МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет
імені Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

ДОПУЩЕНО ДО ЗАХИСТУ
Завідувач кафедри інтелектуальних
інформаційних систем, д-р техн. наук, проф.
_____Ю. П. Кондратенко
«____»_____2022 р.

БАКАЛАВРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

ВИКОРИСТАННЯ НЕЙРОННИХ МЕРЕЖ ДЛЯ КРЕДИТНОГО СКОРИНГУ ПІДПРИЄМСТВА

Спеціальність 122 «Комп'ютерні науки»

122 – БКР – 401.21810304

Виконав студент 4 курсу, групи 401
_____***О. І. Бородін***
«20» червня 2022 р.

Керівник: ***докт. техн. наук, професор***
_____ ***Ю. П. Кондратенко***
«20» червня 2022 р.

Миколаїв – 2022

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет ім. Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

Рівень вищої освіти **бакалавр**
Спеціальність **122 «Комп'ютерні науки»**
(шифр і назва)
Галузь знань **12 «Інформаційні технології»**
(шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри інтелектуальних
інформаційних систем, д-р техн. наук, проф.
_____ Ю. П. Кондратенко
«____» _____2021 р.

З А В Д А Н Н Я
на виконання кваліфікаційної роботи

Видано студенту групи 401 факультету комп'ютерних наук Бородіну Олександру Ігоровичу.

1. Тема кваліфікаційної роботи «Використання нейронних мереж для кредитного скорингу підприємства».

Керівник роботи Кондратенко Юрій Пантелійович, д-р техн. наук, професор.

Затв. наказом Ректора ЧНУ ім. Петра Могили від «___» _____ 20___ р. № ____

2. Строк представлення кваліфікаційної роботи студентом «___» _____ 20__ р.

3. Вхідні (початкові) дані до роботи: стандартизована фінансова звітність (балансові рахунки) 10 тисяч німецьких підприємств.

Очікуваний результат: система прогнозування кредитного рейтингу підприємства.

4. Перелік питань, що підлягають розробці (зміст пояснювальної записки):

– аналіз сучасного стану задачі прогнозування кредитного рейтингу підприємства;

– огляд існуючих методів та моделей вирішення поставленої задачі;

– проектування системи прогнозування кредитного рейтингу, використовуючи штучні нейронні мережі;

– порівняльний аналіз результатів застосування обраних методів машинного навчання для розв’язання поставленої задачі.

5. Перелік графічних матеріалів: сторінок – 97, таблиць – 8, рисунків – 14, посилань – 44, додатків – 1, презентація.

6. Завдання до спеціальної частини: «Визначення оптимальних умов праці на робочому місці»

7. Консультанти розділів роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис
Спеціальна частина з охорони праці	Макарова О.В., старший викладач	

Керівник роботи д-р техн. наук, проф. Кондратенко Ю. П.
(наук. ступінь, вчене звання, прізвище та ініціали)

_____
(підпис)

Завдання прийнято до виконання Бородін О. І.
(прізвище та ініціали)

_____
(підпис)

Дата видачі завдання «23» листопада 2021 р.

КАЛЕНДАРНИЙ ПЛАН
виконання бакалаврської кваліфікаційної роботи

Тема: Використання нейронних мереж для кредитного скорингу підприємства.

№	Найменування роботи	Початок	Закінчення	Примітки
1	Подання заяви на затвердження теми та керівників БКР	30.10.2021	30.10.2021	
2	Отримання завдання на виконання БКР	25.11.2021	25.11.2021	
3	Складання календарного плану роботи на весь період виконання БКР	10.12.2021	10.12.2021	
4	Отримання завдання на переддипломну практику	23.05.2022	23.05.2022	
5	Проходження переддипломної практики, збір та аналіз матеріалів до БКР	23.05.2022	04.06.2022	
6	Розробка звіту з переддипломної практики	04.06.2022	06.06.2022	
7	Виконання БКР: 1. Аналіз сучасного стану задачі прогнозування кредитного рейтингу підприємств. 2. Огляд існуючих моделей. 3. Проектування системи. 4. Моделювання та аналіз результатів.	28.02.2022 та 06.06.2022	27.03.2022 та 19.06.2022	
8	Попередній захист БКР на засіданні комісії кафедри	30.05.2022	31.05.2022	
9	Доробка та остаточне оформлення БКР	02.06.2022	20.06.2022	
10	Подання БКР рецензенту	16.06.2022	18.06.2022	
11	Подання БКР, її електронної копії та інших документів (відгуку, рецензії) до захисту	20.06.2022	22.06.2022	
12	Захист БКР перед екзаменаційною комісією (ЕК)	27.06.2022	29.06.2022	

