз відхилених заявок, де метод або його вдосконалення обираються у налаштуваннях головного вікна програми як на рисунку 4.6, при цьому перший варіант відповідає методу ітеративної класифікації (рисунок 3.7), а другий – його

вдосконаленню, яке запропоновано називати «методом ітеративного обчислення ймовірностей» (рисунки 3.8 та 3.9).

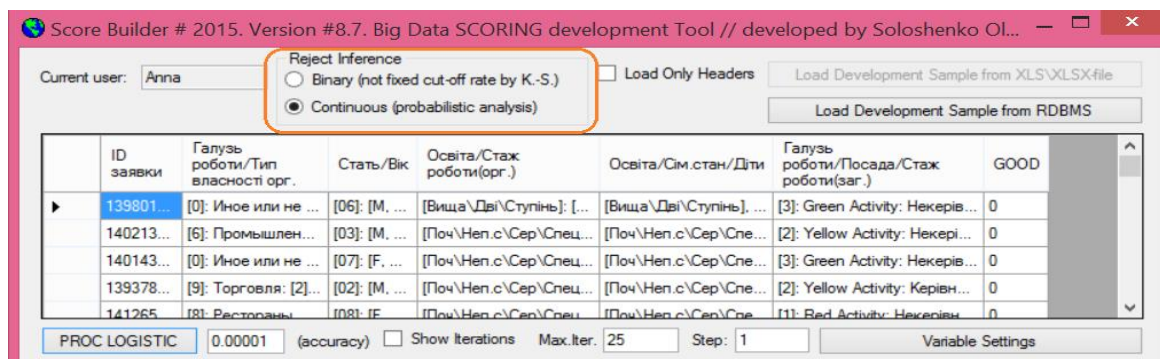


Рисунок 4.6 – Налаштування методу включення і аналізу відхилених заявок

Результати виконання логістичної регресії та аналізу відхилених заявок відображаються в окремому вікні, зображеному на рисунках 4.7, 4.8 та 4.9.

Name	Coefficient	StdDev	Wald Chi-Square	p-value	StdDev(WOE)	Influence(WOE coef)
Intercept	2.65817577386769	0.0093635080462769	80591.681126385	<=0.0001	0	0%
Галузь роботи/Тип власності орг.	0.508406067253878	0.0256475378291849	392.943483889807	<=0.0001	0.460927327547642	20.76%
Стать/Вік	0.597316083071573	0.0195517851769654	933.330769009904	<=0.0001	0.522319642101263	27.64%
Освіта/Стаж роботи(орг.)	0.464335289146334	0.0223447360296897	431.830114117154	<=0.0001	0.521480745775043	21.45%
Освіта/Сім.стан./Діти	0.49399957817232	0.0208727519556649	560.136317485838	<=0.0001	0.52338439996768	22.91%
* Галузь роботи/Посада/Стаж роботи(зар.)	0.209020022855618	0.0253457202358359	68.0090186395856	<=0.0001	0.390597640855787	7.23%

Рисунок 4.7 – Коефіцієнти логістичної регресії та їх аналіз

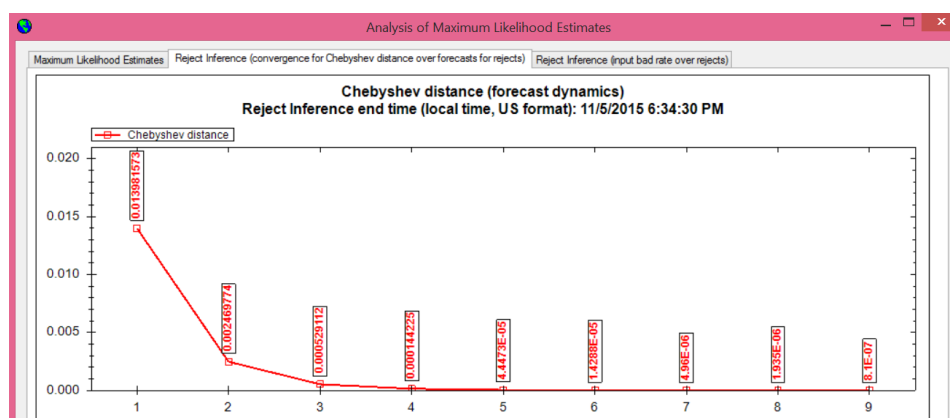


Рисунок 4.8 – Демонстрація збіжності методу аналізу відхилених заявок

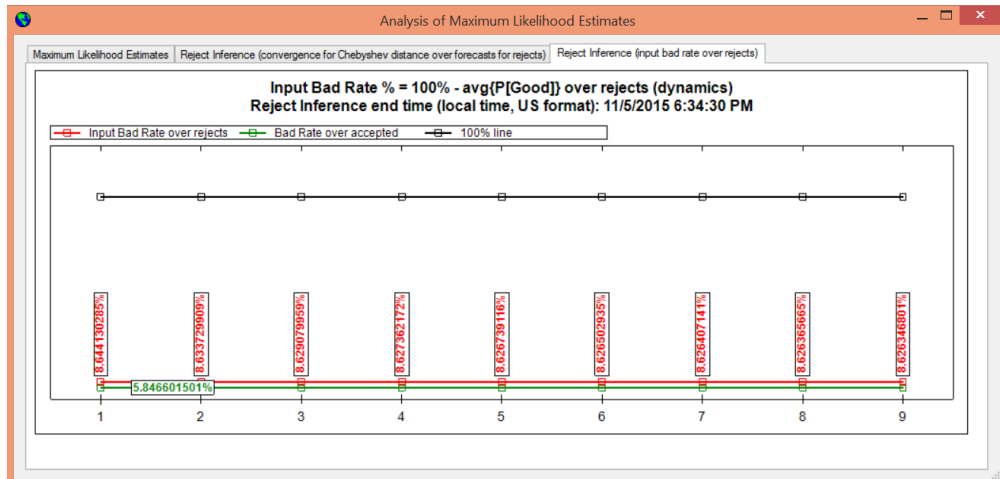


Рисунок 4.9 – Динаміка ймовірності нульового класу для відхилених заявок

Коефіцієнти логістичної регресії дублюються у головному вікні (рисунок 4.10), де відображається максимізований логарифм функції правдоподібності.

Variable	Intercept	Галузь роботи/Тип власності орг.	Стать/Вік	Освіта/Стаж роботи(орг.)	Освіта/Сім.стан./Діти	Галузь роботи/Посада/Стаж роботи(sag.)
*	2.65817577...	0.50840606725...	0.597316083...	0.464335289146334	0.49399957817232	0.209020022855618

Log Likelihood: -48055.8052844407

Рисунок 4.10 – Коефіцієнти регресії та значення критерію оптимальності

Результати переобчислення значень IV, кореляцій наводяться у додатку В.

4.4.4 Аналіз якості прогнозів на тестовій вибірці

Завантаження тестових даних з СКБД MS SQL Server відбувається аналогічно навчальним даним, але після натискання кнопки «Load Validation (or Recent) Sample from RDBMS» у головному вікні програми. Далі заповнюється форма завантаження зображена на рисунку 4.11. У даній задачі тестова вибірка має відому бінарну цільову змінну, оскільки складається з 20% прийнятих заявок.

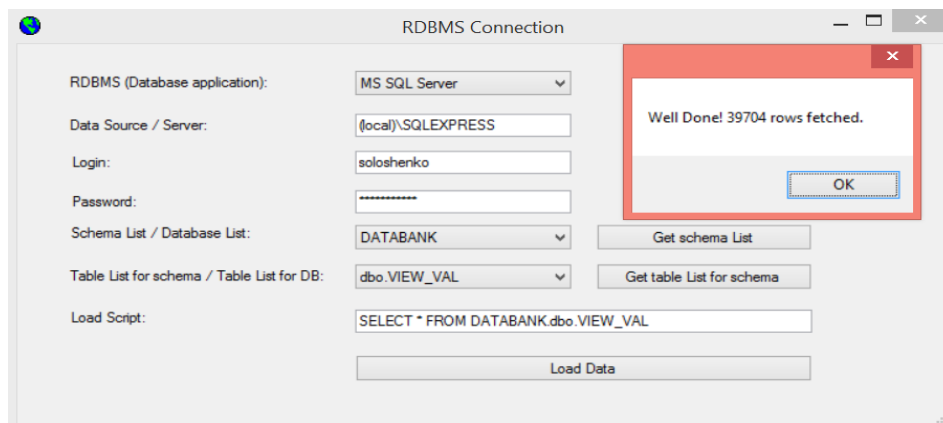


Рисунок 4.11 – Завантаження 39704 тестових заявок (20% від всіх прийнятих)

Далі для обчислення та виводу показників якості прогнозів на тестовій вибірці використовується кнопка «Forecasting using Validation sample», що розміщена на головній формі програми. При цьому на окремій формі виводиться п'ять графіків, що були вже наведені у попередньому розділі: рисунок 3.13 (індекс Джині), рисунок 3.12 (індекс Джині), рисунок 3.10 (індекс Джині), рисунок 3.14 (статистика Колмогорова-Смирнова), рисунок 3.16 (відстань Махаланобіса). Також на формі виводиться таблиця значень двох показників (рисунок 4.12).

