НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 004.942:519.216.3

До захисту допущено
В. о. завідувача кафедри ММСА
О.Л.Тимощук
«___» _____ 2018 р.

Магістерська дисертація

на здобуття ступеня магістра за спеціальністю 124 Системний аналіз
на тему: «Система кредитного скорингу позичальників кредитів на основі інтелектуального аналізу даних»

Виконав:
студент ІІ курсу, групи КА-72 мп
Ревва Роман Володимирович _____

Керівник: Професор кафедри ММСА
д.т.н, професор, Бідюк П.І. _____

Рецензент: професор кафедри ЗЗІ
КПІ ім. І.Сікорського,
д.т.н., професор Архипов О.Є. _____

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

Київ
2018

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти — другий (магістерський)
Спеціальність (спеціалізація) — 124 «Системний аналіз» («Системний аналіз і управління»)

ЗАТВЕРДЖУЮ
В. о. завідувача кафедри
ММСА
О. Л.
Тимощук
«___» _____ 2018 р.

ЗАВДАННЯ

на магістерську дисертацію студенту Ревві Роману Володимировичу

1. Тема дисертації: «Система кредитного скорингу позичальників кредитів на основі інтелектуального аналізу даних», науковий керівник дисертації Бідюк Петро Іванович, доктор технічних наук, професор, затверджені наказом по університету від «07» листопада 2018 р. № 4121-с

2. Термін подання студентом дисертації: _____

3. Об'єкт дослідження: позичальники кредитів, представлені статистичними даними з вибраними характеристиками

4. Предмет дослідження: принципи та методи побудови та аналізу скорингових моделей, математичні моделі, методи і критерії оцінювання адекватності скорингових моделей та методи побудови скорингової карти

5. Перелік завдань, які потрібно розробити:

1) Огляд технічної літератури за темою роботи;
2) Вивчення сучасних процесів при впровадженні системи кредитного скорингу позичальників кредитів у банку;
3) Дослідження актуальності теми дослідження;

4) Вибір методів для побудови моделі та набору вхідних даних;
5) Опис архітектури системи;
6) Детальний огляд всіх обраних методів;
7) Вибір параметрів тренування для кожного з них та критеріїв їх ефективності;
8) Проведення аналізу ринкових можливостей запуску стартап-проекту.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:

1) Схема системи аналізу кредитоспроможності клієнтів у банку;
2) Архітектура системи;
3) Вигляд моделі скорингу в пакеті ПЗ SAS Enterprise Miner;
4) Знімки екрану з результатами аналізу скорингових моделей;

7. Орієнтовний перелік публікацій:

(1) Прогнозування кредитоспроможності клієнтів за скоринговим методом;

8. Дата видачі завдання: _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Отримання завдання на магістерську дисертацію	07.09.2018	
2	Огляд технічної літератури за темою	07.09.2018 - 01.10.2018	
3	Дослідження актуальності вибраної теми	02.10.2018 - 07.10.2018	
4	Вибір методів для побудови моделі	07.10.2018 – 17.10.2018	
5	Пошук наборів вхідних даних	17.10.2018 – 19.10.2018	
6	Опис архітектури системи	19.10.2018 – 23.10.2018	
7	Детальний огляд всіх обраних методів	24.10.2018 – 29.10.2018	
8	Вибір параметрів тренування для кожного з них та критеріїв їх ефективності	30.10.2018 – 02.11.2018	
9	Проведення аналізу ринкових можливостей запуску стартап-проекту	02.11.2018 – 19.11.2018	
10	Написання слайдів для доповіді	20.11.2018 – 26.11.2018	

Студент Р.В. Ревва

Науковий керівник дисертації П.І. Бідюк

РЕФЕРАТ

Магістерська дисертація: 102 с., 25 рис., 25 табл., 1 додаток, 13 джерел.

Об’єкт дослідження – позичальники кредитів, представлені статистичними даними з вибраними характеристиками.

Предмет дослідження – математичні моделі, методи інтелектуального аналізу даних, критерії оцінювання адекватності скорингових моделей та методи побудови скорингової карти.

Методи дослідження – методи інтелектуального аналізу даних, нейронні мережі, регресійний аналіз, статистичні методи аналізу даних, методи класифікації, методи побудови скорингових моделей.

Метою роботи є аналіз системи кредитного скорингу на основі методів та моделей інтелектуального аналізу даних, а також, їх порівняння з існуючими методами кредитного скорингу.

В роботі проведено огляд основних підходів побудови скорингових моделей, розглянуто та проаналізовано методи нейронних та байєсівських мереж. Було проаналізовано результати моделювання та оцінювання задля обґрунтуваного вибору найкращої моделі для оцінки кредитоспроможності клієнтів.

Результатом роботи є визначення методів кредитного скорингу та розробка архітектури системи, розробка якої вплине на зниження кредитного ризику банків, а, відповідно, і підвищення його кредитного рейтингу, що в свою чергу матиме системний вплив на банківську систему України.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, КРЕДИТНИЙ СКОРИНГ, ПРОГНОЗУВАННЯ, СКОРИНГОВА МОДЕЛЬ, РЕГРЕСІЯ, СКОРИНГОВА КАРТА, НЕЙРОННА МЕРЕЖА.

ABSTRACT

The topic: The credit scoring system of credit borrowers based on the intelligent data analysis.

Master's thesis: 103 p., 25 fig., 25 tab., 1 application, 13 sources.

Object of the study – loans borrowers represented by the statistics of selected characteristics.

Subject of research - mathematical models, methods of data analysis, criteria for assessing the adequacy of scoring models and methods of constructing a scorecard.

Methods of research - methods of data analysis, neural networks, regression analysis, statistical methods of data analysis, classification methods, methods of constructing scoring models.

The aim of the work is to analyze the system of credit scoring on the basis of methods and models of intellectual data analysis, as well as their comparison with existing methods of credit scoring.

In the work the review of the main approaches of constructing scoring models was carried out, methods of neural and Bayesian networks were considered and analyzed. The results of modeling and evaluation were analyzed in order to justify the choice of the best model for assessing the creditworthiness of clients.

The result of the work is to determine the methods of credit scoring and the development of the architecture of the system, the development of which will affect the reduction of credit risk of banks, and, accordingly, increase its credit rating, which in turn will have a systemic impact on the banking system of Ukraine.

DATA MINING, CREDIT SCORING, FORECASTING, SCORING MODELS, SCORECARDS, NEURAL NETWORK.

ЗМІСТ

ПЕРЕЛІК ПРИЙНЯТИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

МНК – метод найменших квадратів;

БСУ – банківська система України;

ВВП – валовий внутрішній продукт;

ІАД – інтелектуальний аналіз даних;

НБУ – Національний банк України;

ПП – програмний продукт;

ПЗ – програмне забезпечення;

САПП – середня абсолютна похибка в процентах;

СКП – сума квадратів похибок;

СПП – середня похибка в процентах;

ФЕП – фінансово-економічні процеси;

AIC – Akaike info criterion (інформаційний критерій Акайке);

ABT – Analytical Base Table;

BSC – Bias-Schwarz criterion (критерій Байєса-Шварца);

DW – Darbin-Wotson (статистика Дарбіна-Уотсона);

MAPE – mean absolute percent error (середня абсолютна похибка в процентах);

ІАД – інтелектуальний аналіз даних

MAE – mean absolute error (середня абсолютна похибка);

R^2 – коефіцієнт множинної детермінації;

RSME – root mean squared error (стандартне відхилення залишків, середньоквадратична помилка);

SSE – sum of squared errors (сума квадратів похибок);

ВСТУП

Робота присвячена системному вирішенню актуальної для України проблеми банківського ризик-менеджменту, зокрема задач анлізу та прогнозування кредитоспроможності клієнтів банку. Основною сучасною системною методологією прогнозування кредитних ризиків є кредитний скоринг, що полягає у розробці математичних моделей спеціального типу – скорингових моделей та скорингових карт, метою яких є прогнозування майбутнього стану заборгованості позичальника або прогнозування довільних поведінкових показників по договору, клієнту, виходячи з соціально-демографічних характеристик, параметрів кредитного продукту, минулих поведінкових індикаторів, даних щодо транзакцій та ін. Оскільки найбільш важливими складовими системи кредитного скорингу є скорингові моделі, метою роботи є вибір, побудова та навчання найбільш універсальної моделі, що допоможе прийняти зважене рішення щодо видачі або невидачі кредиту клієнтові. Також у роботі розглядається перспективна можливість застосування методів текстової аналітики соціальних мереж для підвищення точності скорингових моделей у режимі реального часу.

Відомими сучасними теоретиками і практиками в області управління ризиками є професори Ю.П. Зайченко, В.М. Подладчіков, Н.Д. Панкратова, Джонатан Н. Крук, Лін С. Томас, Девід Дж. Хенд, Л.М. Любчик, доктор Елізабет Мейз, дослідники Наім 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». 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.

РОЗДІЛ 1 АКТУАЛЬНІСТЬ ДОСЛІДЖЕННЯ ТА ПРОБЛЕМАТИКА КРЕДИТОСПРОМОЖНОСТІ НАСЕЛЕННЯ

1.1 Аналіз кредитного ринку та банківської системи україни та актуальність дослідження

У фінансовому ризик-менеджменті, у відповідності зі стандартною класифікацією, головними загрозами для благополуччя фінансового інституту або установи 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 ризики, бухгалтерські ризики, податкові ризики, ризики репутації, ризики дій регулюючих органів) [1].

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 [1].

Кредитний ринок в Україні є основним сегментом фінансового ринку, що передбачає можливість швидкої мобілізації фінансових ресурсів суб'єктами господарювання. Наявність розвинутого та ефективно-функціонуючого кредитного ринку є основою активізації підприємницької діяльності та економічного розвитку держави в цілому. Переваги кредитного ринку зумовлені функціональним потенціалом основних суб'єктів цього ринку — комерційних банків, які не тільки опосередковують рух фінансових ресурсів, а й певною мірою продукують їх. Концентрація і координація фінансових ресурсів на кредитному ринку здійснюється комерційним банком. У такому випадку банки виступають як у ролі покупця, так і ролі продавця на кредитному ринку. Кількість банків у банківській системі України за період 2012-2016 рр. зменшилась майже удвічі (табл. 1.1). Станом на 01 січня 2017 року ліцензію Національного банку України мали 96 банківських установ (в т.ч. 38 банків з іноземним капіталом). З початку 2016 року кількість функціонуючих банківських установ скоротилася на 21. Загалом, з початку 2014 року внаслідок погіршення платоспроможності до 82 банківських установ було запроваджено тимчасову адміністрацію. В чотирьох тимчасова адміністрація продовжує працювати, щодо одного банку (ПАТ «АСТРА БАНК») прийнято рішення про припинення тимчасової адміністрації та призначення куратора.

Таблиця 1.1 – Основні показники функціонування кредитного ринку впродовж 2012–2016 рр.

Показники	**2012**	**2013**	**2014**	**2015**	**2016**	**Темп приросту 2016/2012**
Кількість банків, од.	176	180	163	117	96	0,56
ВВП, млн.грн.	1404669,00	1465198,00	1586915,00	1988544,00	2383182,00	1,70
Активи, млн.грн	1127179,38	1277508,65	1316717,87	1252570,44	1274731,58	1,13

Кредитний портфель, млн.грн	721062,90	832633,87	895117,78	779402,66	630009,60	0,87

Продовження таблиці 1.1

Середньозважена річна процентна ставка за кредитами банків України, у національній валюті, %	18,30	15,90	17,20	21,30	18,30	1,00
Частка проблемних кредитів у заг. обсязі кредитів, %	16,54	12,89	13,25	28,03	30,47	1,84
Частка кредитів банків у ВВП, %	51,33	56,83	56,41	39,19	26,44	0,51
Частка кредитів банків у активах, %	63,97	65,18	67,98	62,22	49,42	0,77

В цілому, щодо 80 банків вже було прийнято рішення про ліквідацію. Щодо 9 банківських установ рішення про відкликання банківської ліцензії та ліквідацію було прийняте без попереднього запровадження тимчасової адміністрації. Також, починаючи з 01.06.2016 р. Правлінням Національного банку України було прийнято декілька рішень про надання згоди на самоліквідацію банківських установ. Така ситуація є наслідком виконання Комплексної програми розвитку фінансового сектору України, яку проводить НБУ. Програма передбачає очищення банківського сектору, а саме: виведення з ринку неплатоспроможних фінансових установ; оцінка якості активів, стрес-тестування, та рекапіталізація банків; оприлюднення кінцевого власника, кредити пов’язаним особам.

Обсяг активів банківської системи України у 2012-2014 роках мав тенденцію до збільшення. Проте за підсумками 2015 року активи банківської

системи зменшились на 5,12% (64 147 млн.грн.). Водночас у 2016 році цей показник зріс на 1,77 %. Ключовим фактором коливання обсягу активів банківської системи в досліджуваному періоді виступала динаміка курсу національної валюти, поряд із виведенням частини комерційних банків з ринку, відтоком клієнтських коштів та скороченням кредитних портфелів. Переважний вплив курсових різниць пояснюється тим, що станом на 01.04.2016р. частка валютних активів БСУ відповідає 48,3%. Протягом 2012-2016 рр. кредитний портфель БСУ зменшився на 13% (91,05 млрд. грн.). Стрімке зменшення почалося з 2015 року – 12,9% (115,7 млрд.грн.) відносно 2014 р. За підсумками 2016 року обсяг кредитного портфелю банківської системи України становив 630 млрд.грн, що на 19% менше обсягу кредитного портфелю у попередньому році (рис. 1.1).

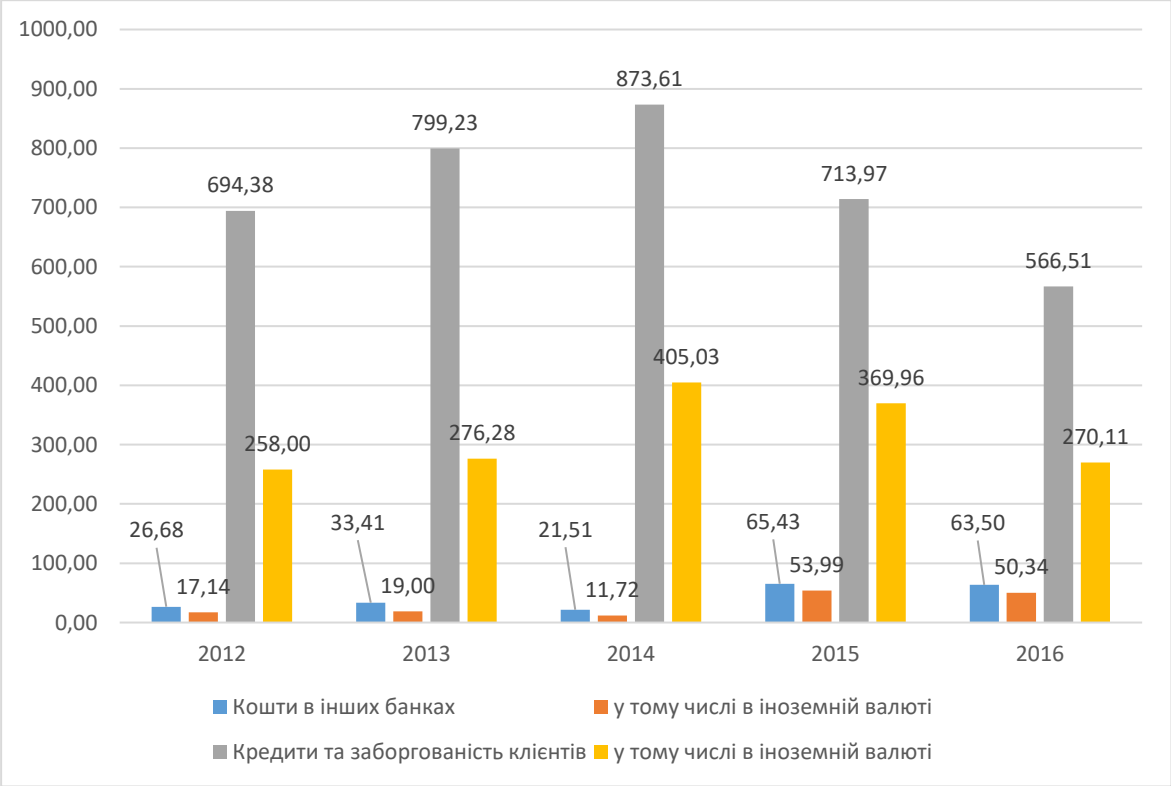

Рисунок 1.1 – Динаміка кредитного ринку впродовж 2012-2016рр.

Кредитна активність банківських установ залишається низькою, що обумовлено обмеженим колом надійних позичальників, високим рівнем

невизначеності щодо подальшого економічного розвитку та високою вартістю кредитного ресурсу. Згідно з даними НБУ середньозважена річна процентна ставка за кредитами банків України, у національній валюті у 2015 році становила 21,3%, проте у 2016 році знизилась до рівня 2012 року (18,3%).

З огляду на девальвацію національної валюти та, враховуючи суттєву частку валютних кредитів в клієнтському кредитному портфелі банків, частка проблемних кредитів у загальному обсязі кредитів БСУ зросла з 16,54% у 2012 році до 30,47% у 2016 році. Частка кредитів банків у ВВП країни також суттєво зменшилась – з 51,33 % у 2012 р. до 26,44% у 2016 р. Також починаючи з 2015 року зменшилась частка кредитів по відношенню до активів БСУ до 49,42% у 2016 році.

Однією з основних причин нинішнього процесу ліквідації банківських установ є збільшення частки проблемних кредитів населення внаслідок недостатньо якісної системи оцінювання кредитоспроможності клієнтів банку.

Обсяг непрацюючих кредитів банківської системи у вересні 2018 року збільшився ще на 6,869 млрд грн порівняно з серпнем 2018 року. Таким чином, станом на 1 жовтня 2018 року обсяг проблемних кредитів в Україні досяг рекордних 662,782 млрд грн. Найбільший приріст проблемних кредитів зафіксований в портфелі ПриватБанку – на 3,5% або 8,236 млрд грн (до 243,696 млрд грн).

Таким чином, одним із найбільш істотних недоліків та слабких місць банківської системи України на сьогоднішній день є неможливість адекватної оцінки кредитних ризиків для більшості банків, результатом чого є неефективна політика управління ризиками. Натомість, формування ефективної політики управління кредитними ризиками банку потребує вдосконалення методів його оцінки для досягнення фінансової стійкості кожної кредитної організації та стабільного розвитку банківської системи вцілому.

За останні роки стало можливим накопичувати великі масиви даних, які надалі можна використовувати для побудови моделей, що нададуть можливість аналізувати та прогнозувати кредитоспроможність позичальника в режимі «онлайн» з максимальною точністю, а також навіть як відповідь прямий запит позичальника, що дозволить зменшити навантаження на спеціалістів з видачі кредитів та, власне, на локальні відділення банків. Такі складні системи потребують розробки спеціального програмного забезпечення, що складається з багатьох взаємопов'язаних компонентів, серед яких найважливішим є створення скорингової моделі та скорингової карти, що описують процес прийняття рішення щодо видачі кредиту клієнтові.

Сьогодні, майже кожна банківська установа використовує пакети прикладного програмного забезпечення або власні засоби автоматизації та підтримки процесу прийняття рішень на основі не тільки інформації, а й історичних знань, отриманих з внутрішніх або зовнішніх баз даних, що містять дані щодо власної історії кредитування, клієнтів, транзакцій, опису спеціальної інформаціїї, інтерпретацій, внутрішньої термінології та ін. В умовах стрімкого збільшення об'ємів накопичених даних та розвитку технологій обробки великих масивів даних (BigData), поглиблена автоматизація процесу оцінювання кредитоспроможності населення може вивести роботу кредитних відділів банківських установ України на якісно новий рівень.