Розробив студент Бородін О. І. _____
(прізвище, ім'я, по батькові студента) (*підпис*)

Керівник роботи д.т.н., проф. Кондратенко Ю. П. _____
(*посада, прізвище, ім'я, по батькові)* *(підпис)*

«13» грудня 2021 р.

АНОТАЦІЯ

до бакалаврської кваліфікаційної роботи

студента 401 групи ЧНУ ім. Петра Могили

Бородіна Олександра Ігоровича

на тему: **«ВИКОРИСТАННЯ НЕЙРОННИХ МЕРЕЖ ДЛЯ КРЕДИТНОГО СКОРИНГУ ПІДПРИЄМСТВА»**

Прогнозування кредитного рейтингу має величезний потенціал для сприяння економічному зростанню світової економіки і може бути цінним інструментом для підвищення фінансової інклюзії та ефективності. **Метою роботи** є визначення кредитного рейтингу підприємства на основі методів машинного навчання. **Об'єкт дослідження** – процес оцінки фінансового ризику з використанням штучного інтелекту та методів машинного навчання. **Предметом дослідження** є застосування багатошарових нейронних мереж та випадкового лісу для визначення кредитоспроможності компанії на основі показників стандартизованої фінансової звітності, в умовах часткової відсутності даних.

Робота складається з п'яти розділів. У першому розділі проведено аналіз наявних аналогів та публікацій, поставлена задача кредитного скорингу. У другому розділі детально розглянуто математичні моделі, розібрані принципи роботи нейронних мереж. У третьому розділі проаналізовано вхідний набір даних, обрано програмні засоби та описано принципи роботи нейронної мережі та випадкового лісу. У четвертому розділі проведено моделювання та описано результати дослідження. У п'ятому розділі описано норми до умов праці та основи техніки безпеки.

Загальний обсяг роботи: 97 – сторінок. Бакалаврська робота містить один додаток, 14 рисунків, 8 таблиць і посилання на 44 джерела.

Ключові слова: кредитний скоринг, машинне навчання, нейронні мережі, фінансовий аналіз.

ABSTRACT

to the bachelor's qualification work by the student of the group 401
of Petro Mohyla Black Sea National University

Oleksandr Borodin

"APPLICATIONS OF ARTIFICIAL NEURAL NETWORKS FOR ENTERPRISE CREDIT SCORING"

Forecasting a credit rating has a great potential for promoting the economic growth of the world economy and can be a valuable tool for promoting financial inclusion and efficiency. The goal of the current work is to assign a business credit rating based on machine learning methods. The object of this study is the process of assessing financial risk using artificial intelligence and machine learning methods. The subject of the study is the establishment of a deep neuron network and a random forest to determin the company's creditworthiness on the basis of indications in a standardized financial rate, in conditions of missing values.

The work is made up of five paragraphs. At the first paragraph, an analysis of recent analogues and publications was carried out, the task of credit scoring was set. In the second, the mathematical models, the principles of the work of neural networks were examined in detail. In the third, the input data was analyzed, the program features were collected, and the principles of the work of the neural network annd the random forest were described. In the fourth paragraph, a simulation was carried out and the results of the research were discussed. The fifth paragraph describes the norms of working practice and the basics of safety technology.

Pages – 97, figures – 14, tables – 8, references – 44, appendices – 1.

Key words: credit scoring, machine learning, neural measures, financial analysis.