	INDICATOR	VALUE %
▶	Gini (Goods curve VS. Bads curve)	40.1128712661315%
	Gini (Lorenz Curve)	40.1128712659511%
	Gini (ROC Curve)	40.1128712660505%
	K-S.	30.6194030259618%

Рисунок 4.12 – Таблиця значень деяких показників якості прогнозів

4.4.5 Порівняння з класичним методом ітеративної класифікації

У даному підрозділі наводиться порівняння пропонованого вдосконалення методу ітеративної класифікації (названого «методом ітеративного обчислення ймовірностей») з класичним методом. Також слід зауважити, що включення та аналіз відхилених заявок призначені для покращення стійкості розподілів, тоді як

якість прогнозів на тестовій вибірці прийнятих заявок може бути дещо нижчою. Тому треба мати на увазі, що для даного прикладу модель без урахування відхилених заявок (лише на прийнятих) має показники: GINI = 40,32%; K.-S. = 30,82%; MD = 0,7828; PA = 89,4943%. На рисунку 4.13 наведено порівняння. Тут

псевдоточність – це показник $PA = 100\% - \frac{1}{N} \sum_{i=1}^N |y_i - P(\mathbf{c}, \mathbf{X}(i))| = 100\% - E |error_i|$.

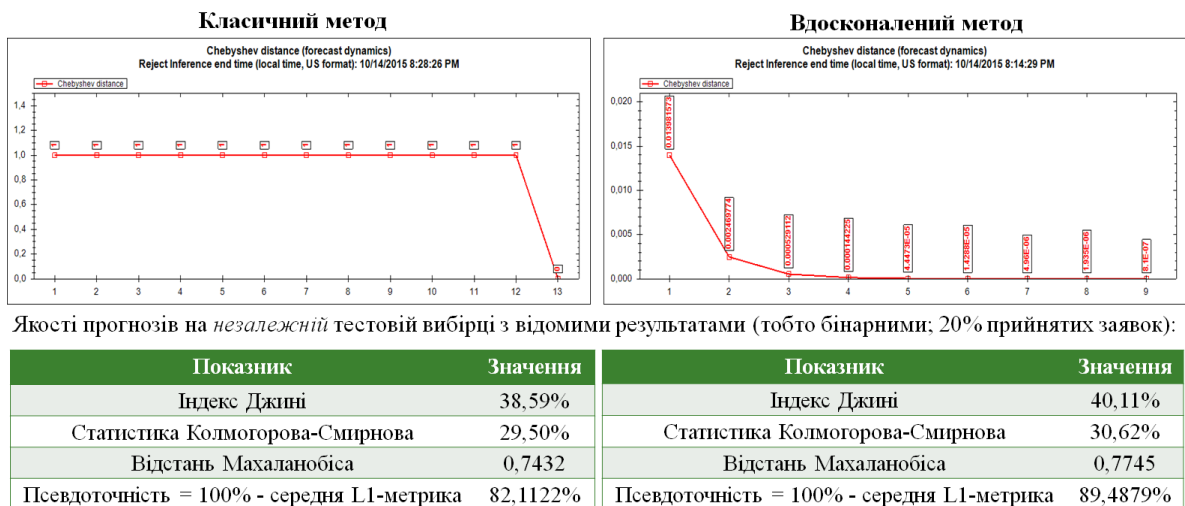


Рисунок 4.13 – Переваги вдосконаленого методу ітеративної класифікації

4.4.6 Представлення скорингової карти балів

Для того, щоб, наприклад, отримати скорингову карту балів (*scorecard*), де сумарний бал виражає натуральний логарифм співвідношення шансів (*log of odds*) помножений на 100, треба у головному вікні програми натиснути кнопку «Scores», а далі натиснути кнопку «Show simplified Score = 100*Clog*WoE». Тоді буде зображено вікно балів, яке зображене на рисунку 4.14. Також дане вікно містить налаштування границь скорингових груп, які зокрема використовуються при виборі режиму дослідження стійкості скорингової моделі у згаданому раніше вікні налаштування змінних. При повторному аналізі предикативної сили, виводі графіків, кореляції вже задіяна навчальна вибірка з відхиленими заявками.

Variable TYPE	Variable NAME	Coef. Logistic	Variable VALUE	WoE	SCORE
	#Intercept	2.658175773867...			265.8175774
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[0]: Иное или не указано; [1] Государственное	0.688121285029...	34.9845036
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[0]: Иное или не указано; [2] Негосударственное	-0.25198315163...	-12.8109763
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[1]: Коммунал. службы, муницип. управл., услуги и т.д.; [1] Го...	0.461567372808...	23.4663653
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[1]: Коммунал. службы, муницип. управл., услуги и т.д.; [2] Не...	-0.10975296535...	-5.5799073
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[2]: СМИ, издательства, туризм, салоны, шоу-бизнес; [1] ...	0.375480024590...	19.0896323
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[2]: СМИ, издательства, туризм, салоны, шоу-бизнес; [2] ...	-0.00278389515...	-0.1415349
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[3]: Банки, финансы, ИТ; [1] Государственное	1.130798659223...	57.4904899
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[3]: Банки, финансы, ИТ; [2] Негосударственное	0.671521292176...	34.1405499
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[4]: МВД, юрид. услуги, вооруж. силы, охр. и детектив. п...	0.413683367727...	21.0319134
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[4]: МВД, юрид. услуги, вооруж. силы, охр. и детектив. п...	-0.29506574861...	-15.0013217
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[5]: Здравоохранение; [1] Государственное	0.773668184200...	39.3337599
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[5]: Здравоохранение; [2] Негосударственное	0.550197569938...	27.9723783
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[6]: Промышленность, строительство, сельское хоз. и ...	0.258880489868...	13.1616412
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[6]: Промышленность, строительство, сельское хоз. и ...	-0.30688161260...	-15.6020474
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[7]: Наука и культура, образование; [1] Государственное	1.023284640948...	52.024412
Categorical	Галузь роботи/Тип власності орг.	0.508406067253...	[7]: Наука и культура, образование; [2] Негосударствен...	0.395264286341...	20.0954761

Score_Group	Max Prob. Def.	auto MinEdge(>=)	auto MaxEdge(<)
[1]: Best	0.02	389.1820298	
[2]: Excellent	0.04	317.805383	389.1820298
[3]: Good	0.08	244.2347035	317.805383
[4]: Not Bad	0.16	165.8228077	244.2347035
[5]: Bad		165.8228077	

Рисунок 4.14 – Вікно фінальної скорингової карти балів

4.4.7 Збереження налаштувань змінних

Збереження налаштувань вхідних змінних при потребі може здійснюватися за допомогою кнопки «Save (Serialization)» на формі налаштування вхідних змінних, а завантаження – за допомогою кнопки «Open (Deserialization)», яка знаходиться поряд. При цьому реалізація має справу з серіалізацією і відповідно десеріалізацією об'єктів .NET [102], використовуючи файли з розширенням .dat.

4.5 Результати застосування вдосконалення методу k -найближчих сусідів

Таблиця 4.1 – Порівняння моделей, побудованих лише на прийнятих заявках

Метод моделювання на прийнятих заявках	Індекс Джині	Кількість параметрів, що оптимізуються	Кількість безумовно заданих параметрів
Логістична регресія	40,32%	6	0
« k -plus-NN» ($k = 10$)	30,45%	0	1
« k -plus-NN» ($k = 50$)	36,58%	0	1

Результати, що наведені у вигляді таблиці 4.1, свідчать про високу ефективність вдосконаленого методу k -найближчих сусідів у сенсі достатньої якості прогнозів при нульовій кількості коефіцієнтів моделі, що потребують пошуку, задаючи лише один безумовний параметр – кількість найближчих сусідів. При цьому більш широке вдосконалення даного методу, яке має високу обчислювальну складність (див. підрозділ 3.1.9), передбачає пошук оптимального значення кількості найближчих сусідів. Але навіть при емпіричному значенні кількості найближчих сусідів, вдосконалений метод має достатньо високу якість прогнозів на незалежній тестовій вибірці, що є не набагато нижчою за якість прогнозів логістичної регресії (див. таблицю 4.1). При порівнянні моделей використовуються дані споживчого кредитування з підрозділу 4.4.

4.6 Деякі характеристики моделі для індивідуального кредитування

У даному підрозділі представлені деякі характеристики скорингової моделі для індивідуального кредитування. Головними завданнями даного підрозділу є більш детальне дослідження статистичної значимості коефіцієнтів і дослідження перехресного тестування моделі. З даної навчальної вибірки повністю виключена попередньо означена невизначена або «сіра» зона (див. підрозділ 2.1).