1.2 Огляд існуючих методів і моделей розв'язання задачі

Кредитний скоринг (credit scoring, від англ. score – рейтинг) або аплікаційний скоринг (application scoring) – це методологія оцінювання кредитоспроможності потенційних позичальників у ризик-менеджменті [4],

або методологія класифікації потенційних клієнтів (контрагентів) банку по ступеню (рівню) ризику [2], або набір моделей прийняття рішень та основоположних технік, що допомагають кредиторам в процесі вирішення питання надання споживчого кредиту [3]. Скоринг – методологія оцінювання кредитоспроможності або майбутньої поведінки на рівні клієнтів або договорів, як потенційних, так і існуючих [4], тому існує багато категорій скорингу: кредитний (аплікаційний) скоринг, поведінковий скоринг, скоринг виявлення та попередження шахрайства, колекторський скоринг, інші численні категорії скорингу, що залежать насамперед від набору вхідних параметрів:

- аплікаційний скоринг – оцінка кредитоспроможності позичальників, що звернулися до банку для отримання кредиту за результатами аналізу анкети, отриманої від позичальника під час подання заявки;
- колекторський скоринг – визначення пріорітетних напрямків роботи з позичальниками, стан яких класифіковано як «незадовільний». Цей вид скорингу дозволяє проводити системну роботу з простроченною заборгованістю до моменту її передачі у колекторське агентство;
- поведінковий скоринг – динамічна оцінка стану кредитоспроможності існуючого клієнта-позичальника на основі даних щодо рахунків і транзакцій, таким як графік погашення заборгованості, обороти за поточними рахунками, наявність нових кредитів, графік зміни тарифних планів, тощо. Такий вид скорингу дозволяє визначити ліміти кредитування, маркетингові ходи та дії, що можуть бути застосовані до клієнта;
- fraud-скоринг – оцінка ймовірності шахрайства потенційного позичальника.

З іншого боку, при побудові скорингових моделей можуть використовуватись дані різної природи, тому розглянемо також наступну класифікацію:

- апріорний скоринг – побудова моделей на основі статистичних даних (макроекономічні показники, дані держстатистики, результати перепису населення, тощо), що використовуються для оцінки параметрів моделі, що в свою чергу використовується для оцінювання кредитоспроможності позичальника.
- апостеріорний скоринг – побудова моделей на основі історичних даних про клієнтів кредитної організації. Такі дані являють собою таблицю агрегованих показників, як-от кількість непогашених кредитів, середній термін погашення, дані щодо заявок, тощо).

У зв’язку зі стрімким зростанням популярності кредитних карт, стало очевидно суттєве збільшення часу на прийняття рішення щодо доцільності кредитування клієнта. Цей факт відкрив двері для скорингового методу, що являв собою формалізацію знань кредитора про позичальника. У кінці 50-х років почав розвиватися ринок розробки та підтримки автоматизованих скорингових систем. Першою компанією, що запропонувала скорингову систему, була, заснована у 1956-му році, Fair, Isaac & Co. (зараз відома як FICO). Основні риси концептуальної основи сучасного споживчого кредитного скорингу були встановлені у 1964-му році, про що свідчить праця «Опис реалізації і функціонування системи кредитного скорингу у фінансовій установі» (BOGESS, 1967). [5] Після введення в систему, заявки автоматично оброблювалися, отримували відповідний скоринговий бал, та у разі відповідності внутрішній політиці, автоматично схвалювалися. Після введення системи в експлуатацію, середній час прийняття рішення щодо кредитування одного клієнта зменшився з одного тижня до 24 годин. Система стала першою автоматизованою комп’ютерною скоринговою системою, що дозволяла аналізувати набори данних за допомогою складних алгоритмів множинної регресії, та зумовила розвиток більш точних моделей у наступні десятиріччя.

З розвитком скорингових моделей, реалізація масових кредитних продуктів охарактеризувалася алгоритмізацією. Більша частина знань кредитора про позичальника отримується з бази даних, використовуючи велику кількіть методів статистики, математичного аналізу та інтелектуального аналізу даних (ІАД). Скорингові моделі почали створюватися, використовуючи граничні умови, задані кредитором, а також, математично формалізовані правила та формули. Усі вищенаведені особливості дозволили значно підвищити рівень автоматизації процесу прийняття рішень, що в свою чергу дозволило істотно зменшити час, необхідний для прийняття рішень.

Таким чином, серед переваг скорингових систем оцінювання кредитних ризиків можна виділити швидкість прийняття рішень щодо видачі кредиту, можливість постійного вдосконалення оцінок кредитних ризиків, зменшення суб'єктивності процесу прийняття рішень, точність оцінки ризику за результатами аналізу ретроспективних даних інших клієнтів як юридичних, так і фізичних, підвищення ефективності та швидка адаптація до нових умов ринку. Натомість, серед недоліків можна визначити наслідки невірно вибраної моделі, недостатню точність наданої інформації, а також, неможливість управління якістю рішень, що приймаються. Наведі вище недоліки існуючих методів скорингу кредитного ризику визначають проблеми ефективного впровадження автоматизованих скорингових систем в українську банківську практику. Для вирішення наведених вище проблем необхідно в першу чергу проводити вибір системи кредитного скорингу залежно від потреб і можливостей банку, максимально автоматизувати процес оцінювання кредитоспроможності та прийняття рішень щодо видачі кредиту. Наразі існує декілька підходів до практичної реалізації скорингових систем у банках України:

- розробка скорингу власними ресурсами банку. Такі системи найчастіше являють собою набір правил у Microsoft Excel, що має

мінімальні можливості для моделювання, а також, автоматизованої інтеграції з іншими банківськими системами, однак, для малих банківських установ з невеликими обсягами кредитних операцій, такі системи є раціональним, адже майже не потребують витрат на впровадження;

- розробка скорингової системи та методології сторонніми компаніями-вендорами. Такий варіант дуже розповсюджений у ровзинутих банках Західної Європи. Система створюється для конкретного банку, виходячи з кількості операцій, клієнтів, взаємодії з іншими системти та являє собою повноцінну систему обробки даних, оцінки ризиків, побудови статистичних звітів, моделювання з використанням засобів математичного та статистичного апарату, методів інтелектального аналізу даних, та ін.

Одним з головних недоліків другого підходу для української банківської системи є висока їх вартість та складність впровадження. Зважаючи на це, основною вимогую для таких систем є оптимізація кредитного портфелю шляхом зменшення частки проблемних кредитів та зростання обсягів кредитування. Саме тому найбільш важливим для оцінки таких систем є показник ROI (Return on Investment), адже збільшення прибутку від оптимізації кредитного портфелю за перші кілька років має покривати високу вартість впровадження. Крім того, для забезпечення можливості прийняття рішень виходячи з аналізу тенденцій кредитного ринку країни, необхідна наявність алгоритму визначення актуальності моделі, а також регулярний перегляд скорингової методики. Важливим елементом також є ретроспективний аналіз проведених кредитних операцій.

Для аналізу кредитоспроможності заявника за допомогою скорингової системи, необхідні демографічні, фінансові, а також, соціальні дані, що надаються заявником під час заповнення анкети на отримання кредиту, або під

час діалогу у відділенні в усній формі. Оскільки такі дані є доволі суб'єктивними, необхідна наявність системи перевірки наданих заявником даних, що може бути реалізована використовуючи історичні дані щодо кредитних операцій клієнта із зовнішніх джерел – так званих Бюро Кредитних історій, що містять дані кредитних історій багатьох осіб.

Хоча найбільший вплив на результат роботи скорингової системи має вибір правильної скорингової моделі, досить важливим є формат відображення результатів скорингу для їх подальшої інтеграції з іншими системами. Іншим фактором, що впливає на дизайн системи кредитного скорингу є відповідність регуляторним нормам та стандартам. Саме тому, одним з найбільш популярних форматів відображення результатів кредитного скорингу є скорингові карти, що дозволяють за результатами моделювання співставити певний скоринговий бал кожному клієнтові.

Процесс трансформації неперервних змінних за допомогою групування дозволяє користувачеві аналізувати та враховувати залежність цільової змінної від кожного предиктора, що допомагає не лише визначити які саме змінні найбільше впливають на цільову змінну, але й допомогають визначити характер впливу, що в свою чергу дозволяє користовувачеві перевіряти наявну стратегію, кредитну політику та в подальшому їх вдосконалювати. Наведений вище процес також дозволяє коригування взаємозв'язків з ризиками за результатами експертних оцінок. Наприклад, нестандартна поведінка певних груп та ін. відхилення, що спричинюють сприйняття ризикових груп клієнтів як надіних, можуть бути скориговані за допомогою зменшення відповідних вагових коефіцієнтів Weight of Evidence. Зменшення вагових коефіцієнтів також дозволяє виділити знизити вплив змінних, який неможливо пояснити з точки зору бізнесу.

Отже, формат скорингових карт є дуже простим для інтерпретації, пояснення та подальшого використання. Причини низьких або високих

скорингових балів можуть бути легко представлені регуляторним органам або зовнішнім аудиторам на вимогу, а також, співробітникам.

Цей інтуітивно-зрозумілий формат також є простим при валідації або зміні методики скорингування.

1.3 Деякі комп'ютерні системи для виконання інтелектуального аналізу даних та побудови скорингових моделей

Використання методів кредитного скорингу для оцінювання кредиспроможності клієнтів банків важко уявити без пакетів статистичного прикладного програмного забезпечення та систем ІАД, які поєднують можливості отримання, обробки та завантаження даних (ETL), побудови та кастомізації складних моделей, засобів порівняння моделей використовуючи різні статистичні критерії одначасно та зручний інтерфейс користувача. Розглянемо найбільш поширені програмні засоби, що застосовуються під час розв'язання задач кредитного скорингу: статистичних пакетів SAS та Python.

1.3.1. – SAS

SAS Institute (SAS) – американська компанія-вендор аналітичного програмного забезпечення. SAS розробляє набір прикладного програмного забезпечення, що дозволяє отримати дані з різних джерел, маніпулювати даними, проводити прогнозний аналіз та будувати складні математичні

моделі для подальшого спрощення процесу прийняття рішень в бізнесі. Система SAS об’єднує в собі як новітні технології інтеграції, обробки та збереження даних, так і найпотужніші аналітичні інструменти. Для кредитного скорингу використовуються наступні продукти SAS:

- SAS Credit Scoring for Enterprise Miner
- SAS Credit Scoring for Banking

Розглянемо обидва продукти детальніше. SAS Enterprise Miner - це інтегрований компонент системи SAS, створений для виявлення в масивах даних інформації, необхідної для прийняття рішень.

Розроблений спеціально для пошуку та аналізу прихованих закономірностей у даних (data mining) SAS Enterprise Miner включає в себе ефективні методи статистичного аналізу, відповідну методологію виконання проектів дослідження даних (SEMMA) і зручний графічний інтерфейс користувача. SEMMA пропонує ряд загальних принципів побудови концепції проекту, а SAS Enterprise Miner надає аналітику багатий набір інструментів для використання на кожному етапі проекту.

Компонент SAS Credit Scoring for Enterprise Miner дозволяє використовувати для побудови скорингових моделей додаткові функції, такі як “Reject Inference”, “Scorecard”, “Credit Exchange”, “Interactive Grouping”, що призначені виключно для задач кредитного скорингу.

Компонент SAS Credit Scoring for Banking – корпоративне рішення, призначене виключно для кредитного скорингу та являє собою пакет програмного забезпечення, незалежного від SAS Enterprise Miner, що дозволяє поєднати найбільш ефеткивні методи прогнозного моделювання, обробки даних та механізму побудови звітів для спрощення процесу прийняття рішень щодо кредитоспроможності клієнтів за скоринговим методом.

За своєю суттю SAS, як програмний продукт, призначений для обробки даних і вирішує чотири основні завдання:

- отримання даних;
- керування даними;
- аналіз даних;
- представлення даних.

Отримання даних – під цим терміном будемо розуміти різні способи придбання даних – це і безпосереднє введення, і читання зовнішніх файлів з даними та передача наборів даних з інших прикладних програм.

Керування даними – впорядкування даних їх перевірка і виправлення помилок, зберігання, перетворення.

Аналіз даних – побудова звітів, графіків, таблиць, статистична чи інша обробка даних, на підставі якої можна прийняти те чи інше рішення.

Представлення даних – відображення (візуалізація) результатів аналізу.

При роботі з даними SAS використовує різні типи файлів:

- вихідні («сирі») дані зберігаються у файлах з різними розширеннями;
- Найбільш типові – *. dat, *. txt, *. csv;
- програми, створювані розробником, мають розширення. SAS;
- набори даних, що створюються програмно –. SAS7BDAT.

Крім того SAS дозволяє працювати з даними інших систем з обробки даних (Oracle, Dbase, Excel).

Можливість працювати з широким спектром даних дозволяє SAS бути в першій десятці виробників програмного забезпечення протягом усього часу існування компанії.

Ще однією підставою для успіху SAS на ринку програмного забезпечення є можливість роботи SAS-програм на практично всіх відомих на даний момент платформах, починаючи від мейнфреймів і закінчуючи персональними комп'ютерами.

Така «переносимість» програм з платформи на платформу обумовлюється тим, що 90 % програмного коду є незалежним і тільки 10 % слугує для налаштування на конкретне апаратне забезпечення.

Робоче середовище SAS розроблене для полегшення використання та швидкої розробки і налагодження програм.

Вихідні дані можуть бути представлені в багатьох різних формах. Дані можуть міститися в друкованій формі або в будь-якому комп'ютерному файлі, вони можуть бути представлені у вигляді файлів на великих машинах або міститися в базах і банках даних, створених іншими програмами. Проте, де б вони не перебували, коли вони є, то завжди знайдеться спосіб використання їх в системі SAS.

Всі методи доступу до даних можна розділити на чотири основні категорії:

- введення даних безпосередньо у SAS набір;
- створення набору даних SAS з зовнішнього файлу;
- перетворення «чужого» набору в SAS набір;
- читання «чужого» набору безпосередньо (без перетворення).

Який метод обрати аналітику, залежить від того, де знаходяться дані, якими засобами для їх доставки володіє дослідник і багатьох інших факторів.

До того як почати аналізувати будь-які дані з використанням програмного забезпечення SAS, необхідно, щоб ці дані були організовані у вигляді SAS набору. Отримати SAS набір можна з зовнішнього файлу на кроці даних, зчитавши його, або розмістити дані усередині кроку даних. І в тому і в іншому випадку необхідно вказати, де знаходяться дані для аналізу.

Аналітик може використовувати SAS програми для всіх перерахованих завдань: для доступу до даних, керування ними, аналізу або формування звітів [13].

Найбільшою перевагою SAS як системи для кредитного скорингу є його повна інтеграція з іншими продуктами SAS, такими як SAS Marketing

Automation, SAS Real-Time Decision Manager та ін. Отримані в SAS скорингові моделі можуть бути використані в вищенаведених програмних продуктах для досягнення максимальної автоматизації та точності в процесі прийняття маркетингових, скорингових та інших рішень.

1.3.2 – Python

Python – інтерпретована об'єктно-орієнтована мова програмування високого рівня з строгою динамічною типізацією.[9] Розроблена в 1990 році Гвідо ван Россумом. Структури даних високого рівня разом із динамічною семантикою та динамічним зв'язуванням роблять її привабливою для швидкої розробки програм, а також як засіб поєднання існуючих компонентів. Python підтримує модулі та пакети модулів (бібліотеки), що сприяє модульності та повторному використанню коду. Інтерпретатор Python та стандартні бібліотеки доступні як у скомпільованій так і у вихідній формі на всіх основних платформах. В мові програмування Python підтримується декілька парадигм програмування, зокрема: об'єктно-орієнтована, процедурна, функціональна та аспектно-орієнтована. Завдяки безкоштовності, високій швидкості, інтерпретованості та великій кількості потужних бібліотек, мова Python швидко стала одним з найбільш розповсюджених засобів ІАД. Використання Python для задач кредитного скорингу можливе за наявності бібліотек NumPy, Pandas, SciPy та scikit-learn, які безкоштовно встановлюються разом з пакетом Anaconda. Серед найбільш суттєвих переваг мови Python для розв’язання задач кредитного скорингу можна назвати її кросс-платформеність, можливість повної кастомізації, безкоштовність та велика кількість потужних бібліотек, що містять майже всі відомі на сьогоднішній день алгоритми побудови

прогнозних скорингових моделей. Найбільшим недоліком є віднутність готового in-box рішення для промислового кредитного скорингу: всі операції виконуються через команди мовою Python, що суттєво ускладнює використання кінцевим користувачем.

1.3.3 – FICO

Розробник програмного забезпечення, направлених на прийняття зважених та обґрунтованих рішень. В портфель продуктів FICO входять аналітичні системи для автоматизації, оптимізації та уніфікації процесів прийняття рішень у всіх сферах діяльності підприємства (організації). Штаб-квартира компанії знаходиться в Міннеаполісі, Міннесота (США). Продукти FICO використовуються провідними світовими банками та фінансовими установами США. В фінансовій галузі продукти компанії застосовуються в першу чергу для прийняття рішень на всіх етапах роботи з кредитами. Загалом рішеннями компанії користуються клієнти з понад 80 країн. Компанія FICO також допомагає приватним особам покращити свій кредитний рейтинг через спеціалізований сайт компанії

FICO® Score, доступний в трьох основних споживачів звітності установ, допомагає кредиторам зробити точні, надійні і швидкі рішення кредитного ризику по життєвим циклом клієнта. Основною перевагою є власний критерій оцінки ризику, що відіграє критичну роль в мільярдах рішень кожного року. Для задач кредитного скорингу також використовуються наступні продукти компанії:

- FICO Capstone Decision Accelerator;
- FICO® Score Open Access;
- FICO Score Adoption Services.

Через надто високу вартість, недостатню універсальність та неможливість використання обмеженої версії для академічних робіт, система FICO не розглядається у цій роботі.

1.4 Поняття інтелектуального аналізу даних як області досліджень системного аналізу

Термін «інтелектуальний аналіз даних» (ІАД) походить від англомовного поняття «data mining» [5, 8, 9, 13, 16], 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 [14]. Сучасне поняття «виявлення знань у базах даних» (KDD – Knowledge Discovery in Databases) можна вважати синонімом інтелектуального аналізу даних [15]. 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 [15]. До цього часу обробка та аналіз даних здійснювалися в рамках прикладної статистики, при цьому в основному вирішувалися завдання обробки невеликих баз даних [15]. Інтелектуальний аналіз даних – це процес підтримки прийняття рішень, який ґрунтується на пошуку в даних прихованих закономірностей [9]. 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» [13, 14]. 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 [15]. 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 [15, 16]: (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]: (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.

1.5 Місце моделей оцінювання кредитоспроможності в інтелектуальному аналізі даних

Кредитний скоринг можна вважати одночасно основопороджуючим 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 [16]. 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 [16]. 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 [16]. 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) [16]. Пов’язаною з методологією скорингу є розробка антиспамних фільтрів за допомогою ймовірнісного наївного байєсівського класифікатора [16]. 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) [12]. 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) [16]. 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. У даній дисертаційній роботі надалі наводяться нові можливості бінарного скорингового моделювання.

Висновки до розділу

В умовах курсу на очищення БСУ, а також, поступового зміцнення економіки та зростання показників купівельної спроможності населення, проблема підвищення точності скорингових моделей та розробки потужного, точного, легкого у використанні та, разом з тим, дешевого програмного забезпечення для побудови скорингових моделей та скорингових карт є наразі актуальною для України. Існуючі системи частіше за все є занадто складними – для використання більшої частини таких систем необхідно додатково навчати персонал, користуватися консультаційними послугами вендорів або їх партнерів, що в умовах нинішньої економічної ситуації в Україні призводить до збільшення додаткових витрат. В таких умовах виникає потреба створення простого у використанні програмного забезпечення, що може в режимі реального часу надавати інформацію щодо доцільності кредитування клієнта.

На сьогоднішній день в спеціальній літературі описана досить велика кількість методів прогнозування кредитоспроможності клієнтів. Найбільш популярними є методи, що базуються на лінійних та логістичних регресійних моделях, нейронних мережах, деревах класифікації, методи дискримінантного аналізу, методи математичного моделювання та ін.

Однак кожен з методів має свої переваги і недоліки, які є суттєвими у випадку моделей, що будуть використовуватися в режимі реального часу. Також сьогодні відсутній систематизований підхід до вибору структури математичних моделей та методів прогнозування, а також рекомендацій щодо їх застосування.

Постановка задачі

- Аналіз існуючих методів побудови скорингових моделей
- Розробка методики побудови скорингових моделей

- Вибір методу оцінювання параметрів моделей
- Вибір оптимальної моделі для побудови скорингових моделей в реальному часі.
- Розробка програмного забезпечення
- Виконання і аналіз результатів виконаних обчислювальних експериментів.

РОЗДІЛ 2 ВИБІР МЕТОДІВ І МОДЕЛЕЙ ДЛЯ АНАЛІЗУ КРЕДИТОСПРОМОЖНОСТІ

2.1 Вибір методів та моделей

Для побудови скорингових моделей будемо користуватися анонімізованими даними одного з українських банків, що подані у вигляді вже сформованної ABT-таблиці. Кожен рядок ABT-таблиці відповідає унікальному ідентифікатору клієнта. Кожен стовпець таблиці містить числове значення певного атрибуту клієнта, іноді в агрегованій формі.