ЗМІСТ

ВСТУП 4
1 ПОСТАНОВКА ЗАДАЧІ КРЕДИТНОГО СКОРИНГУ ПІДПРИЄМСТВА 6
1.1 Актуальність проблеми 6
1.2 Огляд та аналіз наявних аналогів та публікацій 10
1.3 Постановка задачі 12
Висновки до розділу 1 13
2 МАТЕМАТИЧНІ МОДЕЛІ КРЕДИТНОГО СКОРИНГУ 14
2.1 Фінансові коефіцієнти 14
2.2 Дискримінантний аналіз 23
2.3 Моделі бінарної відповіді 25
2.4 Методи машинного навчання 27
Висновки до розділу 2 32
3 ПРОЄКТУВАННЯ СИСТЕМИ КРЕДИТНОГО СКОРИНГУ 33
3.1 Підготовка та аналіз набору даних 34
3.2 Вибір бібліотек машинного навчання 40
3.3 Характеристика штучних нейронних мереж 44
3.4 Характеристика моделі випадкового лісу 55
Висновки до розділу 3 57
4 МОДЕЛЮВАННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ СИСТЕМИ 58
4.1 Реалізація нейронної мережі 58
4.2 Аналіз результатів нейронної мережі 60
4.3 Аналіз результатів випадкового лісу 62
Висновки до розділу 4 68
5 ОХОРОНА ПРАЦІ 71
ВИСНОВКИ 87
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 89
ДОДАТОК А Програмний код класу нейронної мережі 93

ПЕРЕЛІК СКОРОЧЕНЬ

ЗНМ	– згорткові нейроні мережі
ІІС	– інтелектуальна інформаційна система
МСП	– малі та середні підприємства
НМ	– нейронна мережа
ШІ	– штучний інтелект
ШНМ	– штучні нейроні мережі
ANN	– artificial neural network
CR	– current ratio
DT	– decision tree
EM	– expectation maximization
MSE	– mean squared error
PD	– probability of default
QR	– quick ratio
RF	– random forest
RMSE	– root mean squared error
ROA	– return on assets
ROE	– return on equity
SVM	– support vector machines

Пояснювальна записка

до кваліфікаційної роботи

на тему:

«ВИКОРИСТАННЯ НЕЙРОННИХ МЕРЕЖ ДЛЯ КРЕДИТНОГО СКОРИНГУ ПІДПРИЄМСТВА»

Спеціальність 122 «Комп’ютерні науки»

122 – БКР – 401.21810304

Виконав студент 4 курсу, групи 401

_____***О. І. Бородін***

«20» червня 2022 р.

Керівник: ***докт. техн. наук, професор***

_____ ***Ю. П. Кондратенко***

«20» червня 2022 р.

Миколаїв – 2022

ВСТУП

Запорукою успіху будь-якого бізнесу є правильна оцінка кредитоспроможності потенційних ділових партнерів. Ця умова діє для усіх суб'єктів господарювання: інвесторів, постачальників, замовників, фінансових установ тощо. Зазвичай кредитоспроможність вимірюється з точки зору кредитного рейтингу, який традиційно визначався спеціальними аналітичними агентствами, до яких можна віднести Moody's, S&P, Fitch та інші. Ця система передбачає, що компанії, які прагнуть отримати оцінку власного кредитного рейтингу, вимушені сплатити комісію переліченим вище агентствам, що є досить коштовним та тривалим процесом. Таким чином, тільки великі корпорації, які, як правило, котуються на фондових біржах, можуть дозволити собі послуги відомих кредитних агентств.

Однак, переважна частина економіки складається із малих підприємств, які також потребують доступ до кредитів для своєї підприємницької діяльності [8]. Це породжує необхідність у знаходженні альтернативних способів обчислення кредитної оцінки для великої кількості суб'єктів господарювання, враховуючи обмеженість доступних економічних даних для малих підприємств. Враховуючи дану потребу, виникли PD моделі або моделі «ймовірності дефолту» (PD – Probability of Default) [35], які можна застосувати до великих наборів даних майже автоматично.