4.6.1 Завантаження навчальних даних з електронної таблиці MS Excel

У даній задачі прискорене завантаження даних відбувається у два етапи:

- 1) завантаження лише структури електронної таблиці (див. підрозділ 4.2), встановлюючи відмітку «Load Only Headers» у головному вікні та натискаючи кнопку «Load Development Sample from XLS\XLSX-file» з подальшим вибором тут фільтру файлів «Excel 2007 (XLSX) (*.xlsx)» та вибором власне файлу;

- 2) копіювання та вставка лише даних з MS Excel у структуру dataGridView.

4.6.2 Дослідження статистичної значимості коефіцієнтів моделі

На рисунку 4.15 зображено аналіз статистичної значимості коефіцієнтів.

Analysis of Maximum Likelihood Estimates							
Maximum Likelihood Estimates							
	Name	Coefficient	StdDev	Wald Chi-Square	p-value	StdDev(WOE)	Influence(WOE coef)
▶	Intercept	2.58570693354666	0.02984329892464	7506.97452171426	<=0.0001	0	0%
	РЕГІОН ПРОЖИВАННЯ	0.888239281874073	0.0828432620314209	114.959685960639	<=0.0001	0.390106242069425	14.41%
	ЗНАЧЕННЯ ПЕРЕПЛАТИ	0.574792048225286	0.0699918095838607	67.4414747965218	<=0.0001	0.480303259431753	11.48%
	ДОЛЯ СТРАХОВИХ ВНЕСКІВ	0.688128597693688	0.0685422370356523	100.791211347101	<=0.0001	0.408927897800664	11.7%
	ГАЛУЗЬ ПРАЦЕВПАШТУВАННЯ	0.698173462737055	0.138645373605376	25.3580532880241	<=0.0001	0.252513698363723	7.33%
	СІМЕЙНИЙ СТАН	0.666169055929094	0.0966558674574025	47.502062345269	<=0.0001	0.25789222483937	7.14%
	ЧАС РЕЕСТРАЦІЇ І ПЕРЕВІРКА ДОМ.ТЕП.	0.564589696730834	0.132332540350324	18.2025652935473	<=0.0001	0.177761005780934	4.17%
	ОСВІТА ТА РОБ.СТАЖ В ОРГАНІЗАЦІЇ	0.60736314698592	0.0994312871551392	37.3121908718046	<=0.0001	0.276355251980717	6.98%
	СТАТЬ ТА ВІК	0.808334105258056	0.0780546229628663	107.246850063085	<=0.0001	0.40010810911591	13.45%
	ТИП ОРГАНІЗАЦІЇ ПРАЦЕВЛ. І ЗАГ.РОБ.СТАЖ	0.210577125755369	0.108135902737566	3.7921243206492	0.0515	0.290621577727704	2.54%
*	ПОПЕРЕДНЯ КРЕДИТНА ІСТОРІЯ (ДНІ ПРОСТР.)	0.820312510959433	0.0649489327067736	159.51981337063	<=0.0001	0.609834950906378	20.8%

Рисунок 4.15 – Аналіз статистичної значимості коефіцієнтів моделі регресії

Значення p -value можна порахувати за допомогою класичного алгоритму з використанням системи Wolfram Mathematica[®] або Wolfram|Alpha [30, 104, 105] (рисунок 4.16).

WolframAlpha computational knowledge engine

1-integrate[1/sqrt(2*Pi*x)*exp(-x/2), x, 0, (0.210577125755369/0.108135902737566)^2]

Input interpretation:

$$1 - \int_0^{\left(\frac{0.210577125755369}{0.108135902737566}\right)^2} \frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{x}{2}\right) dx$$

Result:

0.0514942555030

Computation result:

$$1 - \int_0^{\left(\frac{0.210577125755369}{0.108135902737566}\right)^2} \frac{\exp\left(-\frac{x}{2}\right)}{\sqrt{2\pi x}} dx = 0.0514943$$

Download page POWERED BY THE WOLFRAM LANGUAGE

Рисунок 4.16 – Розрахунок p -value за допомогою класичного алгоритму

4.6.3 Результати перехресного тестування моделі

Для проведення перехресного тестування моделі за допомогою методу послідовного виключення одного елемента (методу «*leave-one-out*») у головному вікні програми треба натиснути відповідну кнопку «Cross-validation (“leave-one-out” method)». Отримані показники якості моделі: (а) 50% – індекс Джині, (б) 38% – статистика Колмогорова-Смирнова, (в) 0,97 – відстань Махаланобіса.

4.7 Приклад дискретизації при побудові моделі колекторського скорингу

У даному прикладі цільова змінна – це індикатор виходу будь-коли в 60+ днів прострочення за 2 майбутні місяці, а вхідна змінна – поточна кількість днів прострочення (від 3 до 59 включно). Вхідна змінна завжди визначена. Розмір вибірки складає 27194. При виборі числового типу даних («Numerical») програма сповіщає про 41 унікальне непусте значення. Введені налаштування для методу дискретизації такі: $k = 3$ (кількість інтервалів), $s = 10\% = 0,1$ (допустима доля), $t = 10$ (кратність границь). Результат дискретизації змінної за допомогою базової системи підтримки рішень, яка реалізує пропонований метод (див. підрозділ 2.4), зображено на рисунку 4.17. Програмну реалізацію методу також здійснено за допомогою таких трьох взаємно інтегрованих засобів розробки з метою простої ефективної автоматичної візуалізації орієнтованого графу (рисунок 4.18): Python [19, 37, 38, 106, 107], Linux Ubuntu [108] та GraphViz. У результаті, рисунок 4.18 чітко відображає ідею динамічного програмування Беллмана [56, 109, 110], де критерієм максимізації при трьох формалізованих умовах є інформаційна статистика як сума відстаней Кульбака-Лейблера [46–48, 111]. Кількість вершин кожного рівня орієнтованого графу, не враховуючи нульовий та останній рівні:

$$|layer| = \left(\frac{K_{\max} - K_{\min}}{t} + 1 \right) - (k - 2) = K - (k - 2) = \left(\frac{50 - 10}{10} + 1 \right) - (3 - 2) = 4.$$

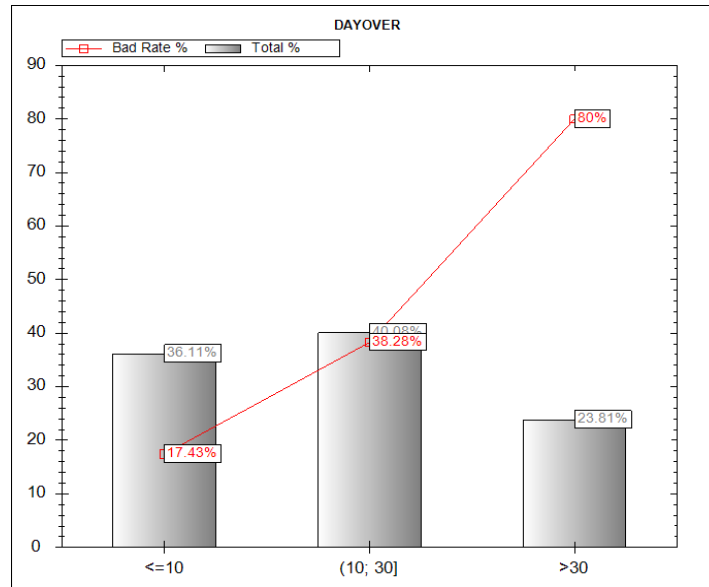


Рисунок 4.17 – Результат дискретизації неперервної змінної

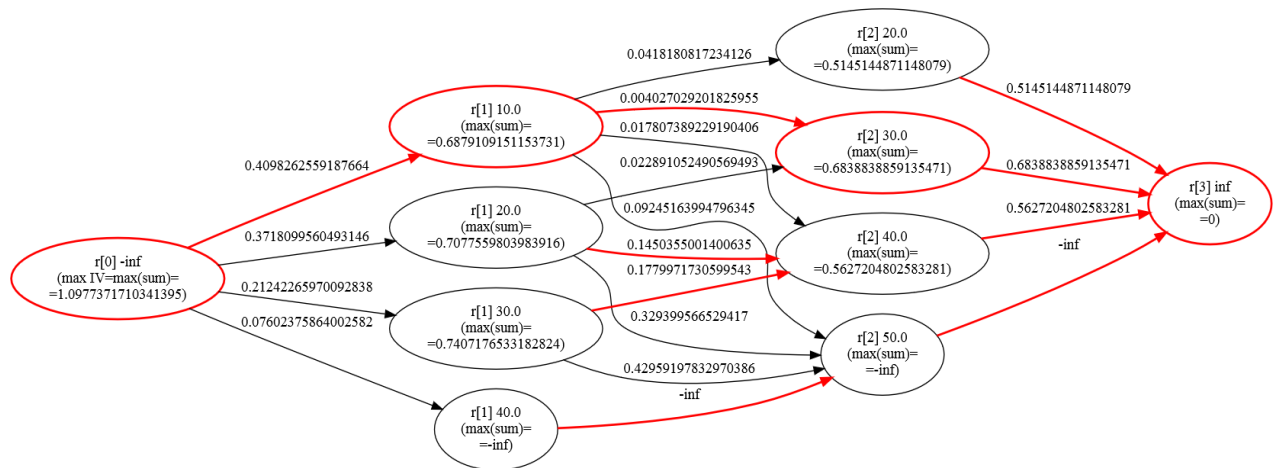


Рисунок 4.18 – Результат реалізації методу дискретизації засобами Python, GraphViz (використовує мову опису графів DOT) та Linux Ubuntu

Перспективним напрямком є програмна реалізація автоматичної побудови графів засобами С# [112] у рамках СППР або засобами R [34–36, 113]. Реалізації методу тут оперують узагальненою інформаційною статистикою [13, 43, 114].