Побудова (тренування) моделі відбувається на накопичених історичних даних клієнтів та результатів їх кредитування, використовуючи певний, визначений, алгоритм. Після чого, натренована модель може використовуватися на даних, що надходять у систему в режимі реального часу, для визначення кредиспроможності клієнта в режимі онлайн.

Для кредитного скорингу використовуються декілька класів моделей, що грунтуються на принципово різних математичних методах та підходах.

Розглянемо найбільш поширені класи більш детально.

2.1.1 – Лінійні регресійні моделі. Метод найменших квадратів

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

$$E(y \mid \boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{c})$$

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

Модель множинної лінійної регресії записується таким чином [13]:

$$E(y \mid \boldsymbol{x}) = c_0 + \sum_{i=1}^{M} c_i x_i$$

де M – кількість вхідних змінних, а також $x^T = (x_1\ x_2\ \dots x_i\ \dots x_M)$,

$$c^T = (c_0\ c_1\ c_2\ \dots c_i\ \dots c_M)$$

Якщо доповнити вектор змінних одиничним входом (нульова координата):

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

де для вектору спостережень $\boldsymbol{x}_{obs}$ розмірості вже 1+M маємо $x_0 = 1$.

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

$$\hat{y}(\mathrm{i}) = E\big(y(i)\big|\boldsymbol{X}(i)\big) = \boldsymbol{X}(i)\boldsymbol{c}_{LS}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{c}_{LS}$$

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

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}_{LS}\|^2 \to min$$

Формула обчислення оптимального вектору за допомогою МНК [8]:

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

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).

2.1.2 – Узагальнена ймовірнісна нелінійна регресія. Метод максимальної правдоподібності

У термінах підрозділу 2.1.1 функція регресії f є неперервним диференційовним монотонно зростаючим відображенням F від лінійної функції змінних, тобто $f(\boldsymbol{x}, \boldsymbol{c}) = F(x_{obs}^T, c)$, де $F : R \rightarrow (0; 1)$
Це означає, що функція F є деякою кумулятивною функцією розподілу $F(z) = P(\varepsilon \leq z)$. Введемо позначення $P(c, x_{obs}) = f(x, c) = F(x_{obs}^T c)$, тоді функція правдоподібності матиме вигляд:

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

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 ($\ln L(c) \rightarrow max$):

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

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

Для пошуку c_{MLE} зазвичай використовується метод Ньютона:

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

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

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

Суть логістичної регресії полягає у використанні логістичного перетворення (сигмоїдальної функції), що відповідає логістичному розподілу з нульовим середнім та середньоквадратичним відхиленням, що дорівнює числу $\frac{\pi}{\sqrt{3}}$ [7].

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

Це означає, що логарифм відношення шансів для конкретного прогнозу:

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

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 2.5), то суть моделі логістичної регресії при використанні WoE-перетворень концептуально можна записати таким чином:

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

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 [6-8]:

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

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

Отже, з точки зору впровадження та застосування в системі кредитного скорингу, моделі регресійного типу є простими в застосуванні та інтерпретації, а також не потребують істотних потужностей обчислювальних машин. Крім того, серед переваг регресійних методів, зокрема, методу логістичної регресії, можна виділити наступні:

- можливість роботи з великою кількістю атрибутів;
- надвисока швидкість роботи (робота на великих вибірках);
- за допомогою логістичної регресії легко визначити ймовірності належності до кожного класу.

Натомість, існують і недоліки, серед яких:

- слабка предиктивна здатність у вирішенні задач, у яких залежність цільової змінної від предикторів є нелійною;

Оскільки, для задач кредитного скорингу в українській банківській системі одним із найважливіших критеріїв є легкість інтерпретації залежностей та коефіцієнтів моделей, а також, беручи до уваги кількість клієнтів середньостатистичного банку України, яка становить близько 1 млн. фізичних осіб, логістична регресія є оптимальним вибором для побудови

скорингової моделі. При цьому, дані для прогнозування кредитоспроможності зазвичай не містять нелінійних зв’язків, тому недолік логістичної регресії у цьому випадку не є критичним.

2.2 Сучасний підхід до побудови нейромережевих моделей

Дослідження нейронних мереж було розпочате в 1943 році з публікації, написаної МкКулочем та Піттом [14]. Згідно з проголошеним принципом, необхідно було розробити модель, яка могла б моделювати природну роботу нейрона.

Дендрити, які використовуються для прийняття сигналів нейроном та аксони, що допомагають транслювати оброблювану інформацію в інші нейрони, є найбільш важливими частинами нейрона для застосування у методах інтелектуального аналізу даних. Аксони з'єднуються з дендритами інших нейронів через синапси. За допомогою заданної функції синапси обробляють інформацію, отриману від дендритів, і якщо вхідний сигнал перевищує так званий поріг стимулу, вони направляють інформацію через аксона. Найбільш важливою властивістю нейрона є те, що він безперервно змінюється (тобто змінює свою внутрішню функцію) на основі отриманих дані. Цей процес отримав назву навчання. Синапси грають важливу роль в цьому процесі навчання, оскільки вони здатні посилювати або послаблювати сигнали, що надходять від інших нейронів. У процесі навчання, ваги (коефіцієнти посилення сигналу) змінюються під час проходження через синапси. Зміна ваг у нейронній мережі є основою процесу навчання на математичному рівні.

Нейронні мережі – ефективний інструмент для вирішення задач класифікації, однак має недолік високої схильності до перенавчання (overfitting) [5, 13, 14].

Найпопулярнішою формою архітектури нейронної мережі є багатошаровий персептрон (MLP. Багатошаровий персептрон може мати будь-яку кількість входів та певну кількість прихованих шарів з будь-якою кількістю вузлів. Також, багатошаровий персептрон:

- використовує лінійні функції комбінації у прихованих і вихідних шарах;
- використовує сигмоїдні функції активації у прихованих шарах;
- може мати будь-яку кількість виходів з будь-якою функцією активації;
- має зв'язки між вхідним шаром і першим прихованим шаром, між прихованими шарами та між останнім прихованим шаром та вихідним шаром.

У системі SAS Enterprise Miner можна створювати безліч варіантів моделей багатошарового персептрону. Наприклад, підтримуються прямі зв'язки між входами та виходами.

За умови достатньої кількості даних, достатньої кількості прихованих елементів та достатнього часу для тренувань, MLP з одним прихованим шаром може навчитися наближати практично будь-яку функцію до будь-якого ступеня точності. З цієї причини MLP відомі як універсальні апроксиматори, і можуть бути використані, коли немає достатньо даних про взаємозв'язки вхідних змінних.

Метод зворотного поширення помилки (англ. backpropagation) — метод навчання багатошарового перцептрону. Це ітеративний градієнтний алгоритм, який використовується з метою мінімізації помилки роботи багатошарового перцептрону та отримання бажаного виходу. Основна ідея цього методу полягає в поширенні сигналів помилки від виходів мережі до її входів, в напрямку, зворотному прямому поширенню сигналів у звичайному режимі роботи. Барц і Охонін запропонували відразу загальний метод («принцип подвійності»), який можна застосувати до

ширшого класу систем, включаючи системи з запізненням, розподілені системи, тощо [1]. Для можливості застосування методу зворотного поширення помилки функція активації нейронів повинна бути диференційованою.

Навчання нейронних мереж можна представити як задачу оптимізації. Оцінити — означає вказати кількісно, добре чи погано мережа вирішує поставлені їй завдання. Для цього будується функція оцінки. Вона, як правило, явно залежить від вихідних сигналів мережі і неявно (через функціонування) — від всіх її параметрів. Найпростіший і найпоширеніший приклад оцінки — сума квадратів відстаней від вихідних сигналів мережі до їх необхідних значень:

$$H = \frac{1}{2} \sum_{\tau \in v_{out}} (Z(\tau) - Z^*(\tau))^2$$

,

де $Z^*(\tau)$ — необхідне значення вихідного сигналу.

Метод найменших квадратів далеко не завжди є найкращим вибором оцінки. Ретельне конструювання функції оцінки дозволяє на порядок підвищити ефективність навчання мережі, а також одержувати додаткову інформацію — «рівень впевненості» мережі у відповіді [2].

На кожній ітерації алгоритму зворотного поширення вагові коефіцієнти нейронної мережі модифікуються так, щоб поліпшити рішення одного прикладу. Таким чином, у процесі навчання циклічно вирішуються однокритеріальні задачі оптимізації. Навчання нейронної мережі характеризується чотирма специфічними обмеженнями, що виділяють навчання нейромереж із загальних задач оптимізації: астрономічне число параметрів, необхідність високого паралелізму при навчанні, багато-критеріально вирішуваних завдань, необхідність знайти досить широку область, в якій значення всіх функцій, що мінімізуються близькі до мінімальних. Стосовно решти проблему навчання можна, як правило,

сформулювати як завдання мінімізації оцінки. Обережність попередньої фрази («як правило») пов'язана з тим, що насправді нам невідомі і ніколи не будуть відомі всі можливі завдання для нейронних мереж, і, може, десь в невідомості є завдання, які не зводяться до мінімізації оцінки. Мінімізація оцінки — складна проблема: параметрів астрономічно багато (для стандартних прикладів, що реалізуються на PC — від 100 до 1000000), адаптивний рельєф (графік оцінки як функції від підлаштовуваних параметрів) складний, може містити багато локальних мінімумів.

Незважаючи на численні успішні застосування алгоритму зворотного поширення помилки, він не є панацеєю. Найбільше неприємностей приносить невизначено довгий процес навчання. У складних завданнях для навчання мережі можуть знадобитися дні або навіть тижні, вона може і взагалі не навчитися. Причиною може бути одна з описаних нижче:

- Параліч мережі: У процесі навчання мережі значення ваг можуть в результаті корекції стати дуже великими величинами. Це може призвести до того, що всі або більшість нейронів будуть функціонувати при дуже великих значеннях OUT, в області, де похідна стискаючої функції дуже мала. Так як помилка, що посилається назад у процесі навчання, пропорційна цій похідній, то процес навчання може практично завмерти. У теоретичному відношенні ця проблема погано вивчена. Зазвичай цього уникають зменшенням розміру кроку η, але це збільшує час навчання. Різні евристики використовувалися для запобігання від паралічу або для відновлення після нього, але поки що вони можуть розглядатися лише як експериментальні.
- Локальні мінімуми: зворотне поширення використовує різновид градієнтного спуску, тобто здійснює спуск вниз по поверхні помилки, безперервно підлаштовуючи ваги в напрямку до мінімуму. Поверхня помилки складної мережі сильно порізана і

складається з пагорбів, долин, складок і ярів в просторі високої розмірності. Мережа може потрапити в локальний мінімум (неглибоку долину), коли поруч є набагато більш глибоких мінімумів. В точці локального мінімуму всі напрямки ведуть вгору, і мережа нездатна з нього вибратися. Статистичні методи навчання можуть допомогти уникнути цієї пастки, але вони повільні.

- Розмір кроку: уважний розбір доведення збіжності [15] показує, що корекції ваг передбачаються нескінченно малими. Ясно, що це нездійсненно на практиці, тому що веде до безкінечного часу навчання. Розмір кроку повинен братися кінцевим. Якщо розмір кроку фіксований і дуже малий, то збіжність надто повільна, якщо ж він фіксований і занадто великий, то може виникнути параліч або постійна нестійкість. Ефективно збільшувати крок до тих пір, поки не припиниться поліпшення оцінки в даному напрямку антиградієнта і зменшувати, якщо такого покращення не відбувається. П. Д. Вассерман [9] описав адаптивний алгоритм вибору кроку, який автоматично коректує розмір кроку в процесі навчання. Слід також відмітити можливість перенавчання мережі, що є скоріше результатом помилкового проектування її топології. При дуже великій кількості нейронів втрачається властивість мережі узагальнювати інформацію. Весь набір образів, наданих до навчання, буде вивчений мережею, але будь-які інші образи, навіть дуже схожі, можуть бути класифіковані невірно.

РБФ-мережі, або мережі на основі радіально-базисних функцій – це особливий клас нейронних мереж, що грунтуються на використанні спеціальних радіально-симетричних функцій. Їхня характерна властивість полягає у тому, що відповідь функції монотонно зростає (спадає) з віддаленням від центральної точки. Типовим прикладом радіально-симетричної функції є функція Гауса:

$$h(x) = exp^{(-\frac{(x-c)^2}{r^2}}$$

Радіально-симетричні функції можуть бути використані у великій кількості моделей, але традиційно термін РБФ-мережа застосовується до одношарових мереж з радіально-симетричними функціями, що мають структуру, зображену на рис. 2.2.1.

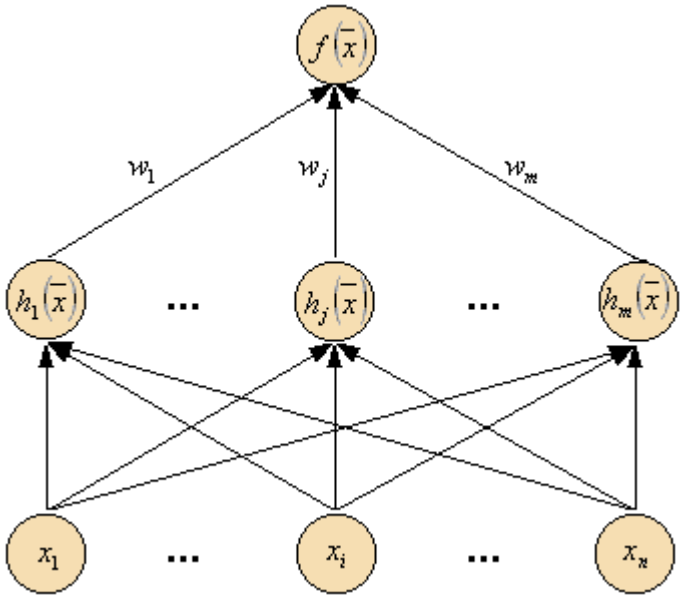

Рисунок 2.2.1 – Схематична модель РБФ-мережі.

Таким чином, вихід мережі RBF є является лінійною комбінацією деякого набору базисних функцій:

$$f(\bar{x}) = \sum_{j=1}^{m} w_j h_j(\bar{x}),$$

де

$$h(\bar{x}) = \exp\left(-\frac{\|\bar{x} - \bar{c}\|^2}{r^2}\right)$$

Таким чином, РБФ-мережа:

- має будь-яку кількість входів;
- як правило, має прихований шар з певною кількістю елементів;

- використовує комбінації радіальних функцій у прихованому шарі на основі квадрата евклідової відстані між вхідним вектором та ваговим вектором;
- як правило, використовує експоненціальну або softmax функції активації в прихованому шарі;
- використовує лінійні функції комбінації у вихідному шарі;
- має будь-яку кількість виходів з будь-якою функцією активації.
- має зв'язки між вхідним шаром і прихованим шаром, а також між прихованим шаром і вихідним рівнем.

Багатошаровий персептрон часто називають розподіленою мережею обробки, оскільки ефект прихованого елемента може бути розподілений по всьому вхідному простору. З іншого боку, гауссові мережі RBF є локальними мережами обробки, оскільки ефект прихованого елемента зазвичай концентрується в локальній зоні, з центром у ваговому векторі.

2.3 Дерева рішень

Основним альтернативним методом класифікації по відношенню до регресійних методів моделювання є рекурсивно-партиційні алгоритми (recursive partitioning algorithms, RPA) у вигляді дерев класифікації (classification trees) – дерев рішень (*decision trees*) [5, 8, 9, 14]. Основу дерев рішень становить індекс ентропії інформації.

Аналогічно інформаційній статистиці вводиться поняття приросту інформації (information gain) при кожному розбитті підмножини (множини) A на менші підмножини Ai , 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 розподілу «*один \ нуль*» («*Good \ Bad*») означена як:

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

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

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 [13], яка може бути спрощена описаними раніше способами.

Дерева рішень – простий інструмент включення корельованих змінних.

2.4 Методика побудови Байєсівських мереж

Теорію мереж Байєса почали активно розвивати на початку 80-х років XX-го століття. Сьогодні їх успішно використовують для розв’язання таких практичних задач: ймовірнісне математичне моделювання процесів і об’єктів різної природи, які функціонують в умовах наявності невизначеностей, прогнозування динаміки розвитку процесів та технологічне передбачення, автоматичне діагностування в техніці та медицині, прийняття рішень у бізнесі і на виробництві, менеджмент фінансово-економічних та інших

ризиків, розпізнавання образів, аналіз причинно-наслідкових зв’язків при дослідженні функціонування складних ієрархічних систем, для створення систем автоматичного керування технічними об’єктами та деяких інших задач. Успішне застосування математичних моделей у формі мереж Байєса зумовлене їх високою гнучкістю, можливістю врахування великої кількості категоріальних і числових змінних (у тому числі невимірюваних), наявністю широкого спектру методів структурного і параметричного навчання, а також формування ймовірнісного висновку у прямому і зворотному напрямах.

Головною перевагою байєсівських моделей перед розглянутими вище моделями є простота моделі з точки зору легкості інтерпретації аналітиком. Наприклад, Байєсівська мережа може явно представляти взаємозв’язки залежності розподілу між будь-якими довільними змінними, тим самим дозволяючи виявляти залежності та причинно-наслідкові зв’язки між усіма змінними, а також, умовного розподілу цільової змінної, на противагу методам нейронних мереж, найбільш сучасні методи яких виявляють зв’язки, які неможливо інтерпретувати з точки зору бізнес-логіки. Оскільки регуляторний тиск на банки з боку НБУ збільшується, легкість інтерпретування процесів стає більш важливим критерієм вибору автоматизованих систем.

Методологія байєсівського аналізу даних та експертних оцінок цілком узгоджується з логікою дій особи, що приймає рішення при аналізі процесів довільної природи, формуванні альтернатив і прийнятті рішень. Апріорна інформація про досліджуваний процес доповнюється даними експерименту, додатковою інформацією якісного або кількісного характеру, що може бути отримана з різних джерел, і на основі інтегрованих знань і даних формується апостеріорний імовірнісний висновок стосовно змінних, параметрів, станів, ситуацій та ін. Байєсівські методи успішно застосовуються на всіх етапах аналізу даних при моделюванні, прогнозуванні та прийнятті рішень. На етапі попередньої обробки даних застосовують ймовірнісну фільтрацію

спостережень, заповнення пропусків, на етапі моделювання — формування структур моделей і оцінювання їх параметрів, а на етапі формування альтернатив — обчислення ймовірнісних висновків (рішень) за допомогою побудованих раніше моделей. Байєсівські методи мають такі переваги: можливість урахування невизначеностей статистичного, структурного і параметричного характеру, поєднання в одній моделі (наприклад, у байєсівській мережі) великої кількості різнорідних змінних, наявність досить гнучких процедур оцінювання параметрів і станів досліджуваних процесів, а також наявність широкого спектра методів формування точних і наближених висновків. До недоліків можна віднести труднощі з отриманням апріорної інформації та відносну складність деяких Байєсівські мережі у системах підтримки прийняття рішень обчислювальних процедур, пов'язаних з числовим інтегруванням, оцінюванням параметрів і формуванням імовірнісних висновків. Стосовно недоліків можна сказати, що в деяких випадках вони дійсно існують і створюють труднощі для дослідника, але з набуттям досвіду використання цих методів та підвищенням якості відповідних обчислювальних процедур аналізу даних і знань рівень та обсяг цих труднощів істотно зменшується.

2.4.1 – Статичні моделі Байєсівських мереж

У загальному випадку байєсівська мережа являє собою пару ⟨G, B⟩, у якій перша компонента G — це спрямований нециклічний граф, що відповідає випадковим змінним і записується як набір умов незалежності: кожна змінна незалежна від її батьків у G. Друга компонента пари B — це множина параметрів, що визначають мережу. Наївний (naïve) та доповнений деревом (tree-augmented, або TAN) байєсівські класифікатори — це

ймовірнісні графічні моделі, що використовуються для формалізованого опису великих масивів даних, які містять невизначеності серед своїх взаємозалежних наборів характеристик. Ці моделі широко використовуються для розв’язання задач сегментації зображень, медичної діагностики та інших задач кластеризації і класифікації на основі статистичних даних. Проблема класифікації полягає у виявленні, до якого класу належить конкретний об’єкт, на основі знань, отриманих після аналізу подібних об’єктів. Кожний елемент описується за допомогою множини змінних, які називають характеристиками або параметрами. Використання наївного байєсівського класифікатора ґрунтується на тому, що всі змінні (характеристики) є незалежними одна від іншої. Це дуже просте уявлення стосовно характеристик системи, але у той же час незалежність, що визначається цією моделлю, не завжди є реалістичною. Модель TAN — це модифікований наївний класифікатор Байєса, який враховує ще один рівень взаємодії між параметрами досліджуваної системи, тобто кожна змінна може залежати від деякої іншої змінної. При цьому залежність між характеристиками моделі TAN є більш реальною, ніж у наївному класифікаторі [7, 8].