Для порівняння, при використанні бінарних розбиттів (модифікація ID3), де локальний критерій – інформаційний приріст (*Information Gain*), $IV=1,078585$, тобто розбиття при забезпеченні трьох умов отримується гіршим (рисунок 4.19).

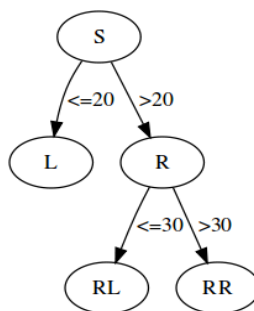


Рисунок 4.19 – Рекурсивне дерево рішень

4.8 Висновки до четвертого розділу

У даному розділі вирішено задачу розробки архітектури системи підтримки прийняття рішень для побудови моделей скорингу та її реалізації засобами Visual C# (друга частина завдання № 7), а також здійснено порівняльні *експерименти* з побудови конкретних моделей для етапів життєвого циклу кредитування, *такі як*:

2.1) побудова моделей аплікаційного кредитного скорингу для споживчого кредитування з урахуванням відхилених заявок за допомогою вдосконаленого методу ітеративної класифікації та класичного методу з подальшим порівнянням результатів та динаміки збіжності при використанні критерію відстані Чебишева рядів вхідних та вихідних оцінок відхилених заявок з обмеженням 10^{-6} з метою доведення: а) вищої якості прогнозів; б) швидшої збіжності при обмеженні;

2.2) побудова моделі для споживчого кредитування лише на прийнятих заявках за допомогою вдосконаленого методу k -найближчих сусідів і комплексне порівняння з логістичною регресією, що доводить не набагато нижчу якість прогнозів на тестовій вибірці при відсутності параметрів, які оптимізуються;

2.3) побудова моделі для індивідуального кредитування, обчислення рівня значимості для гіпотези рівності коефіцієнта нулю за допомогою пропонованого алгоритму і порівняння з класичним: демонстрація точності, простоти, швидкодії;

2.4) демонстрація методу дискретизації для колекторського скорингу, порівняння з рекурсивним деревом рішень, де критерій є іншим та локальним.

ВИСНОВКИ

У дисертаційній роботі проведено наукове дослідження по розробці та вдосконаленню моделей і методів оцінювання кредитоспроможності фізичних осіб, розроблено та реалізовано оригінальну програмну архітектуру, застосовано відповідну розроблену систему для побудови конкретних моделей оцінювання кредитоспроможності. За результатами можна зробити наведені нижче висновки.

1. Розроблений метод дискретизації неперервних змінних на основі ідеї динамічного програмування Беллмана дозволяє формалізувати постановку задачі дискретизації вхідної неперервної змінної відносно цільової змінної (бінарної або також ймовірнісної, враховуючи подальші узагальнення), знайти глобальний максимум інформаційної статистики при заданих умовах на відміну, наприклад, від дерев рішень на прикладі моделі збору заборгованості.

2. Розроблений метод розрахунку статистики Колмогорова-Смирнова, ваги категорії змінної та інформаційної статистики при відомому розподілі категорій та умовному розподілі цільової змінної дозволяє оцінити (відновити) значення показників предикативної сили, виходячи лише з графіку аналізу характеристик.

3. Розроблений алгоритм обчислення рівня статистичної значимості (p -value) спростування нуль-гіпотези для коефіцієнтів моделі шляхом інтегрування розкладу в ряд Тейлора з використанням особливостей мов програмування третього покоління дозволяє просто, швидко, точно оцінювати значення p -value у вигляді безкінечного числового ряду, що залежить лише від математичного сподівання і середньоквадратичного відхилення коефіцієнта логістичної регресії. Алгоритм передбачає просте обмеження на останній врахований член ряду, що дозволяє обчислювати наближене значення суми ряду, опираючись лише на особливості типів даних, без використання чисельних методів інтегрування.

4. Вдосконалення методу моделювання за допомогою моделі логістичної регресії, ваг категорій змінних, інформаційної статистики, індексу Джині та

статистики Колмогорова-Смирнова дозволяє цілісно вирішувати задачу побудови та тестування моделей для випадку ймовірнісної цільової змінної або довільної цільової змінної, що набуває неперервних значень з інтервалу $[0\%; 100\%]$, наприклад, задачу прогнозування відносних втрат у випадку дефолту. А відповідний вдосконалений метод ітеративної класифікації для включення та аналізу відхилених заявок має значні переваги перед класичним методом, що, зокрема, включають в себе ймовірнісні значення проміжних оцінок, відсутність необхідності застосування порогу відсікання для проміжних оцінок, кращу якість прогнозів на тестовій вибірці, швидшу збіжність при встановленні досить точного критерію збіжності, переоцінку ваг категорій на рівні конкретних категорій окремих змінних. Вдосконалений метод нечіткого доповнення для включення та аналізу відхилених заявок має суттєву перевагу перед класичним методом, яка полягає у відсутності необхідності дублювання та зважування відхилених заявок, а при додатковій переоцінці та узагальненні для випадку ймовірнісної цільової змінної для ваг категорій на рівні конкретних категорій всіх змінних результати класичного і вдосконаленого методів далі співпадають і також дорівнюють проміжним результатам після нульової та першої ітерації вдосконаленого методу ітеративної класифікації.

5. Розроблений алгоритм нормування ваг змінних моделі з урахуванням варіації вхідних параметрів дозволяє виключити фактор масштабу окремих вхідних змінних у регресійних моделях з метою об'єктивної оцінки їх реального впливу, наприклад, на рівні лінійної частини (логіта) моделі логістичної регресії. Алгоритм прискорення пошуку оптимального вектору коефіцієнтів логістичної регресії дозволяє означити початкове значення вектору у просторі коефіцієнтів.

6. Вдосконалення методу k -найближчих сусідів вирішує основні недоліки і неоднозначності класичного методу та його модифікацій, зокрема, мають місце урахування рівновіддалених груп найближчих сусідів, ймовірнісна класифікація, узагальнені числові перетворення для використання множини категоріальних

вхідних змінних згідно з методами оцінювання кредитоспроможності фізичних осіб. Вдосконалення має близьку з логістичною моделлю прогностичність навіть при фіксованому заданому параметрі у рамках першого вдосконалення методу.

7. Розроблений алгоритм обчислення показника індексу Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса засобами мов програмування четвертого покоління на прикладі мови SQL дозволяє простіше та наочніше продемонструвати власне суть математичних формул обчислення ключових показників якості прогнозів. Основні переваги розробленої системи моделювання засобами мови Visual C# перед існуючими програмними рішеннями включають:

(а) портативність: для використання системи достатньо відповідного *exe*-файлу, кількох стандартних *dll*-бібліотек, наявності *.NET Framework 3.5+* на ПК;

(б) низьку вартість програмного забезпечення, коли вартість ліцензії SAS[®] Credit Scoring for Banking становить щонайменше десятки тисяч євро в рік;

(в) реалізацію вдосконаленої технології для ймовірнісної цільової змінної:
 (1) дискретизацію (динамічне програмування); (2) аналіз вдосконаленого IV, кореляційної матриці, *p*-value, нормованих ваг; (3) вдосконалену логістична регресія; (4) вдосконалений метод ітеративної класифікації; (5) аналіз відстані Махаланобіса, узагальнених індексу Джині, статистики Колмогорова-Смирнова; (6) аналіз індексу стійкості розподілу; (7) стандартні методи калібрування балів;

(г) простоту графічного інтерфейсу по принципу мінімального втручання.

Побудовані за допомогою розробленого програмного продукту моделі демонструють, зокрема, кращу якість прогнозів з урахуванням відхилених заявок.