2.4.2 – Наївний байєсівський класифікатор

Наївний байєсівський класифікатор — особлива форма ймовірнісної моделі у вигляді байєсівської мережі, яка характеризується тим, що має сильні припущення стосовно незалежності змінних. Ця модель широко використовується для розв’язання задач класифікації. Змінна C являє собою цільову змінну, що може набувати значення $\{1, \dots, C_k\}$, де C_i – стан цільової змінної, а $\{X_1, \dots, X_n\}$ — множина незалежних вхідних змінних процесу, які можуть впливати на цільову. Основним припущенням моделі є те, що всі

вхідні змінні незалежні між собою, але на практиці, при розв’язанні задач кредитного скорингу, виконання умови незалежності спостерігається дуже рідко [8].

2.4.3 – Доповнений деревом байєсівський класифікатор

У реальному світі змінні будь-якої досліджуваної системи корелюють між собою. Якщо модель враховує кореляції між змінними, то точність класифікації, як правило, поліпшується. Одним із варіантів вирішення цієї проблеми є доповнена деревом байєсівська модель. Вона підтримує структуру наївного байєсівського класифікатора і доповнює її додаванням ребер між вершинами, які являють собою змінні досліджуваного процесу, з метою відображення кореляції між змінними. Одночасно такий підхід призводить до підвищення обчислювальної складності алгоритмів. Разом із тим використання наївного байєсівського класифікатора вимагає зберігання тільки умовних ймовірностей належності до станів цільової змінної, а доповнена модель — перебору всіх можливих топологій мережі Байєса. Для того щоб зменшити обчислювальну складність, а також врахувати кореляції між змінними, необхідно накласти обмеження на рівні взаємодії між змінними. Однією з таких моделей є модель класифікатора, розширеного деревом (TAN).

Ця модель вводить обмеження на кількість батьківських змінних: їх може бути не більше двох. У моделі TAN всі вхідні змінні пов’язані з цільовою змінною за допомогою спрямованих ребер. Отже, при визначенні умовної ймовірності $P(C \mid X_1, \dots, X_n)$ до уваги беруться всі незалежні параметри. При цьому кожна змінна у графі може мати двох батьків, а саме: цільову змінну та іншу вхідну змінну, що не є нащадком.

Обчислювальна складність такої моделі значно зменшується, адже кожна змінна може мати не більше двох батьків. Таким чином, TAN має ненабагато більшу обчислювальну складність, ніж наївний байєсівський класифікатор, але при цьому показує вищу точність класифікації. Ключовою особливістю доповненої деревом моделі є її деревоподібна структура. Для того щоб побудувати дерево, необхідно спочатку визначити батька кожної змінної. Крім того, тільки змінні з максимальною кореляцією повинні бути з'єднані одна з одною.

Ще одним важливим поняттям, що відіграє ключову роль при побудові дерева, є взаємна інформація [10]. Для того щоб побудувати дерево, необхідно оцінити кореляцію між кожною парою змінних у системі і додати ребро тільки між тими змінними, які найбільше корелюють. Якщо у досліджуваній системі є N змінних, то відповідний граф матиме N вузлів. Для того щоб отримати деревоподібну структуру, яка з'єднує всі вузли в графі, необхідно додати N − 1 ребро. Крім того, сума ваг усіх цих ребер повинна бути максимальною вагою серед усіх таких деревоподібних структур. Міра кореляції між двома змінними X і Y називається взаємною інформацією і обчислюється за виразом:

$$I_p(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

Для пари змінних ця функція показує, скільки інформації про одну змінну містить інша. Для того щоб побудувати дерево для моделі TAN, використовується умовна взаємна інформація між двома змінними. Це необхідно для того, щоб визначити ребра, які будуть належати до дерева. Умовна взаємна інформація розраховується за формулою:

$$I_p(X;Y \mid Z) = \sum_{x,y} P(x,y,z) \log \frac{P(x,y|z)}{P(x|z)P(y,z)}$$

Алгоритм побудови дерева для моделі TAN складається з наступних кроків:

1) Обчислити взаємну інформацію $I_p(X_i; X_j|C)$ між кожною парою змінних $i \neq j$;
2) Побудувати повний неорієнтований граф, у якому вершини є параметрами $\{X_1, \dots, X_n\}$ і знайти ваги ребер, що з'єднують пари X_i і X_j з використанням значення взаємної інформації $I_p(X_i; X_j|C)$;
3) Побудувати максимально зважене дерево.
4) Для того щоб перетворити отримане неорієнтоване дерево на орієнтоване, необхідно задати цільову зміну як кореневу вершину та визначити напрямки всіх ребер, що виходять назовні від неї.

Оскільки байєсівські мережі є найбільш легкими для інтерпретації аналітиком, цей метод є перспективним для застосування при побудові скорингових моделей.

2.5 Кореляційний аналіз змінних

Одним з найбільш важливих етапів побудови довільних статистичних моделей, особливо регресійного типу [14], є кореляційний аналіз вхідних змінних з метою зменшення кількості вхідних змінних. У даному випадку кореляційна матриця $R \in Mat(M \times M)$ будується для рядів WoE M предикторів:

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

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

де WoEi (n) – WoE-перетворення змінної під номером і 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 дипломної роботи використовуються такі границі абсолютного значення кореляції:

1) $R_{ij} \in [0\%; 25\%]$ – низька кореляція

2) $R_{ij} \in (25\%; 50\%]$ – середня кореляція

3) $R_{ij} \in (50\%; 75\%]$ – висока кореляція

4) $R_{ij} \in (75\%; 100\%]$ – дуже висока кореляція

Для визначення найбільш змінних, що мають найбільшу предикативну силу, застосовується процес групування неперервних змінних, що здійснюється за допомогою WoE, де WoE:

$$WOE = \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

А також, Information Value (IV), де IV:

$$IV = \sum(\text{Event\%} - \text{Non Event\%}) * \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

Значення Information Value визначається для кожної незалежної змінної та дозволяє визначити, які змінні мають найбільшу предикативну силу та мають істотний вплив на цільову змінну. Таким чином, змінні, що мають малу предикативну силу не включаються до скорингової моделі.

Висновки до розділу

В даному розділі розглянуто основні підходи до побудови математичних моделей, на яких базуватиметься розробка скорингової моделі. Розглянуто регресійні моделі, моделі машинного навчання на основі пам'яті, дерев рішень, а також, методи кластеризації.

В якості основної моделі було запропоновано використання нейронних та байєсівських мереж. Ці моделі є найбільш перспективними та дозволяють виявляти приховані взаємозв'язки. В якості альтернативних варіантів побудови моделей розглянуто моделі на основі регресії. Такі моделі є простими, не потребують значних обчислювальних ресурсів, а також, мають найбільш детальну методологію впровадження. Нейромережі дозволяють будувати доволі точні та потужні моделі, навіть там, де потрібний якісний аналіз взаємозв’язків факторів, що впливають на результат.

РОЗДІЛ 3 ЗАСТОСУВАННЯ СИСТЕМИ КРЕДИТНОГО СКОРИНГУ НА ОСНОВІ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ

3.1 Система кредитного скорингу

Застосування кредитного скорингу для банківських установ є складним економічним та технологічним процесом, що потребує системного підходу. Повноцінна система кредитного скорингу має включати в себе засоби обробки та збереження даних, формуванню вітрин даних, середовище моделювання та широкий набір аналітичних інструментів для побудови та аналізу скорингових моделей, а також, широка система побудови інтерактивних звітів для вирішення задач оцінки роботи скорингових моделей та стану кредитного портфелю. Така система надає можливість банківському аналітику-експерту в сфері інтелектуального аналізу даних та кредитних ризиках створювати та аналізувати скорингові моделі для споживчих кредитів, кредитних карт, іпотечних, автомобільних та інших банківських кредитних продуктів. Автоматизована скорингова система спрямована на вирішення таких задач як оцінювання ймовірності дефолту позичальника, оцінювання ймовірності використання банківських рахунків з метою шахрайства, створення рейтингової системи для регулятора та ін.

Оскільки найважливішим компонентом будь-якого аналізу даних є дані, а від їх якості залежить адекватність та точність скорингових моделей, система повинна мати вбудовану серед для розробки процесів сбору, обробки та завантаження даних, що надходять із бази даних, зовнішніх джерел, даних щодо транзакцій, тощо, а також, наявність відповідної структури збереження даних.

З іншого боку, система має бути інтегрованою з іншими системами, наприклад, системою управління маркетинговими кампаніями у режимі реального часу. Така комбінована система має можливість використовувати результати скорингу клієнта для релевантної маркетингової кампанії щодо кредитного продукту, який відповідає вимогам клієнта та є прийнятним з точки зору рівня ризику його неповернення. Наприклад, якщо клієнт залишив заявку на отримання іпотечного кредиту та отримав в результаті роботи скорингової системи низький скоринговий бал, що свідчить про високий ризик дефолту клієнта та неповернення іпотечного кредиту, комбінована система запропонує клієнтові автомобільний кредит на значно меншу суму (у випадку відсутності автомобіля), при цьому використовуючи вже визначений скоринговий бал. Таким чином, банк не втрачає клієнта, але зменшує ризик неповернення кредиту до мінімального, що істотно покращує операційну діяльність банківської установи.

На рис. 3.1 наведено схему роботи повноцінної скорингової системи, а також її інтеграцію з джерелами даних та іншими банківськими системами.

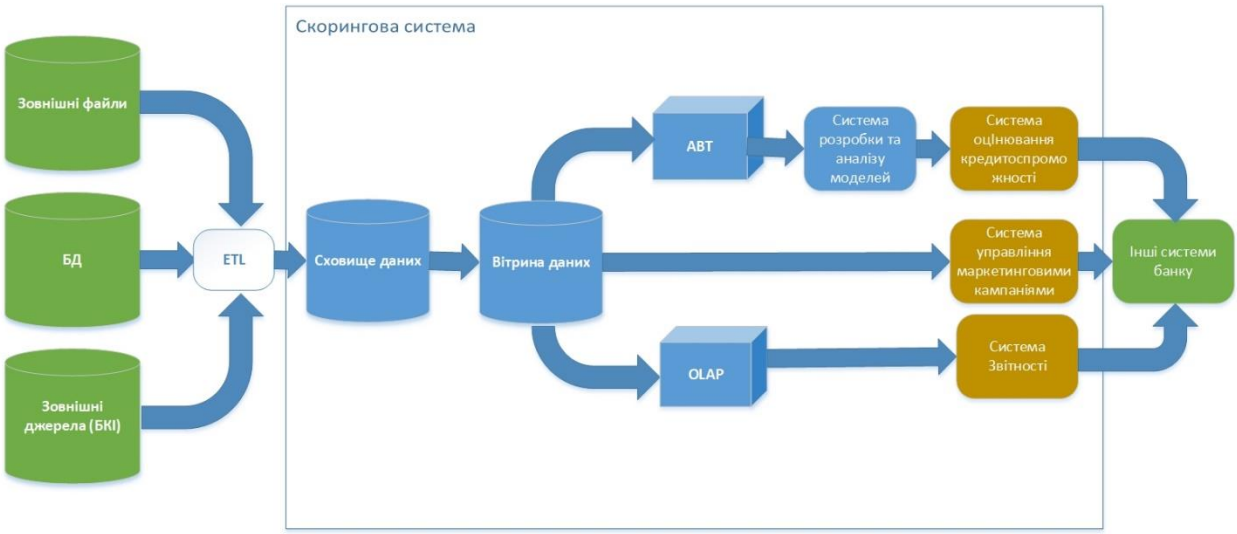

Рисунок 3.1 – Схема повноцінної промислової скорингової системи

Розглянуті особливості існуючих пакетів програмного забезпечення для роботи з математичними моделями, особливості деяких типів моделей, а також, досвід роботи з банківськими даними клієнтів дозволяє нам висунути ряд вимог для скорингової системи, а також, вибору платформи для виконання проміжних обчислень під час побудови та навчання математичних моделей.

Отже, для промислової скорингової системи актуальними є наступні вимоги:

- можливість роботи з надвеликими вибірками (до 20 мільйонів клієнтів);
- можливість роботи з різними джерелами даних (як зовнішніми, що реалізовані у виглядів веб-сервісів та зовнішніх файлів, так і внутрішніми, що характеризуються різними СКБД);
- наявність системи надійного захисту персональних даних клієнтів;
- наявність підсистем збору, завантаження, обробки та аналізу даних;
- наявність підсистеми автоматичної звітності;
- інтегрованість з іншими банківськими системами;
- робота у високонавантаженому середовищі;
- наявність зрозумілого для аналітика ризиків та аналітика даних інтерфейсу.

Ключовим показником кредитоспроможності клієнта є скоринговий бал, що визначається в загальному випадку за допомогою скорингової моделі, побудованої на основі історичних даних, що включають соціо-демографічні дані, дані про продукти, кредитну історію, транзакційні дані, та ін. та прогнозується на основі поточного набору вхідних даних, що представлені тією самою структурою, що і історичні дані. Саме тому найважливішим

компонентом ефективної скорингової системи є наявність точної та репрезентативної структури даних про клієнтів.

3.2 Аналіз кредитоспроможності позичальників кредитів за допомогою байєсівських та нейромереж

В даній дисертації розглядається можливість застосування сучасних методів Інтелектуального аналізу даних, таких як байєсовські мережі та нейронні мережі для вирішення задачі прогнозування кредитоспроможності позичальників кредитів та їх порівняння з класичними методами кредитного скорингу, такими як дерева рішень та регресійний метод. В якості середовища проведення побудови й аналізу моделей було вибрано аналітичний пакет ПЗ SAS Enterprise Miner, що дозволяє проводити аналіз даних за методологією SEMMA (Sample, Explore, Modify, Model, Assess) та порівняння моделей безпосередньо у середовищі розробки.

Виходячи з наведених у першому та другому розділах цієї дисертації теоретичних міркувань, для оцінювання математичних моделей визначимо наступні промислові критерії:

- можливість навчання моделі під час роботи з надвеликими вибірками (до 20 мільйонів клієнтів);
- легка інтерпретація отриманих результатів та взаємозалежностей;
- точність моделі на тестовій вибірці.
- Побудова моделей логістичної регресії, автоматичний вибір значень параметрів моделей;
- Побудова моделей нейронних мереж, автоматичний вибір структури моделей, автоматичний вибір параметрів;

- Побудова моделей байєсівських мереж, автоматичний вибір структури виходячи з заданого користувачем типу мережі;
- Побудова кореляційної матриці вхідного набору даних;
- Автоматичне оцінювання адекватності критерію та вибір найкращої моделі;

Для побудови моделей використовувалась скорочена вибірка анонімізованих соціо-демографічних та агрегованих даних рахунків клієнтів одного з банків, представлених у числовому вигляді, що складаються з вісімнадцяти вхідних змінних та однієї цільової та налічують близько двох тисяч полів (табл. 3.1.1).

Таблиця 3.1.1 – Вибірка банківських демографічних та агрегованих даних щодо рахунків позичальника.

Атрибут	**Тип**	**Дискретність**	**Опис**
Age	Вхідний	Неперервний	Вік позичальника (роки)
Amount	Вхідний	Неперервний	Тіло кредиту
Checking	Вхідний	Дискретний	Залишок на поточному рахунку (грн): 1: ... < 0 2: 0 <= … < 10000 3: 10000 <= …
Depends	Вхідний	Неперервний	Кількість залежних осіб
Durations	Вхідний	Неперервний	Термін кредиту
Employed	Вхідний	Дискретний	Стаж на поточному місці роботи: 1: Непрацевлаштований 2: ... < 1 рік 3: 1 <= … < 4 роки

			4: 4 <= … < 7 років 5: ... >= 7 років

Продовження таблиці 3.1.1

Existcr	Вхідний	Неперервний	Кількість поточних непогашених кредитів
Foreign	Вхідний	Дискретний	Працівник іноземної компанії: 1: Так 2: Ні
Good_bad	Цільовий	Дискретний	Кредитний рейтинг: 0: Ненадійний клієнт; 1: Надійний клієнт;
History	Вхідний	Дискретний	Кредитна історія: 0: Клієнт не користувався кредитами / всі кредити погашені; 1: Всі кредиті цього банку, погашені; 2: Поточні кредити станом на сьогодні виплачуються вчасно; 3: В історії клієнта є періоди погашення кредитів із затримкою; 4: Критичний клієнт / наявні непогашені кредити в інших банках;
Housing	Вхідний	Дискретний	Наявність житла: 1: Оренда; 2: Власне;

			3: Безоплатне проживання;

Продовження таблиці 3.1.1

Installp	Вхідний	Неперервний	Ставка (Відсоток від наявного доходу)
Job	Вхідний	Дискретний	Тип зайнятості: 1: Непрацевлаштований / некваліфікований нерезидент; 2: Резидент-неспеціаліст; 3: Працівник-спеціаліст / державний службовець; 4: Менеджер / самовлаштований / висококваліфікований спеціаліст-працівник / урядовець
Marital	Вхідний	Дискретний	Сімейний стан / стать: 1: Чоловічий: розлучений; 2: Жіночий: розлучений; 3: Чоловічий: нежонатий; 4: Чоловічий: жонатий / вдовець; 5: Жіночий: нежонатий;
Other	Вхідний	Дискретний	Інші види розстрочки: 1: Банківська; 2: Розстрочка за товари; 3: Відсутні;

Продовження таблиці 3.1.1

Property	Вхідний	Дискретний	Власність: 1: Нерухомість; 2: У випадку відсутності нерухомості: наявність полісу страхування життя / договір купівлі житла; 3: У випадку відсутності нерухомості, полісу страхування життя та договору купівлі житла: автомобілі, тощо; 4: Дані відсутні / немає власності;
Purpose	Вхідний	Дискретний	Ціль кредитування: 0, 1: Автомобіль; 2: Матеріали / обладнання; 3: Споживче кредитування; 4: Побутова техніка; 5: Ремонт; 6: Освіта; 7: Відпустка; 8: Перекваліфікація; 9: Бізнес; 10: Інше;
Resident	Вхідний	Неперервний	Рік отримання статусу резидента

Продовження таблиці 3.1.1

Savings	Вхідний	Дискретний	Залишок на депозитних та ощадних рахунках (грн): 1: ... < 10000 2: 10000 <= … < 50000 3: 50000 <= … < 100000 4: 100000 <= …

Отже, вибірка даних містить як демографічні дані, отримані як із внутрішніх джерел (База даних), так і з зовнішніх (Бюро кредитних історій, дані анкети позичальника, тощо). Крім того, більшість інтервальних даних представлені у вигляді дискретних груп значень, що робить модель більш точною та дозволяє легко інтерпретувати результати і взаємозв'язки. Приклад даних вибірки наведено на рис. 3.1.2.

Obs	checking	duration	history	purpose	amount	savings	employed	installp	marital	resident	property	age	other	housing	existcr	job	depends	foreign	good_bad
1	1	6	4	3	1169	5	5	4	3	4	1	67	3	2	2	3	1	1	good
2	2	48	2	3	5951	1	3	2	2	2	1	22	3	2	1	3	1	1	bad
3	4	12	4	6	2096	1	4	2	3	3	1	49	3	2	1	2	2	1	good
4	1	42	2	2	7882	1	4	2	3	4	2	45	3	3	1	3	2	1	good
5	1	24	3	0	4870	1	3	3	3	4	4	53	3	3	2	3	2	1	bad
6	4	36	2	6	9055	5	3	2	3	4	4	35	3	3	1	2	2	1	good
7	4	24	2	2	2835	3	5	3	3	4	2	53	3	2	1	3	1	1	good
8	2	36	2	1	6948	1	3	2	3	2	3	35	3	1	1	4	1	1	good
9	4	12	2	3	3059	4	4	2	1	4	1	61	3	2	1	2	1	1	good
10	2	30	4	0	5234	1	1	4	4	2	3	28	3	2	2	4	1	1	bad
11	2	12	2	0	1295	1	2	3	2	1	3	25	3	1	1	3	1	1	bad
12	1	48	2	9	4308	1	2	3	2	4	2	24	3	1	1	3	1	1	bad
13	2	12	2	3	1567	1	3	1	2	1	3	22	3	2	1	3	1	1	good
14	1	24	4	0	1199	1	5	4	3	4	3	60	3	2	2	2	1	1	bad
15	1	15	2	0	1403	1	3	2	2	4	3	28	3	1	1	3	1	1	good
16	1	24	2	3	1282	2	3	4	2	2	3	32	3	2	1	2	1	1	bad
17	4	24	4	3	2424	5	5	4	3	4	2	53	3	2	2	3	1	1	good
18	1	30	0	9	8072	5	2	2	3	3	3	25	1	2	3	3	1	1	good
19	2	24	2	1	12579	1	5	4	2	2	4	44	3	3	1	4	1	1	bad
20	4	24	2	3	3430	3	5	3	3	2	3	31	3	2	1	3	2	1	good
21	4	9	4	0	2134	1	3	4	3	4	3	48	3	2	3	3	1	1	good
22	1	6	2	3	2647	3	3	2	3	3	1	44	3	1	1	3	2	1	good
23	1	10	4	0	2241	1	2	1	3	3	1	48	3	1	2	2	2	2	good
24	2	12	4	1	1804	2	2	3	3	4	2	44	3	2	1	3	1	1	good
25	4	10	4	2	2069	5	3	2	4	1	3	26	3	2	2	3	1	2	good
26	1	6	2	2	1374	1	3	1	3	2	1	36	1	2	1	2	1	1	good
27	4	6	0	3	426	1	5	4	4	4	3	39	3	2	1	2	1	1	good
28	3	12	1	3	409	4	3	3	2	3	1	42	3	1	2	3	1	1	good
29	2	7	2	3	2415	1	3	3	3	2	1	34	3	2	1	3	1	1	good
30	1	60	3	9	6836	1	5	3	3	4	4	63	3	2	2	3	1	1	bad
31	2	18	2	9	1913	4	2	3	4	3	1	36	1	2	1	3	1	1	good
32	1	24	2	2	4020	1	3	2	3	2	3	27	2	2	1	3	1	1	good
33	2	18	2	0	5866	2	3	2	3	2	3	30	3	2	2	3	1	1	good
34	4	12	4	9	1264	5	5	4	3	4	4	57	3	1	1	2	1	1	good
35	3	12	2	2	1474	1	2	4	2	1	2	33	1	2	1	4	1	1	good
36	2	45	4	3	4746	1	2	4	3	2	2	25	3	2	2	2	1	1	bad
37	4	48	4	6	6110	1	3	1	3	3	4	31	1	3	1	3	1	1	good

Рисунок 3.1.2 – Вибірка даних

З візуального представлення даних можна зробити висновок, що дані мають одну цільову змінну, а всі інші змінні є числовими. Оскільки деякі моделі не працюють з даними, що можуть містити пропущені значення, для точності порівняння моделей, перевіримо вибірку на наявність пропущених значень, а також, виведемо на екран всю наявну інформацію, що допоможе краще охарактеризувати дані, за допомогою складової системи SAS – SAS Enterprise Guide (рис. 3.1.3).

good_bad	N Obs	Variable	Mean	Std Dev	Std Error	Variance	Minimum	Maximum	N	N Miss	t Value	Pr > |t|
bad	300	checking	1.9033333	1.0508742	0.0606723	1.1043367	1.0000000	4.0000000	300	0	31.37	<.0001
		duration	24.8600000	13.2826389	0.7668735	176.4284950	6.0000000	72.0000000	300	0	32.42	<.0001
		history	2.1666667	1.0783157	0.0622566	1.1627648	0	4.0000000	300	0	34.80	<.0001
		amount	3938.13	3535.82	204.1406026	12502015.68	433.0000000	18424.00	300	0	19.29	<.0001
		savings	1.6733333	1.3034386	0.0752541	1.6989521	1.0000000	5.0000000	300	0	22.24	<.0001
		employed	3.1700000	1.2245127	0.0706973	1.4994314	1.0000000	5.0000000	300	0	44.84	<.0001
		installp	3.0966667	1.0883953	0.0628385	1.1846042	1.0000000	4.0000000	300	0	49.28	<.0001
		marital	2.5866667	0.7377691	0.0425951	0.5443032	1.0000000	4.0000000	300	0	60.73	<.0001
		resident	2.8500000	1.0946052	0.0631971	1.1981605	1.0000000	4.0000000	300	0	45.10	<.0001
		property	2.5866667	1.0453699	0.0603545	1.0927982	1.0000000	4.0000000	300	0	42.86	<.0001
		age	33.9633333	11.2223792	0.6479244	125.9417949	19.0000000	74.0000000	300	0	52.42	<.0001
		other	2.5566667	0.7930228	0.0457852	0.6288852	1.0000000	3.0000000	300	0	55.84	<.0001
		housing	1.9133333	0.6113384	0.0352956	0.3737347	1.0000000	3.0000000	300	0	54.21	<.0001
		existcr	1.3666667	0.5597021	0.0323144	0.3132664	1.0000000	4.0000000	300	0	42.29	<.0001
		job	2.9366667	0.6689398	0.0386213	0.4474805	1.0000000	4.0000000	300	0	76.04	<.0001
		depends	1.1533333	0.3609105	0.0208372	0.1302564	1.0000000	2.0000000	300	0	55.35	<.0001
		foreign	1.0133333	0.1148893	0.0066331	0.0131996	1.0000000	2.0000000	300	0	152.77	<.0001
good	700	checking	2.8657143	1.2287549	0.0464426	1.5098385	1.0000000	4.0000000	700	0	61.70	<.0001
		duration	19.2071429	11.0795643	0.4187682	122.7567443	4.0000000	60.0000000	700	0	45.87	<.0001
		history	2.7071429	1.0447527	0.0394879	1.0915083	0	4.0000000	700	0	68.56	<.0001
		amount	2985.46	2401.47	90.7671204	5767069.10	250.0000000	15857.00	700	0	32.89	<.0001
		savings	2.2900000	1.6513444	0.0624150	2.7269385	1.0000000	5.0000000	700	0	36.69	<.0001
		employed	3.4757143	1.1904407	0.0449944	1.4171490	1.0000000	5.0000000	700	0	77.25	<.0001
		installp	2.9200000	1.1280784	0.0426374	1.2725608	1.0000000	4.0000000	700	0	68.48	<.0001
		marital	2.7228571	0.6914916	0.0261359	0.4781606	1.0000000	4.0000000	700	0	104.18	<.0001
		resident	2.8428571	1.1083725	0.0418925	1.2284897	1.0000000	4.0000000	700	0	67.86	<.0001
		property	2.2600000	1.0376875	0.0392209	1.0767954	1.0000000	4.0000000	700	0	57.62	<.0001
		age	36.2242857	11.3811447	0.4301668	129.5304537	19.0000000	75.0000000	700	0	84.21	<.0001
		other	2.7257143	0.6587554	0.0248986	0.4339587	1.0000000	3.0000000	700	0	109.47	<.0001
		housing	1.9357143	0.4933131	0.0186455	0.2433579	1.0000000	3.0000000	700	0	103.82	<.0001
		existcr	1.4242857	0.5847210	0.0221004	0.3418986	1.0000000	4.0000000	700	0	64.45	<.0001
		job	2.8900000	0.6469141	0.0244511	0.4184979	1.0000000	4.0000000	700	0	118.20	<.0001
		depends	1.1557143	0.3628435	0.0137142	0.1316554	1.0000000	2.0000000	700	0	84.27	<.0001
		foreign	1.0471429	0.2120959	0.0080165	0.0449847	1.0000000	2.0000000	700	0	130.62	<.0001

Рисунок 3.1.3 – Описовий аналіз вибірки даних

З наведеної вище таблиці видно, що пропущені значення відсутні, але, при цьому, одна зі змінних не була включена до аналізу. Це пов’язано з тим, що пропущені значення цієї змінної були інтерпретовані як символ «X». Тому наступним кроком при моделюванні є виключення пропущених значень з аналізу. Для цього використаємо інший компонент системи SAS, а саме,

середовище програмування SAS Base. Тепер дані інтерпретовані правильно та підготовлені до побудови моделей.

Для подальшого оцінювання якості моделі та уникнення перенавчення моделі необхідно розбити вхідну вибірку даних на навчальну, перевірочну та тестову. Виконаємо цю операцію також в середовищі SAS Enterprise Miner, використовуючи метод простого випадкового розбиття із співвідношенням 50% / 30% / 20% відповідно.

Для визначення остаточних моделей прогнозування кредитоспроможності клієнтів було побудовано (рис. 3.1.4) такі моделі як:

– нейронні мережі;
– байєсівські мережі;
– логістична регресія;
– дерева рішень;

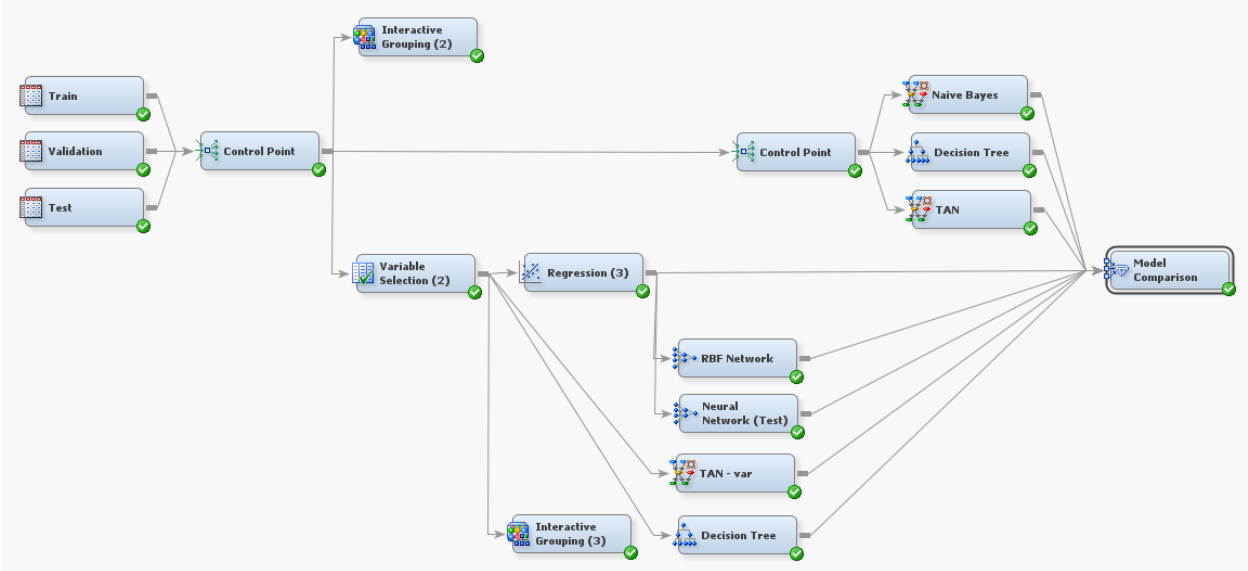

Рисунок 3.1.4 – Побудова скорингових моделей

3.2.1 Побудова скорингових моделей на основі байєсівських мереж

Одними із найбільш легких для сприйняття та інтерпретації моделей є моделі байєсівських мереж. В рамках даної магістерської дисертації будемо розглядати лише статичні байєсівські мережі, адже задача аналізу кредитоспроможності не містить динамічних складових. На сьогодні, найбільш розповсюдженими та вживаними є наступні моделі:

- наївні байєсівські мережі;
- доповнені деревом байєсівські мережі (TAN – Tree Augmented Networks);

Саме ці моделі будуть детально розглянуті для моделювання ризику дефолту позичальників кредитів.

Найпростішою моделлю байєвських мереж є наївний байєсівський класифікатор. В такій моделі умовні ймовірності всіх змінних, включаючи цільові, обчислюються під час навчання моделі. На етапі тестування моделі розраховуються апостеріорні ймовірності для кожного стану цільової змінної. Рішення приймається за максимальним значенням ймовірності. Розглянемо результати навчання побудованої моделі (рис. 3.2.1)

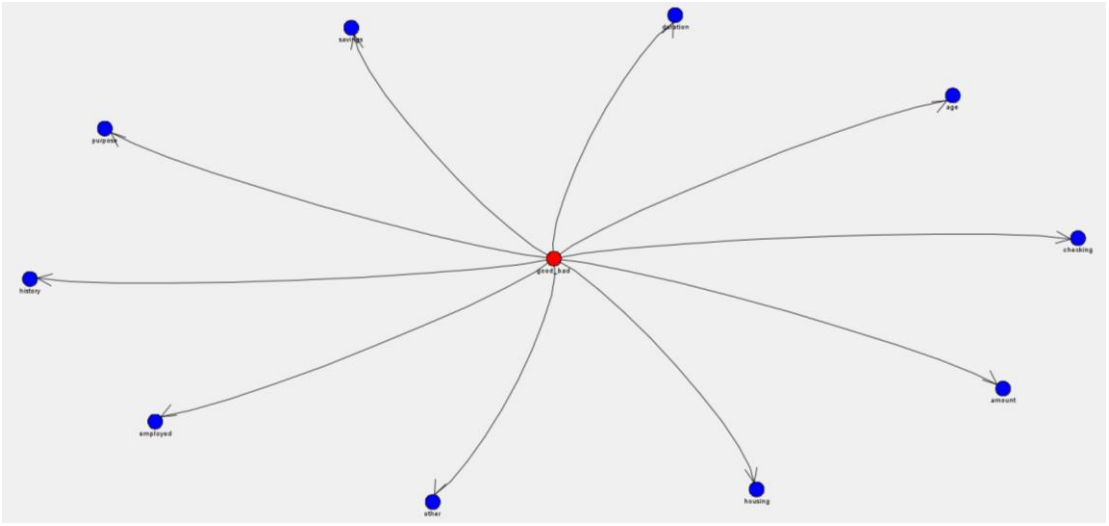

Рисунок 3.2.1 – Наївний байєсівський класифікатор

Для оцінки схильності моделі до помилок першого чи другого роду, розглянемо результати оцінювання ймовірності дефолту клієнта детальніше (рис. 3.2.2), а також у числовому форматі (рис. 3.2.3).

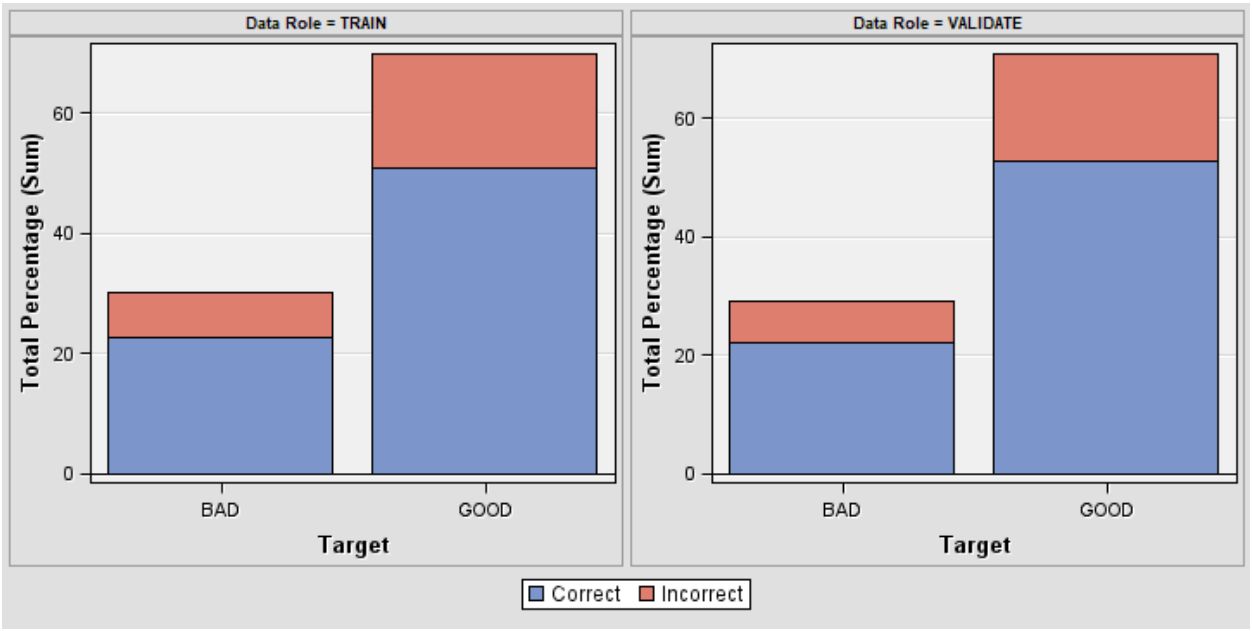

Рисунок 3.2.2 – Візуалізована діаграма помилок 1-го та 2-го роду для моделі наївного байєсівського класифікатора.

False Negative	True Negative	False Positive	True Positive
193	226	74	507

Data Role=VALIDATE Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
109	133	41	317

Рисунок 3.2.3 – Аналіз класифікації позичальників щодо ймовірності дефолту.

Як було зазначено у розділі 2 цієї дисертації, основним припущенням наївного байєсівського класифікатора є те, що всі вхідні змінні незалежні між собою, але це рідко відповідає дійсності. В нашому випадку деякі інтервальні змінні є залежними (наприклад amount та duration, коефіцієнт кореляції між якими складає 0,63). Цим пояснюється порівняно низька точність моделі – значення коефіцієнту міскласифікації складає 0.25.

Розлянемо більш прогресивну модель байєсівських мереж – доповнену деревом байєсівську мережу, що не потребує обов’язкової незалежності вхідних змінних (рис. 3.2.4).

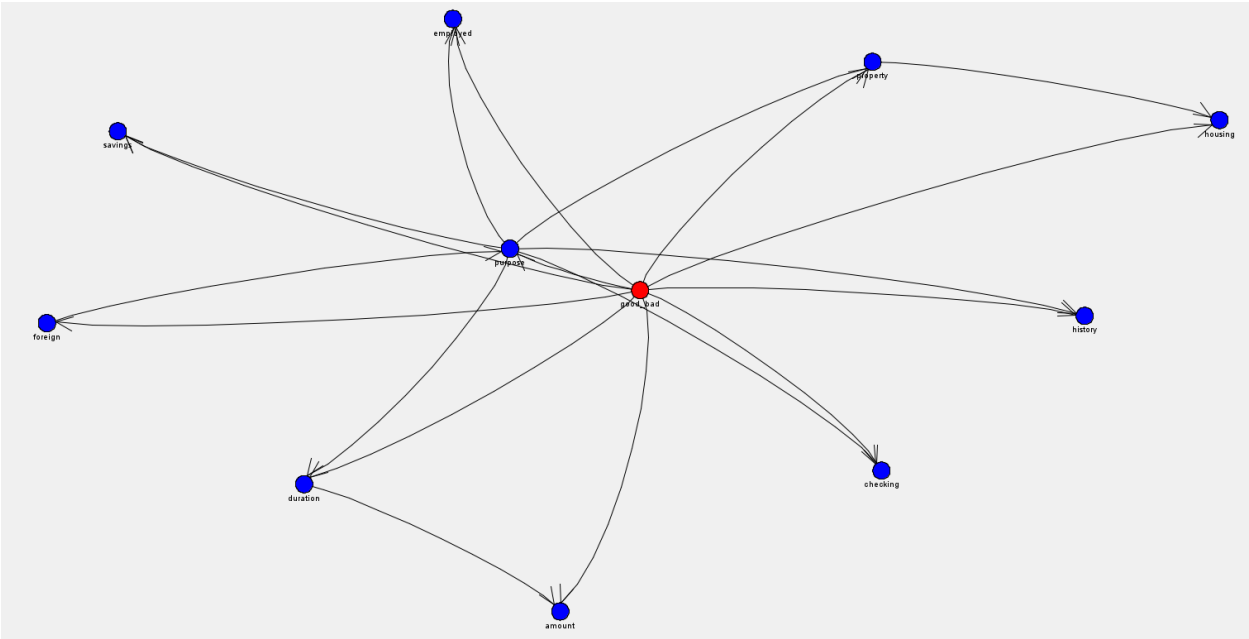

Рисунок 3.2.4 – Доповнений деревом байєсівський класифікатор (TAN)

Побудована модель надає можливість аналізувати залежність не лише цільової та вхідних змінних, а й взаємозв’язки між вхідними змінними. З наведеної на рис. 3.2.4 схеми побудованої мережі можна легко визначити, що наприклад зміні amount та duration є залежними. Оскільки вхідні змінні вибірки, що досліджується, є залежними, можна припустити, що точність моделі TAN буде вищою, ніж у моделі наївного байєсівського класифікатора. Про це свідчить і коефіцієнт міскласифікації, значення якого для побудованої моделі TAN становить 0.22 на валідаційній вибірці.