Перспективи подальших досліджень включають вдосконалення методів мультиваріаційного аналізу вхідних змінних для автоматичного вибору множини предикторів, аналітичне обчислення градієнта та матриці Гессе для зваженої логістичної регресії, її подальше узагальнення для ймовірнісної цільової змінної та її програмну реалізацію, реалізацію запропонованого методу дискретизації також засобами T-SQL та XML, агрегування алгоритмів у прикладні методи та теоретичне доведення коректності узагальнення для другої форми індексу Джині.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Солошенко О. Реалізація алгоритму розрахунку показника Джині та статистики Колмогорова-Смирнова для бінарного класифікатора засобами мови SQL / Солошенко Олександр, Бідюк Петро // «ІНТЕРНЕТ-ОСВІТА-НАУКА-2014», дев'ята міжнародна науково-практична конференція ІОН-2014, 14-17 жовтня 2014: зб. праць. — Вінниця: ВНТУ, 2014. — С. 148-150.
2. Прогнозирование утечки депозитных вкладов физических лиц с использованием технологии Data-Mining / [Терентьев А.Н., Бидюк П.И., Коршевнюк Л.А., Солошенко А.Н.] // Інформаційні процеси і технології «Інформатика – 2012»: матеріали міжнар. наук.-практ. конф. молодих учених і студентів, Севастополь, 23-27 квіт. 2012 р. — Севастополь: СевНТУ, 2012. — С. 28–30.
3. Солошенко О.М. Проблема зміщення співвідношення класів при встановленні порогу відсікання для бінарного класифікатора у задачах кредитного скорингу / О.М. Солошенко, П.І. Бідюк // Інформатика, математика, автоматика «ІМА :: 2015»: матеріали та програма наук.-техн. конф. — Суми: Сумський державний університет, 2015. — С. 216.
4. Энциклопедия финансового риск-менеджмента / [Барбаумов В.Е., Рогов М.А., Щукин Д.Ф. и др.]; под ред. А.А. Лобанова и А.В. Чугунова. — М.: Альпина Паблишер, 2003. — 786 с.
5. Thomas L.C. Credit Scoring and its applications: Monograph / Lyn C. Thomas, David B. Edelman, Jonathan N. Crook. — Philadelphia: SIAM, 2002. — 248 p.
6. Згуровский М.З. Системный анализ: проблемы, методология, приложения: Монография / М.З. Згуровский, Н.Д. Панкратова. — К.: Издательство «Наукова думка», 2005. — 743 с.: ил.
7. Згуровський М.З. Основи системного аналізу: підручник [для студ. вищ. навч. закл., які навч. за напрямками «Системний аналіз», «Прикладна математика», «Інформатика», «Комп'ютерні науки», «Комп'ютерна інженерія», «Системна

- інженерія»] / М.З. Згуровський, Н.Д. Панкратова. — К.: Видавнича група «ВНУ», 2007. — 544 с.: іл.
8. Сиддики Наим. Скоринговые карты для оценки кредитных рисков. Разработка и внедрение интеллектуальных методов кредитного скоринга / Наим Сиддики; [пер. с англ. Евгений Ильичев]. — М.: Манн, Иванов и Фербер, 2014. — 268 с.
9. Siddiqi N. Credit risk scorecards: developing and implementing intelligent credit scoring / Naeem Siddiqi. — Hoboken: John Wiley & Sons, Inc., 2006. — 196 p.
10. Руководство по кредитному скорингу / [Ванг Вэй, Влатса А. Димитра, Гленнон К. Деннис и др.]; под ред. Элизабет Мэйз; [пер. с англ. И.М. Тикота; науч. ред. Д.И. Вороненко]. — Минск: Гревцов Паблишер, 2008. — 464 с.
11. Солошенко О.М. Розробка методу k-plus-найближчих сусідів для задач машинного навчання кредитного скорингу / О.М. Солошенко // Східно-Європейський журнал передових технологій. — 2015. — Т. 3, № 9(75). — С. 29–38.
12. Теория систем и системный анализ в управлении организациями: Справочник: учеб. пособие / [Баринов В.А., Болотова Л.С., Волкова В.Н. и др.]; под ред. В.Н. Волковой и А.А. Емельянова. — М.: Финансы и статистика, 2006. — 848 с.
13. Солошенко О.М. Вдосконалення методу ітеративної класифікації з включення відхилених заявок у кредитному скорингу / О.М. Солошенко // Наук. вісті НТУУ «КПІ». — 2014. — № 5. — С. 63–69.
14. Mok Jie-Men. Reject Inference in Credit Scoring / Jie-Men Mok. — Amsterdam: VMI paper, 2009. — 38 p.
15. Feelders A.J. Credit scoring and reject inference with mixture models / A.J. Feelders // International Journal of Intelligent Systems in Accounting, Finance and Management. — 1999. — Vol. 8, № 4. — Pp. 271–279.
16. Терентьев О.М. Модели и методы построения та анализа байесовских сетей для интеллектуального анализа данных: дис. кандидата техн. наук: 05.13.06 / Терентьев Александр Миколайович. — К., 2009. — 258 с.

17. Байєсівські мережі в системах підтримки прийняття рішень / [Згуровський М.З., Бідюк П.І., Терентьєв О.М., Просянкіна-Жарова Т.І.]. — К.: ТОВ «Видавниче Підприємство «Едельвейс», 2015. — 300 с.
18. Чубукова И.А. Data Mining / И.А. Чубукова. — М.: Бином ЛБЗ, 2008. — 384 с.
19. Сегаран Т. Программируем коллективный разум / Тоби Сегаран; [пер. с англ. А. Слинкин]. — СПб.: Символ-плюс, 2008. — 368 с.: ил.
20. Панкратова Н.Д. Становление и развитие системного анализа как прикладной научной дисциплины / Н.Д. Панкратова // Системні дослідження та інформаційні технології. — 2002. — № 1. — С. 65–94.
21. Антонов А.В. Системный анализ: учебник [для вузов] / А.В. Антонов. — М.: Высш. шк., 2004. — 454 с.: ил.
22. Клир Дж. Системология. Автоматизация решения системных задач / Дж. Клир. — М.: Радио и связь, 1990. — 540 с.
23. Спицнадель В.Н. Основы системного анализа: учеб. пособие / В.Н. Спицнадель. — СПб.: «Изд. дом «Бизнес-пресса», 2000. — 326 с.
24. Клиланд Д. Системный анализ и целевое управление / Д. Клиланд, В. Кинг; под ред. И.М. Верещагина; [пер. с англ. М.М. Горяинова, А.В. Горбунова]. — М.: Советское радио, 1974. — 280 с.: ил.
25. Жариков О.Н. Системный подход к управлению: учеб. пособие для вузов / О.Н. Жариков, В.И. Королевская, С.Н. Хохлов; под ред. В.А. Персианова. — М.: ЮНИТИ-ДАНА, 2001. — 62 с.
26. Кузнєцова Н.В. Системний підхід до аналізу кредитних ризиків з використанням мереж Байєса / Н.В. Кузнєцова, П.І. Бідюк // Наук. вісті НТУУ «КПІ». — 2008. — № 3. — С. 11–24.
27. Бідюк П.І. Адаптивне прогнозування фінансово-економічних процесів на основі принципів системного аналізу / П.І. Бідюк // Наук. вісті НТУУ «КПІ». — 2009. — № 5. — С. 54–61.

28. Матрос Є.О. Впровадження системного підходу до прогнозування обсягів втрат внаслідок реалізації банківських ризиків / Є.О. Матрос // Наук. вісті НТУУ «КПІ». — 2006. — № 3. — С. 37–44.
29. Бідюк П.І. Система підтримки прийняття рішень для аналізу фінансових даних / П.І. Бідюк, Н.В. Кузнєцова, О.М. Терентьєв // Наук. вісті НТУУ «КПІ». — 2011. — № 1. — С. 48–61.
30. Wolfram S. A New Kind of Science / Stephen Wolfram. — 1st ed. — Champaign: Wolfram Media, Inc., 2002. — 1197 p.: il.
31. Терентьєв А.Н. SAS BASE: Основы программирования / А.Н. Терентьєв, В.Н. Домрачев, Р.И. Костецкий. — К.: Эдельвейс, 2014. — 304 с.
32. Allison P.D. Logistic regression using the SAS[®] system: theory and application / Paul D. Allison. — Cary: SAS Institute Inc., 1999. — 287 p.
33. Hosmer D.W. Applied logistic regression / David W. Hosmer, Jr., Stanley Lemeshow. — 2nd ed. — Hoboken: John Wiley & Sons, Inc., 2000. — 375 p.
34. Наглядная статистика. Используем R! / [А.Б. Шипунов, Е.М. Балдин, П.А. Волкова и др.]. — М.: ДМК Пресс, 2014. — 298 с.: ил.
35. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R / Р.И. Кабаков; [пер. с англ. П.А. Волкова]. — М.: ДМК Пресс, 2014. — 588 с.: ил.
36. Мастицкий С.Э. Статистический анализ и визуализация данных с помощью R / С.Э. Мастицкий, В.К. Шитиков. — М.: ДМК Пресс, 2015. — 496 с.: ил.
37. Маккинни У. Python и анализ данных / Уэс Маккинни; [пер. с англ. А. Слинкин]. — М.: ДМК Пресс, 2015. — 482 с.: ил.
38. McKinney W. Python for Data Analysis / Wes McKinney. — Sebastopol: O'Reilly Media, Inc., 2013. — 452 p.: il.
39. Ингерсолл Г.С. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Грант С. Ингерсолл, Томас С. Мортон, Эндрю Л. Фэррис; [пер. с англ. А. Слинкин]. — М.: ДМК Пресс, 2015. — 414 с.
40. Kudyba S. Managing Data Mining: Advice from Experts / Stephan Kudyba. — Hershey: CyberTech Publishing (an imprint of Idea Group Inc.), 2004. — 238 p.: il.