Для оцінки схильності моделі до помилок першого чи другого роду, розглянемо результати оцінювання ймовірності дефолту клієнта детальніше (рис. 3.2.5), а також у числовому форматі (рис. 3.2.6).

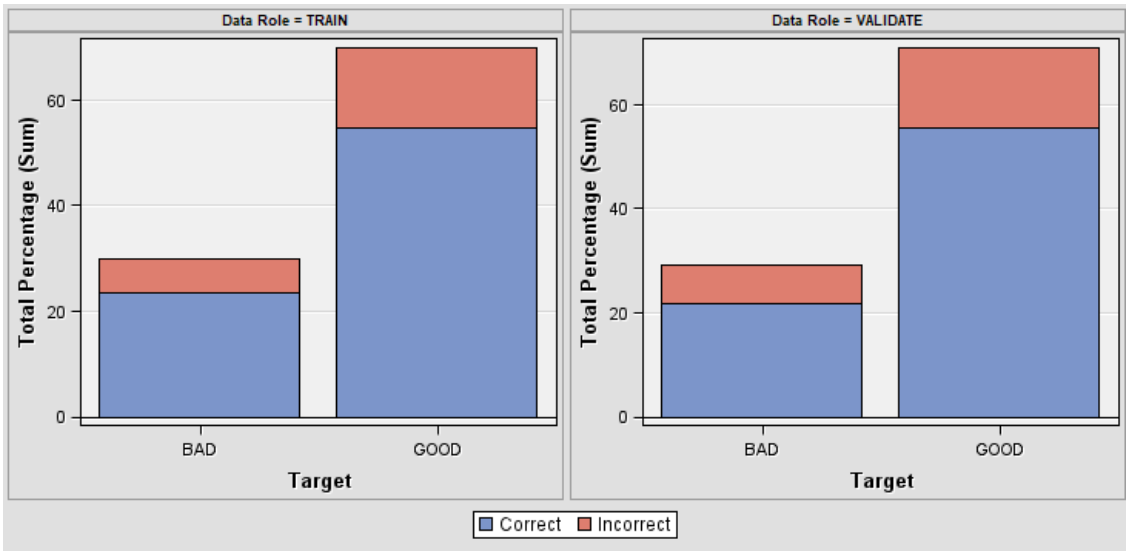

Рисунок 3.2.5 – Візуалізована діаграма помилок 1-го та 2-го роду для моделі TAN

З наведеної вище діаграми видно, що модель TAN має значно менше помилок False Positive, при цьому здатність моделі розпізнати ризикових клієнтів залишається майже незмінною (7,16%), порівняно з наївним байєсовським класифікатором (6,83%).

Data Role=TRAIN Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
154	235	65	546

Data Role=VALIDATE Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
93	131	43	333

Рисунок 3.2.6 – Аналіз класифікації позичальників щодо ймовірності дефолту.

Отже, на валідаційній вибірці різниця в точності між TAN та наївним байєсівським класифікатором не є суттєвою.

При побудові обох наведених вище моделей байєсівських мереж використовалися оригінальні змінні вибірки, що під час навчання моделей були автоматично проаналізовані системою на предмет наявності незалежності між вхідними змінними за наступним алгоритмом, що визначається за критеріями Chi-square:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

та G-square:

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

Де:

$O_{ij} -$ фактична частота $ij -$ го значення таблиці контингентності;

$E_{ij} -$ очікувана частота $ij -$ го значення таблиці контингентності;

Таким чином, дві змінні вважаються незалежними, якщо p-оцінка тестової статистики перевищує задану константу, що в даній роботі набуває значення 0.2. У випадку залежності змінних, змінні виключаються з аналізу. Оскільки в даній роботі розглядається мала вибірка банківських даних лише з 17-ма вхідними параметрами, виключення змінних не призводить до покращення адекватності моделей.

Тому, виходячи з того, що більшість інтервальних даних вибірки вже розділені на агреговані групи, застосуємо підхід “Use Interactions” для створення нових змінних. Такі змінні формуються у результаті комбінації кожної пари існуючих категоріальних змінних, наприклад, «Age/Checking» (рис. 3.2.7).

Variable Name	Role	Measurement Level	Type	Label ▲
GI_checking_coapp	Input	Nominal	Numeric	Grouped Interactions for checking and coapp
GI_checking_employed	Input	Nominal	Numeric	Grouped Interactions for checking and employed
GI_checking_history	Input	Nominal	Numeric	Grouped Interactions for checking and history
GI_checking_housing	Input	Nominal	Numeric	Grouped Interactions for checking and housing
GI_checking_job	Input	Nominal	Numeric	Grouped Interactions for checking and job
GI_checking_marital	Input	Nominal	Numeric	Grouped Interactions for checking and marital
GI_checking_other	Input	Nominal	Numeric	Grouped Interactions for checking and other
GI_checking_property	Input	Nominal	Numeric	Grouped Interactions for checking and property
GI_checking_purpose	Input	Nominal	Numeric	Grouped Interactions for checking and purpose
GI_checking_savings	Input	Nominal	Numeric	Grouped Interactions for checking and savings
GI_employed_foreign	Input	Nominal	Numeric	Grouped Interactions for employed and foreign
GI_employed_history	Input	Nominal	Numeric	Grouped Interactions for employed and history
GI_employed_housing	Input	Nominal	Numeric	Grouped Interactions for employed and housing
GI_employed_job	Input	Nominal	Numeric	Grouped Interactions for employed and job
GI_employed_marital	Input	Nominal	Numeric	Grouped Interactions for employed and marital
GI_employed_other	Input	Nominal	Numeric	Grouped Interactions for employed and other
GI_employed_property	Input	Nominal	Numeric	Grouped Interactions for employed and property
GI_employed_purpose	Input	Nominal	Numeric	Grouped Interactions for employed and purpose
GI_employed_savings	Input	Nominal	Numeric	Grouped Interactions for employed and savings
GI_foreign_history	Input	Nominal	Numeric	Grouped Interactions for foreign and history
GI_foreign_marital	Input	Nominal	Numeric	Grouped Interactions for foreign and marital
GI_foreign_property	Input	Nominal	Numeric	Grouped Interactions for foreign and property
GI_foreign_purpose	Input	Nominal	Numeric	Grouped Interactions for foreign and purpose
GI_foreign_savings	Input	Nominal	Numeric	Grouped Interactions for foreign and savings
GI_history_housing	Input	Nominal	Numeric	Grouped Interactions for history and housing
GI_history_job	Input	Nominal	Numeric	Grouped Interactions for history and job
GI_history_marital	Input	Nominal	Numeric	Grouped Interactions for history and marital
GI_history_other	Input	Nominal	Numeric	Grouped Interactions for history and other
GI_history_property	Input	Nominal	Numeric	Grouped Interactions for history and property
GI_history_purpose	Input	Nominal	Numeric	Grouped Interactions for history and purpose
GI_history_savings	Input	Nominal	Numeric	Grouped Interactions for history and savings
GI_housing_marital	Input	Nominal	Numeric	Grouped Interactions for housing and marital
GI_housing_other	Input	Nominal	Numeric	Grouped Interactions for housing and other

Рисунок 3.2.7 – Результат групування змінних.

Побудуємо, для порівняння, модель TAN на вдосконаленій вибірці (рис. 3.2.8) та проаналізуємо результати прогнозування. Оскільки модель TAN не потребує виконання умов незалежності вхідних змінних, під час побудови не будемо виключати змінні з моделі за критеріями Chi-square та G-square.

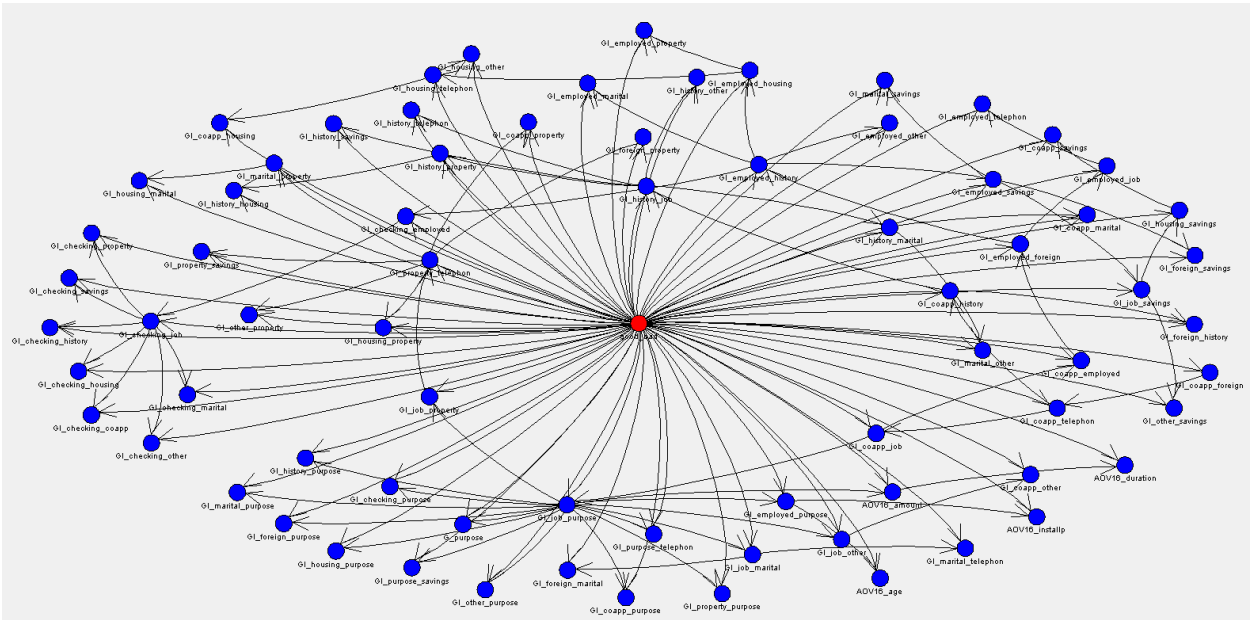

Рисунок 3.2.8 – Модель TAN на розширеній вибірці.

В результаті групування змінних, отримаємо коефіцієнт міскласифікації, що дорівнює 0,128 на валідаційній вибірці, що означає покращення якості моделі більш ніж у два рази, порівняно зі звичайним способом побудови моделі. Аналізуючи результати класифікації на валідаційній вибірці, можна побачити, що модель TAN досить точно розпізнає надійних клієнтів, але схильна вважати ризикового клієнта надійним (рис. 3.2.9).

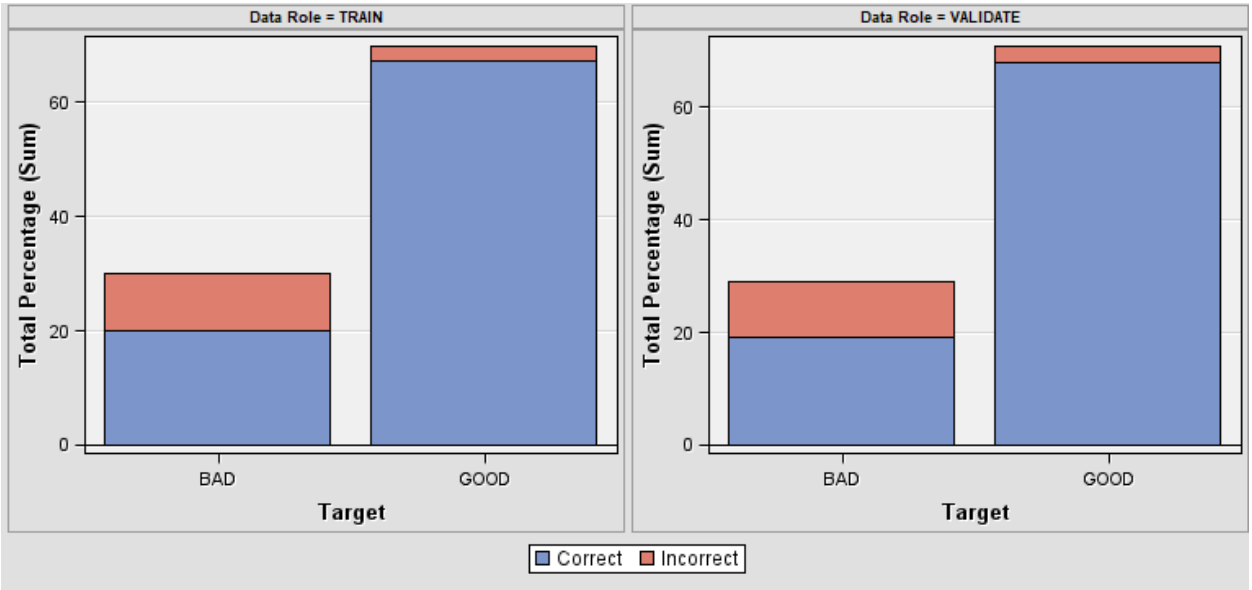

Рисунок 3.2.9 – Візуалізована діаграма помилок 1-го та 2-го роду для моделі TAN на розширеній вибірці.

Data Role=TRAIN Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
26	200	100	674

Data Role=VALIDATE Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
18	115	59	408

Рисунок 3.2.10 – Аналіз класифікації позичальників (TAN)

Отже, байєсівські моделі можуть бути застосовані для вирішення задач кредитного скорингу, але є істотно залежними від кількості змінних у вхідній вибірці та від їх незалежності (наївний байєсівський класифікатор). Перевагою байєсівських мереж (TAN) є висока інтерпретабельність, адже дивлячись на схему 3.2.4 можна легко визначити, які змінні є залежними, а також, визначити типи таких зв'язків.

3.2.2 Побудова скорингових моделей на основі нейромереж

Одним із найбільш сучасних типів математичних моделей та алгоритмів машинного навчання є нейронні мережі. Оскільки одним із найбільш важливих критеріїв для моделей є легкість інтерпретації аналітиком, поглиблені алгоритми нейронних мереж, такі як конволюційні нейронні мережі, не розглядаються в рамках цієї дисертації. Тому для вирішення задачі оцінювання кредитоспроможності позичальників будемо розглядати наступні типи нейронних мереж, що не потребують істотних потужностей для навчання та є легкоінтерпретовними:

- багатошаровий персептрон;
- RBF-мережа;

Розглянемо багатошаровий персептрон, що навчається за алгоритмом зворотного поширення похибки. В якості активаційної функції вихідного шару будемо використовувати функцію Softmax. Аналізуючи точність роботи моделі з різними параметрами встановлено, що найбільш точним є прогноз моделі багатошарового персептрона зі значенням коефіцієнта навчання, що становить 0.756. Коефіцієнт міскласифікації для такої моделі на валідаційній вибірці становить 0.1033, що є свідченням високої точності

моделі. На рис. 3.2.11 показано графік середньоквадратичної похибки для отриманої моделі на тренувальній та валідаційній вибірках:

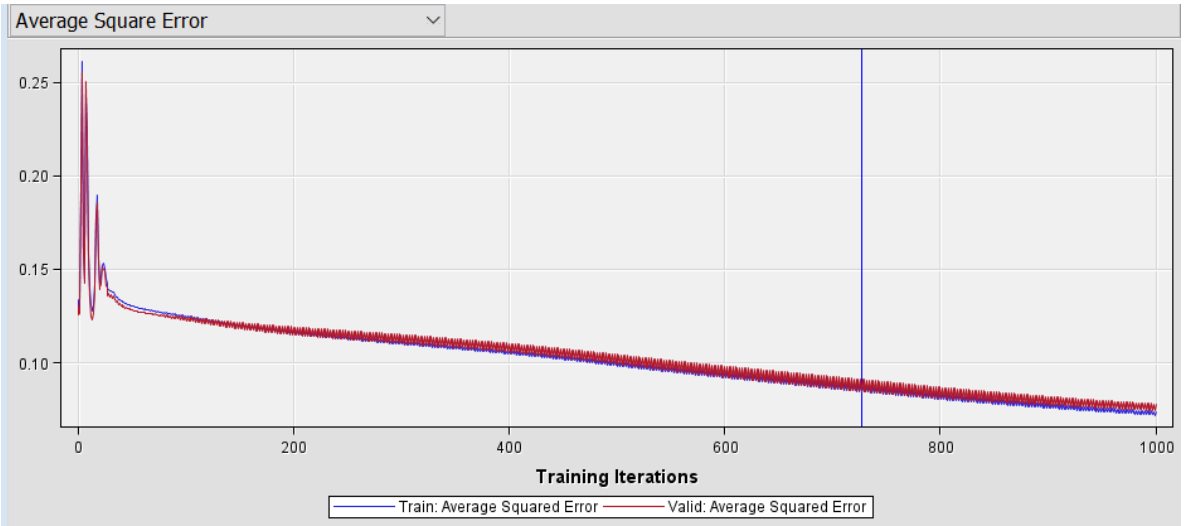

Рисунок 3.2.11 – Графік середньоквадратичної похибки на кожній ітерації.

Аналізуючи результати навчання моделей, розглянемо якість прогнозування дефолту позичальника детальніше. Оскільки для українського банку в умовах економічного спаду, що спостерігається у 2014-2018рр., важливіше зменшити ризик дефолту, у випадку некоректної роботи моделі більш прийнятним результатом прийняття рішень є невидача клієнтові кредиту, якщо клієнт є кредитоспроможним, аніж видача кредиту некредитоспроможній особі. Розглянемо результати оцінювання ймовірності дефолту клієнта. Для цього скористаємося діаграмою, що показує наявність помилок першого та другого роду у більш візуалізованій формі (рис. 3.2.12), а також у числовому форматі (рис. 3.2.13).

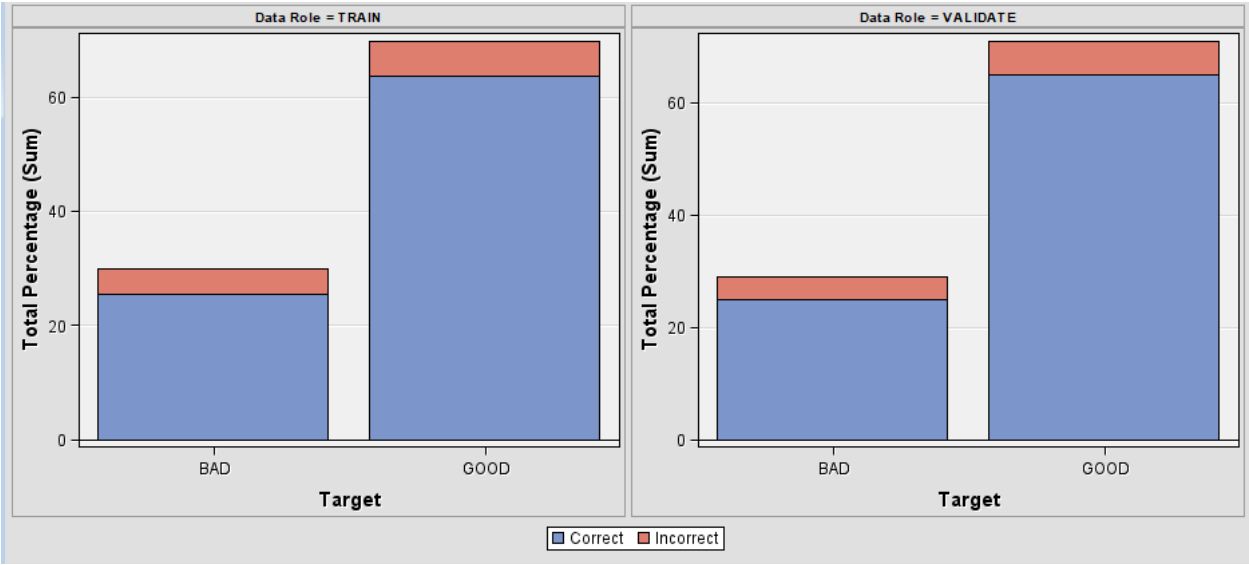

Рисунок 3.2.12 – Візуалізована діаграма помилок 1-го та 2-го роду для моделі багатошарового персептрону зі зворотним поширенням похибки.

Data Role=TRAIN Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
63	256	44	637

Data Role=VALIDATE Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
37	149	25	389

Рисунок 3.2.13 – Аналіз класифікації позичальників щодо ймовірності дефолту.

Наведені вище діаграми демонструють, що приблизно 6% надійних позичальників були класифіковані як ненадійні (FN), але лише 4% ризикових клієнтів були класифіковані як надійні (FP). Отже, модель нейронної мережі «багатошаровий персептрон» зі зворотним поширенням похибки є адекватною, але більш «обережною», про що свідчить FP < FN.

Розглянемо також інший тип нейронної мережі – мережа на основі радіально-базисних функцій. Це нейронна мережа прямого поширення сигналу, яка містить проміжний (прихований) шар радіально симетричних нейронів. Такий нейрон перетворює відстань від даного вхідного вектора до відповідного йому "центру" по деякому нелінійному закону (зазвичай функція Гаусса). В якості активаційної функції вихідного шару будемо використовувати функцію Softmax. Розглянемо результати навчання та валідації РБФ-мережі. Коефіцієнт міскласифікації такої мережі на валідаційній вибірці становить 0.082, що є більш точним результатом, ніж модель нейронної мережі зі зворотним поширенням похибки.

На рис. 3.2.14 показано графік середньоквадратичної похибки для отриманої моделі на тренувальній та валідаційній вибірках:

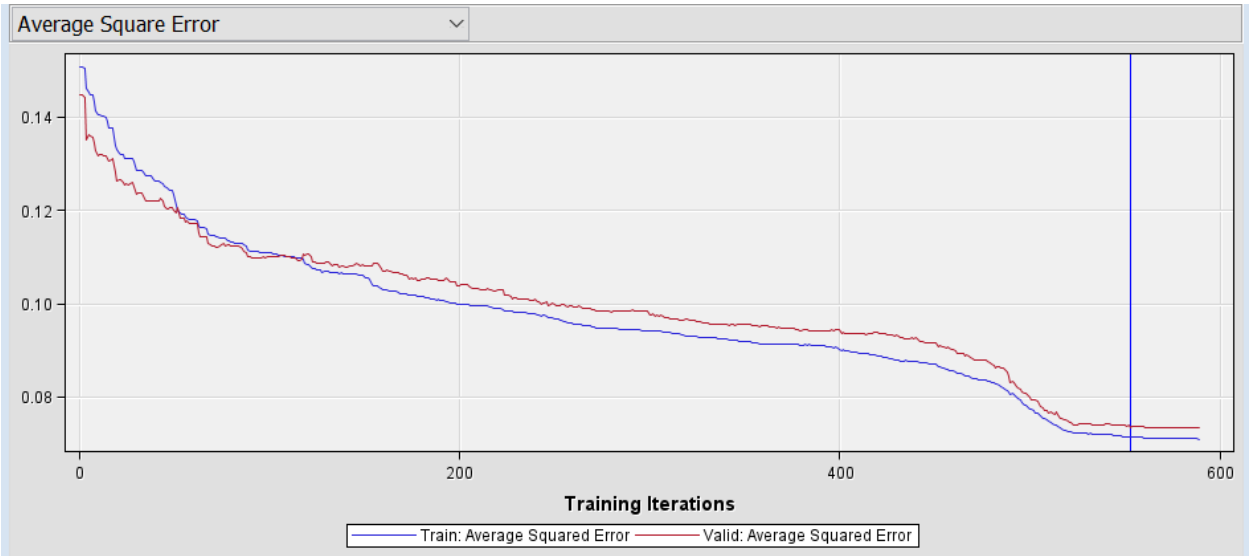

Рисунок 3.2.14 – Графік середньоквадратичної похибки на кожній ітерації.

Проаналізуємо, як і у випадку з багатошаровим персептроном, поведінку моделі у випадку некоректної класифікації позичальників. Для цього скористаємося критеріями кількості помилок першого та другого роду False Negative та False Positive (рис. 3.2.15).

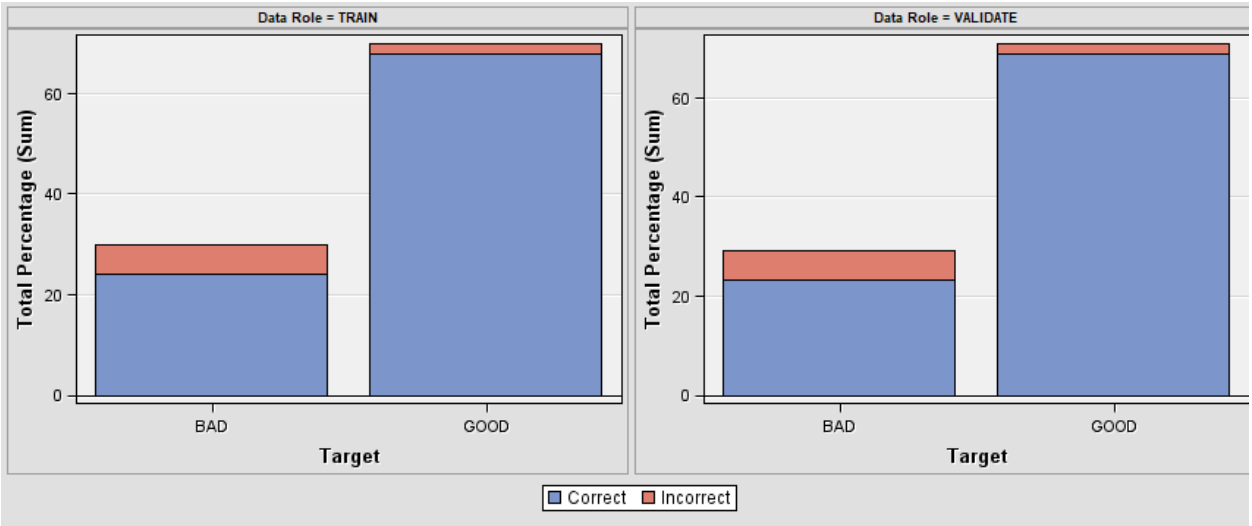

Рисунок 3.2.15 – Візуалізована діаграма помилок 1-го та 2-го роду для моделі РБФ-мережі.

А також, у вигляді числових даних (рис. 3.2.16)

Data Role=TRAIN Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
22	242	58	678

Data Role=VALIDATE Target=good_bad Target Label=' '

False Negative	True Negative	False Positive	True Positive
14	139	35	412

Рисунок 3.2.16 – Аналіз класифікації позичальників щодо ймовірності дефолту.

З наведених вище діграм видно, що не зважаючи на більш високу точність, РБФ-мережа є не такою «обережною», порівняно з багатошаровим персептроном, про що свідчить більш високий відсоток False Positive

помилок (5,83%), ніж False Negative (2,3%). А значить, модель більш схильна видати кредит ризиковому клієнтові, аніж не видати надійному клієнтові.

При цьому, обидві моделі є дуже точними для прогнозування кредитоспроможності, але не такими простими в інтерпретації як більш відомі регресійні та байєсівські моделі.

3.3 Порівняння побудованих моделей та аналіз результатів

Однією з найбільш розповсюджених скорингових моделей є логістична регресія з логіт-функцією активації. Дослідимо адекватність моделі логістичної регресії для даної задачі прогнозування кредитоспроможності. Для визначення найбільш значущих коефіцієнтів будемо на кожному кроці змінювати набір предикторів, додаючи та виключаючи один з предикторів. Таким чином визначимо найбільш вдалий набір предикторів для побудови найбільш ефективної моделі. Після чого, порівняємо результати роботи логістичної регресії з результатами роботи байєсівських та нейронних мереж.

Таким чином, для найкращої комбінаціїї параметрів було отримано наступні значення критеріїв адекватності (рис. 3.3.1).

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
good_bad		AIC	Akaike's Information Criterion	875.9959	.	.
good_bad		ASE	Average Squared Error	0.104207	0.104385	0.10394
good_bad		AVERR	Average Error Function	0.329998	0.335501	0.321743
good_bad		DFE	Degrees of Freedom for Error	892	.	.
good_bad		DFM	Model Degrees of Freedom	108	.	.
good_bad		DFT	Total Degrees of Freedom	1000	.	.
good_bad		DIV	Divisor for ASE	2000	1200	800
good_bad		ERR	Error Function	659.9959	402.6015	257.3944
good_bad		FPE	Final Prediction Error	0.129441	.	.
good_bad		MAX	Maximum Absolute Error	0.979835	0.979835	0.967925
good_bad		MSE	Mean Square Error	0.116824	0.104385	0.10394
good_bad		NOBS	Sum of Frequencies	1000	600	400
good_bad		NW	Number of Estimate Weights	108	.	.
good_bad		RASE	Root Average Sum of Squares	0.322811	0.323086	0.322397
good_bad		RFPE	Root Final Prediction Error	0.359779	.	.
good_bad		RMSE	Root Mean Squared Error	0.341795	0.323086	0.322397
good_bad		SBC	Schwarz's Bayesian Criterion	1406.033	.	.
good_bad		SSE	Sum of Squared Errors	208.4134	125.2615	83.15193
good_bad		SUMW	Sum of Case Weights Times Freq	2000	1200	800
good_bad		MISC	Misclassification Rate	0.143	0.136667	0.1525

Рисунок 3.3.1 – Результати логістичної регресії

Як видно з вирахуваних критеріїв та статистик, модель доволі точно відтворює взаємозв'язок між цільовою змінною та предикторами. Про це свідчать критерії Акайке, а також, загальна точність моделі, що виражена коефіцієнтом міскласифікації. Для моделі логістичної регресії коефіцієнт міскласифікації на валідаційній вибірці має значення 0,136, що свідчить про досить високий рівень адекватності моделі. Розглянемо результати оцінювання ймовірності дефолту клієнта. Для цього скористаємося діаграмою, що показує наявність помилок першого та другого роду у більш візуалізованій формі (рис. 3.3.2).

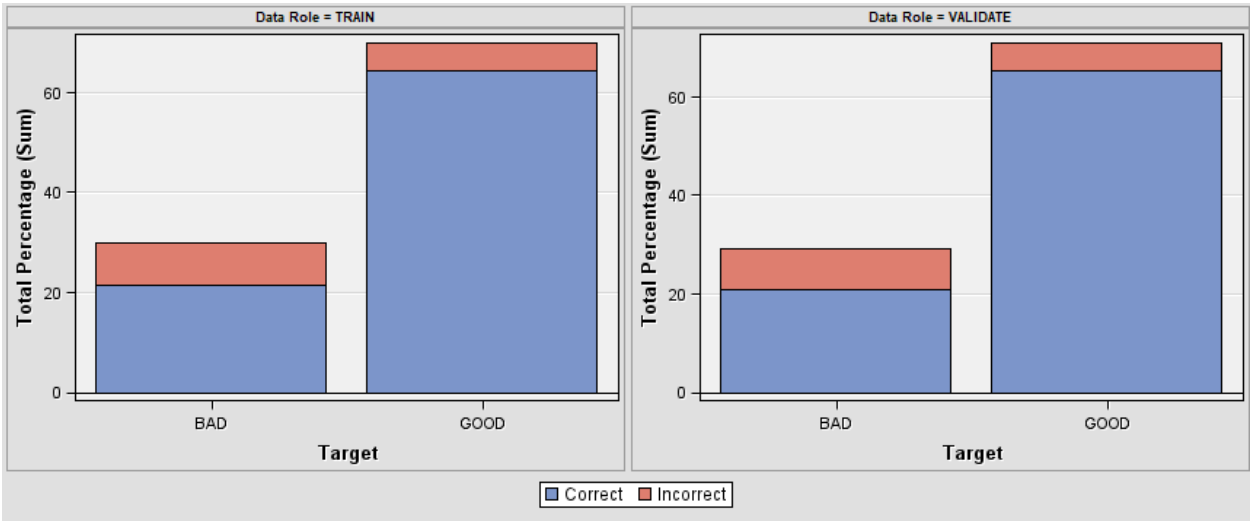

Рисунок 3.3.2 – Аналіз точності класифікації моделі логістичної регресії

Отже, модель логістичної регресії також схильна до більш коректного розпізнавання надійних клієнтів, аніж ризикових.

Для вибору найкращої моделі використовуватимемо компонент Model Comparison в середовищі SAS Enterprise Miner. Як основний критерій будемо використовувати коефіцієнт Міскласифікації (Misclassification Rate). Також для задачі прогнозування кредитоспроможності будемо використовувати статистику Джині (GINI) та кількість правильно спрогнозованих дефолтів клієнтів для тестової вибірки. Для більш повного аналізу додамо також модель дерева рішень, що є найбільш легкою в інтерпретації. Побудуємо

модель дерева рішень на вхідній вибірці без змін, а також, на розширеній вибірці методом групування змінних.

У таблиці 3.3.3 наведено результати аналізу моделей.

Таблиця 3.3.3 – Порівняльна характеристика побудованих моделей

Модель	**Вибірка**	**Misclassification Rate**			**FPR**	**FNR**	**GINI**
		Train	**Validation**	**Test**	**Validation**	**Validation**	**Test**
Наївний байєсівський класифікатор	Початкова	0.267	0.25	0.2925	6,83%	18,6%	0,612
TAN	Початкова	0.219	0.227	0.2075	7,16%	15,5%	0,782
	Розширена	0.126	0.128	0.1225	9,83%	3%	0,876
Багатошаровий персептрон	Розширена	0.107	0.103	0.1125	4,16%	6,16%	0,904
РБФ-мережа	Розширена	0.08	0.082	0.0775	5,8%	2,3%	0,871
Логістична регресія	Розширена	0.143	0.137	0.1525	8%	5,66%	0,846
Дерево рішень	Початкова	0.248	0.236	0.265	16%	8,3%	0,418
	Розширена	0.238	0.243	0.23	9,16%	14,5%	0,473

Отже, з результатів порівняння, неможливо визначити найкращу модель використовуючи лише один критерій Misclassification Rate, адже з одного боку для банку в умовах складної економічної ситуації в Україні найважливішим завданням є визначення ризикових клієнтів та уникнення видачі кредиту, для повернення якого буде необхідно залучати колекторські фірми. З іншого боку, невидача кредитів надійним клієнтам може призвести до відтоку надійних клієнтів, що в свою чергу призводить до фінансових (недоотримання прибутку) та репутаційних ризиків, адже в умовах нинішньої агресивної конкуренції та ринку споживчого кредитування, де велика кількість банків пропонують схожі кредитні продукти, невидача кредиту в одному банку може стати причиною зміни банку клієнтом.

Розглянемо також ROC-криві для кожної побудованої моделі на навчальній, валідаційній та тестовій вибірках (рис. 3.3.4).

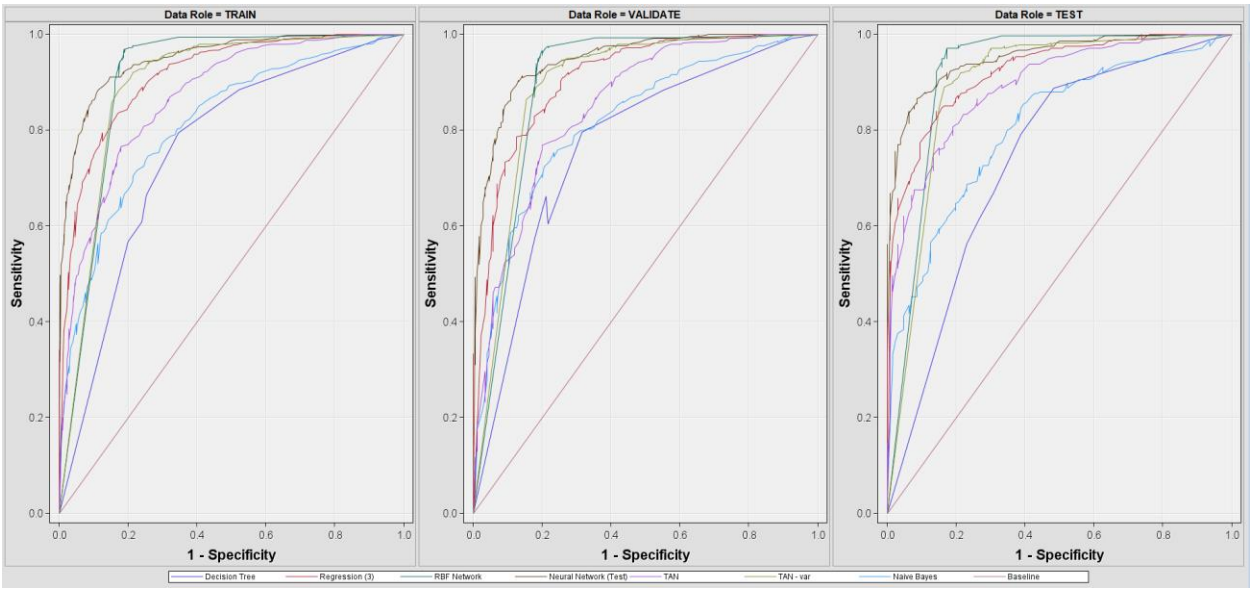

Рисунок 3.3.4 – ROC-аналіз скорингових моделей

За результатами ROC-аналізу, найвище значення коефіцієнта міскласифікації було отримано для моделі РБФ-мережі, що пояснюється насамперед здатністю моделі розпізнавати надійних клієнтів. При цьому, найкраще розпізнає ненадійних клієнтів багатошаровий персептрон, хоча коефіцієнт міскласифікації для нього нижчий.

Тому, для коректного оцінювання та прогнозування кредитоспроможності позичальників, можна використовувати обидві моделі та у випадках розбіжностей між результатами класифікації двох моделей приймати рішення виходячи з відомої для кожної моделі схильності приймати рішення в один чи інший бік.

Також важливою для використання моделей у якості скорингових є здатність моделей «навчатися» на неправильно класифікованих клієнтах. Для цього, порівняємо якість класифікації на навчальній, валідаційній та тестовій вибірках для кожної з досліджуваних моделей. Якість класифікації для усіх моделей, крім РБФ-мережі та мережі TAN, на тестовій вибірці є нижчою, ніж на навчальній. Таким чином, РБФ-мережа та мережа TAN більш здатні до навчання на тренувальній вибірці, ніж інші мережі.

Висновки до розділу

В цьому розділі було розглянуто найбільш популярні методи і підходи Інтелектуального аналізу даних, що використовуються для прогнозування кредитоспроможності позичальників за скоринговим методом, зокрема нейронні мережі та байєсівські мережі. В рамках аналізу, було розглянуто РБФ-мережі, багатошаровий персептрон, наївний байєсівський класифікатор, доповнений деревом байєсівський класифікатор, що були побудовані на основі вибірки даних одного з українських банків з однією цільовою змінною та сімнадцятьма предикторами. Для порівняння адекватності моделей було також побудовано моделі логістичної регресії та дерева рішень, що є найбільш розповсюдженими в сучасних банківських установах України.

За результатами дослідження було виявлено дві найбільш точні моделі: нейронні мережі РБФ та багатошаровий персептрон. Обидві моделі на тестовій вибірці показали результати точності у 90-93 відсотки. В умовах складної економічної ситуації в Україні є сенс використовувати обидві моделі та аналізувати розбіжності класифікації, враховуючи виявлену схильність до поведінки певного типу: багатошаровий персептрон є більш «обережним», а саме у випадку нетипового клієнта має схильність класифікувати його як ненадійного. Натомість, РБФ-мережа схильна до класифікації такого клієнта як надійного.

РОЗДІЛ 4 РОЗРОБКА СТАРТАП-ПРОЕКТУ З ПОБУДОВИ СИСТЕМИ КРЕДИТНОГО СКОРИНГУ ПОЗИЧАЛЬНИКІВ КРЕДИТІВ НА ОСНОВІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В РЕЖИМІ РЕАЛЬНОГО ЧАСУ

4.1 Опис ідеї стартап-проекту

В рамках магістерської дисертації можливе створення стартап-проекту. Більш детальну інформацію про проект наведено в інформаційній карті проекту, в табл. 4.1.