41. Лакин Г.Ф. Биометрия: учеб. пособие [для биол. спец. вузов] / Г.Ф. Лакин. — 4-е изд., перераб. и доп. — М.: Высш. шк., 1990. — 352 с.: ил.
42. Солошенко О.М. Дослідження відстані Кульбака-Лейблера у задачах моделювання у кредитному скорингу / О.М. Солошенко // Развитие информационно-ресурсного обеспечения образования и науки в горно-металлургической отрасли и на транспорте 2014: сб. науч. трудов междунар. конф. — Днепропетровск: НГУ, 2014. — С. 328–333.
43. Солошенко А.Н. Обобщение логистической регрессии, веса категории переменной и индекса Джини для непрерывной целевой переменной, принимающей вероятностные значения / А.Н. Солошенко // Кибернетика и системный анализ. — 2015. — Т. 51, № 6. — С. 174–187.
44. Солошенко О.М. Вдосконалені методи розрахунку статистики Колмогорова-Смирнова, ваги категорії змінної та значення інформації у кредитному рейтингу / О.М. Солошенко // Системні дослідження та інформаційні технології. — 2015. — № 4. — С. 104–113.
45. Солошенко О.М. Адаптація формул підрахунку ваг категорій змінної та значення інформації змінної при відомому розподілі категорій та відомих умовних ймовірностях негативних значень цільової змінної / О.М. Солошенко // Проблеми науки. — 2014. — № 10 (166). — С. 45–47.
46. Kullback S. Information theory and statistics / S. Kullback. — New York: John Wiley and Sons, 1959. — 416 p.
47. Хайкин С. Нейронные сети: полный курс / Саймон Хайкин; под ред. Н.Н. Куссуль; [пер. с англ. Н.Н. Куссуль, А.Ю. Шелестова]. — 2-е изд., испр. — М.: ООО “И.Д. Вильямс”, 2006. — 1104 с.: ил.
48. Haykin S. Neural networks: a comprehensive foundation / Simon Haykin. — 2nd ed. — Delhi: Pearson Education, Inc., 2005. — 823 p.: il.
49. Kullback S. On Information and Sufficiency / S. Kullback, R.A. Leibler // Annals of Mathematical Statistics. — 1951. — Vol. 22, № 1. — Pp. 79–86.

50. Bishop C. Pattern Recognition and Machine Learning / C. Bishop. — New York: Springer, 2006. — 738 p.: il.
51. Колмогоров А.Н. Элементы теории функций и функционального анализа: учеб. пособие / А.Н. Колмогоров, С.В. Фомин. — М.: Наука, 1976. — 544 с.
52. Finlay S. Credit scoring, response modelling and insurance rating: a practical guide to forecasting consumer behaviour / Steven Finlay. — London: Palgrave Macmillan, 2010. — 280 p.: il.
53. Anderson R. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation / Raymond Anderson. — New York: Oxford University Press Inc., 2007. — 731 p.: il.
54. Ben-Gan I. Microsoft® SQL Server® 2012 high-performance T-SQL using window functions / Itzik Ben-Gan. — Sebastopol: O'Reilly Media, Inc., 2012. — 221 p.
55. Ben-Gan I. Microsoft® SQL Server® 2012 T-SQL fundamentals / Itzik Ben-Gan. — Sebastopol: O'Reilly Media, Inc., 2012. — 412 p.
56. Зайченко Ю.П. Исследование операций: учебник / Ю.П. Зайченко. — 6-е изд., перераб. и доп. — К.: Издательский дом «Слово», 2003. — 688 с.
57. Geunes J. Operations Planning: Mixed Integer Optimization Models / Joseph Geunes. — Boca Raton: CRC Press, Taylor & Francis Group, LLC, 2015. — 204 p.: il.
58. Спекторський І.Я. Дискретна математика / І.Я. Спекторський. — 2-ге вид., виправл. і доповн. — К.: ІВЦ “Видавництво «Політехніка»”, ТОВ “Фірма «Періодика»”, 2004. — 220 с.: іл.
59. Кожевников В.Л. Теорія інформації та кодування: навч. посібник / В.Л. Кожевников, А.В. Кожевников. — Дніпропетровськ: Національний гірничий університет, 2011. — 108 с.
60. Солошенко О.М. Спосіб розрахунку показника Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса у кредитному скорингу засобами мови SQL / О.М. Солошенко // Наук. вісті НТУУ «КПІ». — 2015. — № 1. — С. 29–35.

61. Згуровський М.З. Сталий розвиток у глобальному і регіональному вимірах: аналіз за даними 2005 р. / М.З. Згуровський. — К.: НТУУ «КПІ», 2006. — 84 с.
62. Алгоритмы: построение и анализ: пер. с англ. / [Кормен Т.Х., Лейзерсон Ч.Е., Ривест Р.Л., Штайн К.]. — 3-е изд. — М.: ООО «И.Д. Вильямс», 2013. — 1328 с.: ил.
63. Треногин В.А. Функциональный анализ: учебник / В.А. Треногин. — 3-е изд., испр. — М.: ФИЗМАТЛИТ, 2002. — 488 с.
64. Дрейпер Н. Прикладной регрессионный анализ: в 2-х кн. Кн. 1: пер. с англ. / Н. Дрейпер, Г. Смит. — 2-е изд., перераб. и доп. — М.: Финансы и статистика, 1986. — 366 с.: ил.
65. Дрейпер Н. Прикладной регрессионный анализ: в 2-х кн. Кн. 2: пер. с англ. / Н. Дрейпер, Г. Смит. — 2-е изд., перераб. и доп. — М.: Финансы и статистика, 1987. — 351 с.: ил.
66. Бідюк П.І. Прикладна статистика: навч. посібник / П.І. Бідюк, О.М. Терентьєв, Т.І. Просянкіна-Жарова. — Вінниця: ПП "ТД "Едельвейс і К", 2013. — 304 с.
67. Буре В.М. Теория вероятностей и математическая статистика: учебник [для вузов] / В.М. Буре, Е.М. Парилина. — СПб.: Издательство «Лань», 2013. — 416 с.: ил.
68. Фёрстер Э. Методы корреляционного и регрессионного анализа: руководство для экономистов / Э. Фёрстер, Б. Рёнц; [пер. с нем. В.М. Иванова]. — М.: Финансы и статистика, 1983. — 302 с.: ил.
69. Pagès J. Multiple Factor Analysis by Example Using R / Jérôme Pagès. — Boca Raton: CRC Press, Taylor & Francis Group, LLC, 2015. — 251 p.: il.
70. Basilevsky A. Statistical factor analysis and related methods: theory and applications: Monograph / Alexander Basilevsky. — New York: A Wiley-Interscience Publication, John Wiley & Sons, Inc., 1994. — 737 p.: il.
71. Kline P. An easy guide to factor analysis / Paul Kline. — L.: Routledge, 1993. — 194 p.: il.