Таблиця 4.1 – Опис ідеї стартап-проекту

1. Назва проекту	Промислова система прогнозування кредитоспроможності позичальників кредитів у режимі реального часу на основі методів інтелектуального аналізу даних
2. Автори проекту	Ревва Роман Володимирович
3. Коротка анотація	Основне призначення системи – автоматичний аналіз кредитоспроможності нового клієнта-позичальника в момент подачі кредитної заявки на основі вже вибраної моделі Інтелектуального аналізу даних. Підтримка стандарту PMML для завантаження готових моделей. Автоматичне прийняття рішень або система сповіщень аналітику.
4. Термін реалізації проекту	12 місяців.

Продовження таблиці 4.1

5. Необхідні ресурси	1 спеціаліст в галузі Data Science, 1 математик, 1 системний архітектор, 1 архітектор рішень, 3 розробники, 1 тестувальник. Сервер для тестування, патент на ідею, фінансові ресурси.
6. Опис проблеми, яку вирішує проект	Проблема оптимізації проблемних кредитів у банку середнього розміру, що викликана відсутністю легкої до впровадження автоматизованої системи аналізу кредитоспроможності позичальників кредитів у режимі реального часу.
7. Очікувані результати	Створена система аналізу кредиспроможності позичальників кредитів у режимі реального часу, що аналізує ризик дефолту клієнта відповідно до побудованих раніше моделей ІАД на основі даних кредитної заявки та приймає рішення відповідно щодо видачі кредиту.
8. Напрямки застосування	Комерційні банки України, кредитні бюро
9. Вигоди для користувача	Зручний інтерфейс, висока швидкість роботи системи, інтеграція з core-banking системами та низька вартість впровадження.

Сильні, слабкі та нейтральні характеристики ідеї проекту зображено в таблиці 4.2.

Таблиця 4.1 - Визначення характеристик ідеї проекту

№ п/п	**Техніко-економічні характеристики ідеї**	**Потенційні товари/концепції конкурентів**			**W (слабка сторона)**	**N (нейтр. сторона)**	**S (сильна сторона)**
		CRM	Scoring Systems	Core Systems			
1.	ІАД	+/-	+	-			+
2.	Онлайн обробка заявок	-	-	+			+
3.	Комунікації	+	-	+/-			+
4.	Моделювання	-	+	-	+		
5.	Відомість бренду	+	+	-	+		

4.2 Технологічний аудит ідеї проекту

Технологічний аудит ідеї проекту наведений у таблиці 4.3.

Таблиця 4.2 - Технологічна здійсненність ідеї проекту

№ п/п	Ідея проекту	Технології її реалізації	Наявність технологій	Доступність технологій

1	Інтеграція з СКБД та Core-banking системами	C#/Java, JSON, SQL	Технології наявні і не потребують змін. Необхідно реалізувати інтерфейс підключення до СКБД.	Технологія загальнодоступна

Продовження таблиці 4.3

2	Підтримка PMML-моделей	PMML	Необхідно реалізувати підтримку стандарту PMML.	Технологія загальнодоступна
3	Система підтримки роботи в режимі онлайн	C#/Java	Технології наявні, необхідна реалізація системи.	Технологія загальнодоступна
Обрана технологія реалізації ідеї проекту: для реалізації проекту обрана клієнт-серверна архітектура. Мова програмування буде визначена спільно з системним архітектором, виходячи з очікуваного навантаження на систему.				

4.3 Аналіз ринкових можливостей запуску стартап-проекту

Характеристика потенційного ринку стартап-проекту наведена у таблиці 4.4

Таблиця 4.4 – Попередня характеристика потенційного ринку стартап-проекту.

№ п/п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	10
2	Загальний обсяг продаж, грн/ум.од	1 000 000
3	Динаміка ринку (якісна оцінка)	Зростає

Продовження таблиці 4.4

4	Наявність обмежень для входу (вказати характер обмежень)	Швидкість обробки інформації, точність обчислень, робота у високонавантаженому середовищі та в умовах захисту персональних даних
5	Специфічні вимоги до стандартизації та сертифікації	В залежності від потреб клієнта
6	Середня норма рентабельності в галузі (або по ринку), %	70

Характеристика потенційних клієнтів стартап-проекту наведена в таблиці 4.5.

Таблиця 4.5 – Характеристика потенційних клієнтів стартап-проекту

№ п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Необхідність швидкої та точної автоматичної обробки заявок на отримання кредитів	Середній та великий бізнес у сфері комерційних фінансових інститутів	Для великого бізнесу найважливішим критерієм є інформаційна безпека та можливість роботи з	Висока точність прогнозування, висока швидкість обробки даних та прийняття рішень

			великими об’ємами даних.	

Можливі загрози для стартап-проекту наведені у таблиці 4.6.

Таблиця 4.3 – Фактори загроз

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Відсутність швидких та безпечних рішень «під ключ»	Бізнес не готовий до впровадження рішень, що не потребують суттєвої кастомізації.	Акцентувати увагу на клієнтах, що користуються big-enterprise рішеннями.
2	Кредитний скоринг виконується за рахунок open-source рішень	Бізнес може бути не готовий до рішень, де не потребується робота програмістів	Існує штат програмістів, що використовують open-source рішення.

Фактори можливостей наведені у таблиці 4.7.

Таблиця 4.4 – Фактори можливостей

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Розробка пілотного проекту впровадження системи в банки другої ліги.	Пропозиції впровадити системи у пілотну експлуатацію для клієнта для демонстрації можливостей.	Виділення даних на яких може бути побудована система.

Продовження таблиці 4.7

2	Автоматизована система прийняття рішень щодо кредитування фізичних осіб	В залежності від ймовірності дефолту клієнта система автоматично надсилає повідомлення клієнту зі статусом заявки. У випадку ймовірності близької до 50% надсилається сповіщення аналітику.	Підвищення ефективності кредитування та більш поглиблений аналіз «складних» клієнтів

Проведений ступеневий аналіз конкуренції на ринку зображено у таблиці 4.8.

Таблиця 4.5 – Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
1. Чиста конкуренція	Гравці ринку не мають явних переваг один над одним	Більш вигідні умови на тендерах, агресивний маркетинг
2. Регіональна конкуренція	Гравці ринку – інтернаціональні підприємства, або українські підтриємства, що мають офіси в інших країінах.	За рахунок локального менеджменту, процес прийняття рішень є більш гнучким.

Продовження таблиці 4.8

3. Внутрішньогалузева конкуренція	Гравці ринку знаходяться в одній галузі – розробці ПЗ	
4. Товарно-видова конкуренція	Усі продукти гравців ринку мають різне призначення, але виконують схожі функції.	Розробка найбільш простого з точки зору інтеграції продукту.
5. Конкурентні переваги нецінові	Продукти відрізняються гнучкістю, призначенням та ліцензійною політикою.	Підхід демо-версій; у маркетингу наголошувати на тому, що рішення буде зроблене «під ключ»
6. Марочна конкуренція	Значна увага приділяється бренду, що розробив продукт	Партнерство, маркетинг.

Таблиця 4.6 – Аналіз конкуренції в галузі за М. Портером

	Прямі конкуренти в галузі	Потенційні конкуренти	Постачальники	Клієнти	Товари-замінники
Складові аналізу	FICO, Microsoft, SAS	Локальні українські аутсорсингові компанії, компанії-партнери світових вендорів	Від зростання кількості постачальників незначно зростуть продажі	Споживачі замовлюють додаткові модулі до програмного продукту	Замінником товару може стати використання open-source рішень

Продовження таблиці 4.9

Висновки:	Продукти співіснують незалежно і явно не конкурують	Формально, аналогічного продукту на ринку немає, тож можливість виходу на ринок є. Умовою виходу на ринок є система, що не потребуватиме додаткових затрат на впровадження та буде готова «під ключ». Термін виходу на ринок – 1 рік	Постачальники не диктують умови роботи на ринку.	Клієнти диктують умови лише для додаткових модулів, що розроблюються для них. Наприклад, замовник може не захотіти, щоб модуль, розроблений для нього, пропонувався іншим замовникам	Для того, щоб потенційний клієнт обрав саме цей продукт, необхідно, щоб він був максимально простий, швидкий і надійний у використанні, а також був гнучким і маштабованим.

Таблиця 4.7 – Обґрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)
1	Інтеграція	Продукт надає інтерфейс підключення до СКБД та core-banking систем (система веб-сервісів), що дозволяють інтеграцію з більшістю існуючих банківських систем та БД.
2	Модульність	Кожен замовник індивідуально обирає мовленнєві пакети для себе, при цьому не втрачаючи можливість докупити інші пакети в майбутньому
3	Ціна	Оскільки продукт не потребує кастомізації, ціна може бути значно нижча, ніж у прямих конкурентів.
4	Автоматизація	Продукт автоматично виконує оцінку кредитоспроможності та надає відповідь на кредитну заявку у більшості випадків.

Порівняльний аналіз сильних та слабких сторін проекту відображено у таблиці 4.11.

Таблиця 4.8 – Порівняльний аналіз сильних та слабких сторін «FICO»

№ п/п	Фактор конкурентоспроможності	Бали 1-20	Рейтинг товарів-конкурентів у порівнянні з FICO						
			-3	-2	-1	0	1	2	3

1	Інтеграція	12					*		
2	Модульність	8			*				

Продовження таблиці 4.11

3	Ціна	20							*
4	Автоматизація	8			*				

SWOT-аналіз проекту наведено в таблиці 4.12.

Таблиця 4.12 - SWOT-аналіз стартап-проекту

Сильні сторони: Ціна, гнучкість, інтеграція	Слабкі сторони: відсутність середовища розробки власних моделей
Можливості: Рішення «під ключ»	Загрози: Неточність розпізнавання, відсутність попиту

Альтернативи ринкового впровадження проекту розглянуто в таблиці 4.13.

Таблиця 4.13 – Альтернативи ринкового впровадження стартап-проекту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	Розробка системи створення моделей ІАД в системі	Середня	12-24 місяців
2	Хмарний сервіс	Висока	6-9 місяців

4.4 Розроблення ринкової стратегії проекту

Опис та вибір цільових груп потенційних клієнтів зображено в таблиці 4.14

Таблиця 4.14 - Вибір цільових груп потенційних споживачів

№ п/п	Опис цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Кредитні бюро та установи	Середня готовність, через необхідність вкладання великих коштів, але відсутність на ринку рішень для кредитного скорингу за сприятливою ціною	Високий попит	Середня	Вхід в сегмент простий
2	Середні банки (до 1млн клієнтів)	Висока готовність. Банки середнього розміру здебільшого не мають бюджету на відомі міжнародні рішення.	Середній попит	Середня	Вхід в сегмент середній

Продовження таблиці 4.14

3	Великі приватні банки світових груп	Більшість компаній такого рівня застосовували чи планують застосовувати інтелектуальні системи, але надають перевагу великим системами провідних світових вендорів, що експлуатуються в інших офісах банків материнської групи	Дуже високий попит	Висока	Вхід в сегмент дуже складний
4	Великі локальні приватні банки	Висока готовність. Локальні банки готові розглянути альтернативні рішення, адже не мають обмежень використовування лише рішень певних вендорів	Дуже високий попит	Висока	Вхід в сегмент середній
5	Державні банківські установи	Низька готовність. Державні банківські установи мають обмеження з боку регуляторів на вибір вендорів	Середній попит	Висока	Вхіж в сегмент дуже складний
Які цільові групи обрано: 1,2,4					

В таблиці 4.15 зображено вибір базової стратегії розвитку.

Таблиця 4.15 - Визначення базової стратегії розвитку

№ п/п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку
1	Розробка та створення додаткових функціональних модулів (розробка середовища побудови власних моделей)	Таргетні пропозиції бізнесу, проведення презентації функціональних рішень на конференціях	Наявність середовища розробки моделей у конкурентних продуктів	Розробка та удосконалення існуючих модулів на основі потреб ринку та інформації від клієнтів

В таблиці 4.16 наведено визначення базової стратегії конкурентної поведінки.

Таблиця 4.16 - Визначення базової стратегії конкурентної поведінки

№ п/п	Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки

Продовження таблиці 4.16

1	Ні	Можливі обидва варіанти	Стандартні функціональні модулі будуть виконувати схожі функції, що і конкуренти	Унікальна цінова політика, мінімальний об’єм необхідних сервісних робіт для налаштування системи

В таблиці 4.17 наведено визначення стратегії позиціонування.

Таблиця 4.17 - Визначення стратегії позиціонування

№ п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
1	Висока якість прогнозування в клієнтській сфері застосування, надійність, безпека використання, ціна	Розробка та удосконалення існуючих модулів на основі потреб ринку та інформації від клієнтів	Інтеграція з іншими системами, легкість впровадження, автономність	Автоматичне виявлення ненадійних позичальників та комунікація з ними; український банківський продукт «під ключ»

4.5 Розроблення маркетингової програми стартап-проекту

В таблиці 4.18 представлені ключові переваги концепції потенційного товару.

Таблиця 4.18 - Визначення ключових переваг концепції потенційного товару.

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити
1	Автоматизація бізнес-процесів	Спрощення бізнес-процесу кредитного скорингу	Покращує ефективність підрозділу кредитної оцінки фізичних осіб
2	Можливості інтеграції	Спрощення взаємодії між системами	Забезпечує більш ефективне вирішення задач у звуженій сфері застосування
3	Фінансові ресурси	Зменшення витрат на впровадження та підтримку	Забезпечує більш ефективне використання коштів
4	Людські ресурси	Звільнення людських ресурсів від рутинних завдань	Забезпечує звільнення ключових аналітиків від рутинних завдань та їх концентрація на методах роботи із складними клієнтами

Опис трьох рівнів моделі товару відображено у таблиці 4.19.

Таблиця 4.19 - Опис трьох рівнів моделі товару

Рівні товару	Сутність та складові		
I. Товар за задумом	Обробка, аналіз даних. Прогнозування кредитоспроможності позичальників кредитів в режимі реального часу		
II. Товар у реальному виконанні	Властивості/характеристики	М/Нм	Вр/Тх /Тл/Е/Ор
	Швидкодія Ефективність Користувацький інтерфейс	Нм Нм Нм	Тх/Тл/Е Тх/Тл Е
	Якість: стандарти відповідні до законодавства. Моделі Інтелектуального аналізу даних високої точності.		
	Пакування: Власний сайт		
	Марка: BRI Solutions, BRI		
III. Товар із підкріпленням	До продажу: Консультування з питань кредитного скорингу Після прожажу: Консультування з питань підготовки даних		
Патент, закрита реалізація готових методів і моделей ІАД, захищена від можливостей декомпіляції.			

Визначення меж встановлення ціни показано в таблиці 4.20.

Таблиця 4.20 - Визначення меж встановлення ціни

№ п/п	Рівень цін на товари-замінники	Рівень цін на товари-аналоги	Рівень доходів цільової групи споживачів	Верхня та нижня межі встановлення ціни на товар/послугу

1	-	Від 15000$	Рівень доходів підприємств надзвичайно високий	100-2500$/міс в залежності від кількості заявок.

Формування системи збуту зображено в таблиці 4.21.

Таблиця 4.21 - Формування системи збуту

№ п/п	Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник товару	Глибина каналу збуту	Оптимальна система збуту
1	Таргетні пропозиції для компаній	Презентації функціоналу, пілотні проекти	-	Прямі продажі

Концепція маркетингових комунікацій відображена у таблиці 4.22.

Таблиця 4.22 - Концепція маркетингових комунікацій

№ п/п	Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуютьс я цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомленн я	Концепція рекламного звернення
1	Кредитні бюро	Соціальні мережі, прямі комунікації	Рішення «під ключ», автоматизована система фільтрації ризикових клієнтів.	Донести інформацію про важливість скорингових систем	Економія коштів, що витрачаються на колекторські випадки, яких можна було уникнути з використання м скорингу.

Продовження таблиці 4.22

2	Середній український банк	Тендери, внутрішньо-ринкова комунікація, конференції, пряма комунікація	Автоматична обробка вхідних заявок на основі сучасних методів ІАД	Короткий опис переваг продукту, заохочення дізнатись більше. Презентація кейсів.	Оцінювання кредитоспроможності клієнта на етапі його заявки на кредит та автоматична генерація відповіді
3	Великий локальний приватний банк	Конференції, прямі комунікації, тендери	Звільнення аналітиків від необхідності в ручному режимі запускати процеси оцінювання кредитоспроможності та процеси комунікації, легка інтеграція та впровадження	Донести інформацію про оптимальність рішення для бізнесу клієнта	Тільки прямі продажі, фокус на ROI

Висновки до розділу

Отже, за результатами проведеного дослідження ринку банківського аналітичного ПЗ, можна стверджувати про наявність попиту на запропоновану систему на українському банківському ринку, а саме в локальних приватних банках, середніх за розміром банках з іноземним капіталом, а також, кредитних бюро/кафе. Слід зауважити, що конкуренція на ринку банківського аналітичного ПЗ є високою, тому не дивлячись на іноваційність та зручність

впровадження продукту, необхідний фокус на побудові ефективного маркетингу та відділу продажів. Відповідно до сегментації ринку, було розроблено стратегії внутрішньоринкової комунікації та піар-кампанії окремо для кожного сегменту. За умови наявності початкового фінансування, продукт є потенційно конкурентоспроможним на ринку.

ВИСНОВКИ ПО РОБОТІ ТА ПЕРСПЕКТИВИ ДЛЯ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В роботі виконано огляд та застосування методів Інтелектуального аналізу даних для вирішення задачі прогнозування кредитоспроможності клієнтів. Розглянуто найбільш розповсюджені методи побудови скорингових моделей на основі інтелектуального аналізу даних. Виявлено особливості моделювання кредитоспроможності на основі моделей різних типів.

В роботі виконано порівняння двох пакетів ПЗ, що дозволяють виконувати побудову та аналіз скорингових моделей: Python та SAS. Також, пронаналізовано особливості скорингової системи FICO. Розглянуто основні види скорингових моделей, тести на адекватність та критерії якості побудованих моделей. Для оцінки якості моделі вибрано кількість помилок першого та другого роду, коефіцієнт міскласифікації та критерій Джині.

Досліджено та застосовано сучасну аналітичну методологію SEMMA як загальну методику для аналізу економетричних даних.

Для побудови та навчання скорингових моделей було вибрано методи нейронних мереж та байєсівських мереж. Застосовано розроблену методологію до аналізу кредитоспроможності клієнтів за скоринговим методом, використовуючи SAS.

На основі моделювання і прогнозування визначені кращі моделі для аналізу кредитоспроможності клієнтів:

– РБФ-мережа;
– Багатошаровий перспетрон.

Для подальших досліджень є перспективи подальшого вдосконалення та розробки архітектури системи, поліпшення побудованих моделей для прогнозування, а також, застосування інших методів та моделей (метод градієнтного бустінгу, метод опорних векторів та ін.). Також, враховуючи

сучасні тренди у сфері аналізу даних, існує перспектива застосування методів обробки текстових даних, аналізу поведінки клієнта у соціальних мережах з використанням методів текстової аналітики та сентиментального аналізу для подальшого вдосконалення точності, адекватності та легкості моделей кредитного скорингу.

ЛІТЕРАТУРА

1. Вовк, В. Я., Хмеленко, О. В. Кредитування і контроль [Текст] : навч. посіб. / В. Я. Вовк, О. В. Хмеленко. – К.: «Знання», 2008. – 463 с. – ISBN: 978- 966-346-402-2;

2. 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.

3. Thomas L.C. Credit Scoring and its applications: Monograph / Lyn C. Thomas, David B. Edelman, Jonathan N. Crook. — Philadelphia: SIAM, 2002. — 248 p.

4. 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.

5. Boggess, W. P. Screen-test your credit risks. / Harvard Business Review, 1967. – 113-122p.;

6. Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. - Morgan Kaufmann Publishers is an imprint of Elsevier, 2011. – 664 c.

7. Кредитний ризик комерційного банку: Навч. посібник / В.В. Вітлінський, О.В. Пернарівський, Я.С. Наконечний, Г.І. Великоіваненко /За ред. В.В. Вітлінського. – К.: Т-во «Знання», КОО, 2000. – 251с.

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. В.В. Вітлінський Ризик у Менеджменті / Вітлінський В.В., Наконечний С.І. — М.; Борисфен, 1996. – 336с.

10. Siddiqi N. Credit risk scorecards: developing and implementing

intelligent credit scoring / Naeem Siddiqi. — Hoboken: John Wiley & Sons, Inc., 2006. — 196 p.

11. 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.

12. 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.

13. Терентьев А.Н. SAS BASE: Основы программирования / А.Н. Терентьев, В.Н. Домрачев, Р.И. Костецкий. — К.: Эдельвейс, 2014. — 304 с.

14. 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.\

15. 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.

16. 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.