72. Factor analysis at 100: historical developments and future directions / [Cudeck Robert, Bartholomew J. David, Jones V. Lyle et al.]; edited by Robert Cudeck and Robert C. MacCallum. — Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 2007. — 381 p.: il.
73. Kernel-based Data Fusion for Machine Learning: Methods and Applications in Bioinformatics and Text Mining / [Yu S., Tranchevent L.-C., De Moor B., Moreau Y.]. — Berlin: Springer-Verlag, 2011. — 212 p.: il.
74. Амосов А.А. Вычислительные методы: учеб. пособие / А.А. Амосов, Ю.А. Дубинский, Н.В. Копченова. — 4-е изд., стер. — СПб.: Издательство «Лань», 2014. — 672 с.: ил.
75. Lebanon G. Bias, Variance, and MSE of Estimators [Электронный ресурс] / Guy Lebanon. — Atlanta: Georgia College of Tech Computing, 2010. — 2 p. — Режим доступа: <http://www.cc.gatech.edu/~lebanon/notes/estimators1.pdf>.
76. Zhang P. Approximating true relevance model in relevance feedback: a thesis for the degree of Doctor of Philosophy: Computing / Zhang Peng. — Aberdeen, 2013. — 151 p.
77. Hohl W. ARM Assembly Language: Fundamentals and Techniques / William Hohl, Christopher Hinds. — 2nd ed. — Boca Raton: CRC Press, Taylor & Francis Group, 2015. — 420 p.: il.
78. Таненбаум Э.С. Архитектура компьютера / Эндрю С. Таненбаум, Тодд Остин. — 6-е изд. — СПб.: Питер, 2013. — 816 с.: ил.
79. Кузнєцова Н.В. Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування / Н.В. Кузнєцова, П.І. Бідюк // Наук. вісті НТУУ «КПІ». — 2010. — № 1. — С. 42–53.
80. Nisbet R. Handbook of statistical analysis and data mining applications / Robert Nisbet, John Elder, Gary Miner. — San Diego: Academic Press (an imprint of Elsevier), 2009. — 824 p.: il.
81. Егорова И.Н. Программная реализация методов классификации / И.Н. Егорова, С.В. Егоров // Східно-Європейський журнал передових технологій. — 2010. — Т. 1, № 5(43). — С. 52–54.

82. Keller J.M. A fuzzy k-nearest neighbor algorithm / J.M. Keller, M.R. Gray, J.A. Jr. Givens // IEEE transactions on systems, man and cybernetics. — 1985. — Vol. SMC-15, № 4. — Pp. 580–585.
83. Берзлев, О. Ю. Метод прогнозування знаків приростів часових рядів / О. Ю. Берзлев // Східно-Європейський журнал передових технологій. — 2013. — Т. 2, № 4(62). — С. 8–11.
84. Королюк Д.В. Классификация бинарных детерминированных статистических экспериментов с настойчивой регрессией / Д.В. Королюк // Кибернетика и системный анализ. — 2015. — Т. 51, № 4. — С. 163–168.
85. Rokach L. Data mining with decision trees: theory and applications / Lior Rokach, Oded Maimon. — 2nd ed. — Singapore: World Scientific Publishing Co. Pte. Ltd., 2014. — 305 p.: il.
86. Солошенко О.М. Застосування шаблону проектування Composite та технології LINQ до задачі префіксного кодування тексту за допомогою алгоритму Хаффмана / О.М. Солошенко // Наук. вісті НТУУ «КПІ». — 2014. — № 6. — С. 76–82.
87. Михайлюк В.А. О сложности вычисления параметров устойчивости в задачах булева программирования / В.А. Михайлюк, Н.В. Лищук // Кибернетика и системный анализ. — 2015. — Т. 51, № 5. — С. 56–62.
88. Price K.V. Differential Evolution: A Practical Approach to Global Optimization / Kenneth V. Price, Rainer M. Storn, Jouni A. Lampinen. — Berlin: Springer-Verlag, 2005. — 538 p.: il.
89. Yobas M.B. Credit scoring using neural and evolutionary techniques / M.B. Yobas, J.N. Crook, P. Ross // IMA Journal of Mathematics Applied in Business and Industry. — 2000. — № 11. — Pp. 111–125.
90. Солошенко О.М. Застосування системного аналізу до подолання концептуальних невизначеностей цілей у задачах оптимізації портфелю цінних паперів [Електронний ресурс] / О.М. Солошенко // Системні науки та кібернетика. — 2011. — № 1. — С. 14–23. — Режим доступу: <http://mmsa.kpi.ua/ssc/ssc-Vol1-2011/o-m-soloshenko-zastosuvannya-sistemnogo-anal456zu-do-podolannya->

[konceptualnih-neviznachenostei-c456lei-u-zadachah-optim456zac456-portfelyu-c456nnih-paper456v.](#)

91. Солошенко О.М. Застосування системного аналізу до подолання концептуальних невизначеностей цілей у задачах оптимізації портфелю цінних паперів / О.М. Солошенко // Информационно-компьютерные технологии в экономике, образовании и социальной сфере: тезисы докладов V всеукр. науч.-практ. конф. — Симферополь: КРП «Издательство «Крымучпедгиз» , 2010. — С. 81–83.
92. Zeng G. A rule of thumb for reject inference in credit scoring [Електронний ресурс] / Guoping Zeng, Qi Zhao // Mathematical Finance Letters. — 2014. — Vol. 2014, № 2. — Рр. 1–13. — Режим доступу: <http://scik.org/index.php/mfl/article/view/1477>.
93. Montrichard D. Reject Inference Methodologies in Credit Risk Modeling / Derek Montrichard // SESUG 2008: The Proceedings of the SouthEast SAS Users Group, Paper ST-160. — St. Pete Beach: SouthEast SAS® Users Group, 2008. — 10 p.: il.
94. Anderson B. Reject Inference Techniques Implemented in Credit Scoring for SAS® Enterprise Miner™ / Billie Anderson, Susan Haller, Naeem Siddiqi // SAS Global Forum 2009: Proceedings, Paper 305-2009. — Cary: SAS Institute, 2009. — 11 p.: il.
95. McLachlan G.J. Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis / G.J. McLachlan // Journal of American Statistical Association. — 1975. — № 70. — Pp. 365–369.
96. Hastie T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman. — 2nd ed. — New York: Springer-Verlag, 2009. — 745 p.: il.
97. Hollander M. Nonparametric statistical methods / Myles Hollander, Douglas A. Wolfe. — 2nd ed. — New York: John Wiley & Sons, Inc., 1999. — 787 p.: il.
98. Lifshits M.A. Gaussian Random Functions / M.A. Lifshits. — Dordrecht: Springer Science+Business Media, 1995. — 337 p.: il.
99. Weisstein E.W. CRC Concise Encyclopedia of Mathematics / Eric W. Weisstein. — 1st ed. — Boca Raton: CRC Press, 1999. — 1970 p.: il.

100. Бондарь А.Г. Microsoft SQL Server 2014 / А.Г. Бондарь. — СПб.: БХВ-Петербург, 2015. — 592 с.: ил.
101. Троелсен Э. Язык программирования C# 5.0 и платформа .NET 4.5 / Эндрю Троелсен; [пер. с англ. Ю.Н. Артеменко]. — 6-е изд. — М.: ООО “И.Д. Вильямс”, 2013. — 1312 с.: ил.
102. Албахари Дж. C# 5.0. Справочник. Полное описание языка: пер. с англ. / Джозеф Албахари, Бен Албахари. — М.: ООО “И.Д. Вильямс”, 2014. — 1008 с.: ил.
103. C# 5.0 и платформа .NET 4.5 для профессионалов: пер. с англ. / [Нейгел Кристиан, Иввен Билл, Глинн Джей и др.]. — М.: ООО “И.Д. Вильямс”, 2014. — 1440 с.: ил.
104. Дьяконов В.П. Mathematica 5.1/5.2/6 в математических и научно-технических расчетах: Монография / Владимир Павлович Дьяконов. — 2-е изд., перераб. и доп. — М.: Издательство «СОЛОН-Пресс», 2012. — 744 с.: ил.
105. Дьяконов В.П. Mathematica 5/6/7. Полное руководство / Владимир Павлович Дьяконов. — М.: ДМК Пресс, 2010. — 624 с.: ил.
106. Richert W. Building Machine Learning Systems with Python / Willi Richert, Luis Pedro Coelho. — Birmingham: Packt Publishing Ltd., 2013. — 271 p.: il.
107. Karkera K.R. Building Probabilistic Graphical Models with Python / Kiran R. Karkera. — Birmingham: Packt Publishing Ltd., 2014. — 172 p.: il.
108. Граннеман С. Linux. Карманный справочник: пер. с англ. / Скотт Граннеман. — М.: ООО “И.Д. Вильямс”, 2010. — 416 с.: ил.
109. Вентцель Е.С. Исследование операций: задачи, принципы, методология / Е.С. Вентцель. — М.: Наука, 1988. — 208 с.: ил.
110. Вентцель Е.С. Исследование операций / Е.С. Вентцель. — М.: Советское радио, 1972. — 407 с.: ил.
111. Reid M.D. Information, Divergence and Risk for Binary Experiments / Mark D. Reid, Robert C. Williamson // Journal of Machine Learning Research. — 2011. — № 12. — Pp. 731–817.

112. Панкратова Н.Д. Інформаційна система для моделювання та оцінювання фінансових операційних ризиків за допомогою байєсівської мережі / Н.Д. Панкратова, П.І. Бідюк, М.Г. Рубець // Системні дослідження та інформаційні технології. — 2015. — № 3. — С. 7–19.
113. Lantz B. Machine Learning with R / Brett Lantz. — Birmingham: Packt Publishing Ltd., 2013. — 375 p.: il.
114. Soloshenko O.M. Generalizations of Logistic Regression, Weight of Evidence, and the Gini Index for a Continuous Target Variable Taking on Probabilistic Values / O.M. Soloshenko // Cybernetics and Systems Analysis. — 2015. — Vol. 51, № 6. — Pp. 992–1004.

Додаток А

Акт впровадження результатів дисертаційної роботи

forward-bank.com
0 800 300 880



вир. № 15/4 - 05 - 960
від 07.11.14

**Акт впровадження результатів
дисертаційної роботи Солошенка О.М.**

Акт виданий аспіранту Національного технічного університету України «Київський політехнічний інститут» (шифр спеціальності: 01.05.04 – «Системний аналіз і теорія оптимальних рішень») Солошенку Олександрю Миколайовичу в тому, що результати його дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук впроваджені та використовуються в ПАТ «БАНК ФОРВАРД», а саме:

- 1) за допомогою розробленої Солошенком О.М. комп'ютерної програми мовою Visual C# «Система підтримки прийняття рішень для побудови довільних скорингових моделей "СППР ПДСМ"», були отримані ймовірнісні прогнозуючі моделі:
 - 1.1) аплікаційного скорингу для споживчого кредитування;
 - 1.2) поведінкового скорингу прострочення найближчого платежу;
- 2) впроваджені та використовуються такі новітні методи скорингу:
 - 2.1) метод оптимальної дискретизації неперервних змінних за допомогою динамічного програмування Беллмана;
 - 2.2) метод прискорення збіжності вектору коефіцієнтів логістичної регресії за допомогою лінійного прогнозу початкових значень;
 - 2.3) метод обчислення сигма-нормованих ваг коефіцієнтів регресії.

Цей акт не є документом для фінансових розрахунків.

Голова Правління ПАТ «БАНК ФОРВАРД»





А.В. Кисельов

Виконавець:
Солошенко О.М.
(044) 393-04-29

ПАТ «Банк Форвард»
вул. Санжарівського, 705, Київ, 01032, Україна
тел.: +380 (44) 390 95 95, тел./факс: +380 (44) 390 88 11
e-mail: info@forward-bank.com
ідентифікаційний код 34186061

PJSC "Forward Bank"
705 Saksaganskogo str., Kyiv, Ukraine, 01032
tel.: +380 (44) 390 95 95, tel./fax: +380 (44) 390 88 11
e-mail: info@forward-bank.com
Identification code 34186061

Рисунок А.1 – Копія акту впровадження в ПАТ «БАНК ФОРВАРД»

Додаток Б

Принцип збереження чисел у форматі IEEE 754

Даний стандарт для збереження чисел з плаваючою комою передбачає, що кількість біт відведених на збереження числа розбивається на: (а) біт s для збереження знаку, (б) біти $\{e_i\}_{i=1}^E$ для збереження порядку (експоненти), (в) біти $\{m_j\}_{j=1}^M$ для збереження значущої частини (мантиси). Зміщення експоненти розраховується за допомогою такої формули: $bias(E) = 2^{E-1} - 1$, тоді дійсне число або його наближення представляється за допомогою формули:

$$x = (-1)^s \left(1 + \sum_{j=1}^M m_j 2^{-j}\right) \cdot 2^y,$$

де:

$$y = \sum_{i=1}^E e_{E-i+1} 2^{i-1} - bias(E),$$

тобто використано формат «big-endian», де перший розряд – найбільший.

Для типу даних одиначної точності («float») розміру 4 байти, маємо один знаковий біт та $E = 8$ і $M = 23$, а для типу даних подвійної точності («double») розміру 8 байтів маємо один знаковий біт та $E = 11$ і $M = 52$. Приклад формату одиначної точності у випадку збереження числа 1,5 зображено на рисунку Б.1.

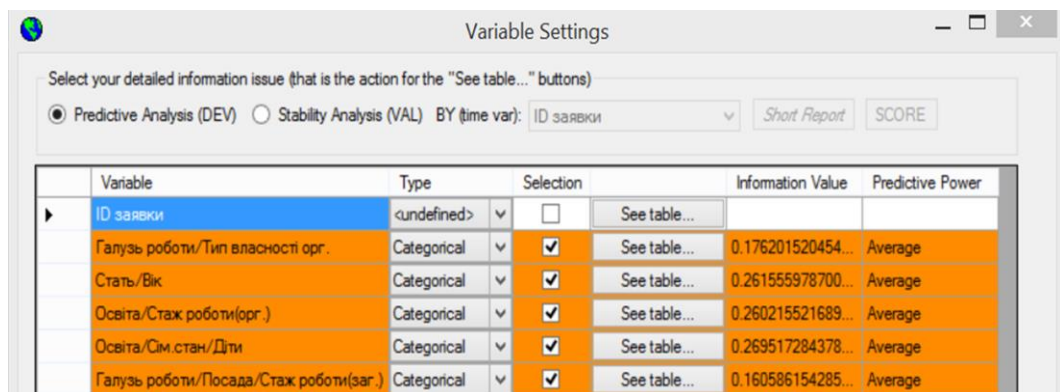
s	e[1]	e[2]	e[3]	e[4]	e[5]	e[6]	e[7]	e[8]	m[1]	m[2]	m[3]	m[4]	m[5]	m[6]	m[7]	m[8]	m[9]	m[10]	m[11]	m[12]	m[13]	m[14]	m[15]	m[16]	m[17]	m[18]	m[19]	m[20]	m[21]	m[22]	m[23]
0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Рисунок Б.1 – Збереження числа 1,5 у форматі одиначної точності («float»)

Додаток В

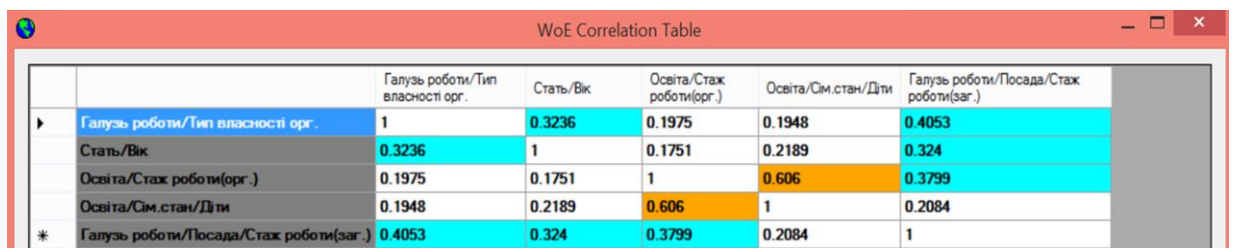
Інформаційна статистика та кореляції на всій множині заявок

На рисунку В.1 наведено переобчислені значення інформаційної статистики з урахуванням відхилених заявок для аплікаційного кредитного скорингу для роздрібного споживчого кредитування, а на рисунку В.2 зображено кореляційну матрицю з урахуванням відхилених заявок відповідно.



Variable	Type	Selection	Information Value	Predictive Power
ID заявки	<undefined>	<input type="checkbox"/>		
Галузь роботи/Тип власності орг.	Categorical	<input checked="" type="checkbox"/>	0.176201520454...	Average
Стать/Вік	Categorical	<input checked="" type="checkbox"/>	0.261555978700...	Average
Освіта/Стаж роботи(орг.)	Categorical	<input checked="" type="checkbox"/>	0.260215521689...	Average
Освіта/Сім.стан./Діти	Categorical	<input checked="" type="checkbox"/>	0.269517284378...	Average
Галузь роботи/Посада/Стаж роботи(sar.)	Categorical	<input checked="" type="checkbox"/>	0.160586154285...	Average

Рисунок В.1 – Переобчислені значення інформаційної статистики з урахуванням відхилених заявок



	Галузь роботи/Тип власності орг.	Стать/Вік	Освіта/Стаж роботи(орг.)	Освіта/Сім.стан./Діти	Галузь роботи/Посада/Стаж роботи(sar.)
Галузь роботи/Тип власності орг.	1	0.3236	0.1975	0.1948	0.4053
Стать/Вік	0.3236	1	0.1751	0.2189	0.324
Освіта/Стаж роботи(орг.)	0.1975	0.1751	1	0.606	0.3799
Освіта/Сім.стан./Діти	0.1948	0.2189	0.606	1	0.2084
* Галузь роботи/Посада/Стаж роботи(sar.)	0.4053	0.324	0.3799	0.2084	1

Рисунок В.2 – Переобчислені значення кореляційної матриці з урахуванням відхилених заявок

Очевидно, що при порівнянні з відповідними рисунками 4.3 та 4.5 дещо простежується тенденція до слабо помітного посилення взаємозв'язків як між вхідними змінними та цільовою змінною, так і між вхідними змінними посередництвом використання ваг категорій змінних (WoE).