У даній роботі було досліджено два різних підходи, які використовують сучасні статистичні методи та методи машинного навчання для прогнозування кредитного рейтингу підприємства. Перший метод заснований на штучній нейронній мережі та добре працює з необробленими наборами даних (у нашому випадку – фінансовими балансами) із пропусками даних, і є підходом «чорного ящика». Другий підхід є більш аналітико-статистичним та відомий як Random Forest. У ньому спочатку виводяться фінансові коефіцієнти, а потім –

використовується статистичне пояснення для заповнення прогалин у даних та визначення кредитного рейтингу.

Додатково у роботі було проаналізовано проблему відсутності даних, з якою стикаються аналітики при спробі застосувати моделі прогнозування кредитного рейтингу компаній: для заповнення пропущених даних було застосовано алгоритм максимізації очікувань (Expectation Maximization, EM).

Метою роботи є визначення кредитного рейтингу підприємств на основі методів машинного навчання.

Об'єктом дослідження виступає процес оцінки фінансового ризику з використанням штучного інтелекту та методів машинного навчання.

Предметом дослідження є застосування багатошарових нейронних мереж та випадкового лісу для визначення кредитної привабливості компанії на основі показників стандартизованої фінансової звітності, в умовах часткової відсутності даних.

Для досягнення поставленої мети потрібно виконати такі завдання:

– дослідити предметну область кредитного скорингу підприємств;

– провести аналіз наявних методів та інструментів для вирішення поставленої мети;

– розробити архітектуру нейронної мережі та навчити її класифікувати компанії відповідно до вхідних показників фінансової звітності;

– підібрати оптимальний алгоритм обробки пропущених даних та навчити модель випадкового лісу вирішувати задачу кредитного скорингу;

– розробити програмний застосунок на базі навчених моделей та провести тестування на запропонованих наборах, проаналізувати отримані результати.

1 ПОСТАНОВКА ЗАДАЧІ КРЕДИТНОГО СКОРИНГУ ПІДПРИЄМСТВА

Зараз нейронні мережі відіграють дуже важливу роль у науковому світі. Вони спрощують автоматизацію виробничих процесів за допомогою класифікації об’єктів у реальному часі, розпізнають та синтезують людську мову, надають рекомендації та поради, враховуючи інтереси споживачів, або ж використовуються у медичних дослідженнях для діагностики захворювань. Методи машинного навчання також активно застосовуються у фінансах та економічному аналізі. Вони можуть передбачати різні часові ряди, оцінювати ймовірність банкрутства компанії або ж виступати консультантами для клієнтів банку.

Важливою частиною діяльності будь-якого гравця фінансового ринку є оцінка ризиків. Останні наукові дослідження показують, що моделі машинного навчання дозволяють точно прогнозувати кредитні ризики, інколи повністю замінюючи роботу досвідченого фінансового аналітика.

1.1 Актуальність проблеми

Прогнози корпоративних дефолтів та моделі кредитного скорингу використовуються у різних сферах економіки. Корпорації можуть діагностувати свій поточний стан на основі моделей прогнозування та визначати свої стратегії. Керівники можуть здійснювати управління бізнесом більш стабільно, відслідковуючи ключові індекси та індикатори, які визначають ризик корпоративного дефолту. Інвестори можуть переглянути свої стратегії та покращити свої портфелі, вивчивши ймовірність корпоративних дефолтів. Крім того, уряди можуть встановлювати макропруденційну політику та покращувати пов'язані з ними фінансові розрахунки, використовуючи корпоративні прогнози дефолту. Таким чином, моделі прогнозування за замовчуванням допомагають у розробці та вдосконаленні фінансової системи. Крім того, використовуючи алгоритми машинного навчання та статистичні моделі, корпоративні прогнози за замовчуванням знаходяться на передньому краї передової фінансової інженерії.

Недавня світова фінансова криза і збільшення кредитного ризику підкреслюють важливість цієї галузі.

Кредитно-рейтингова система.

Кредитний рейтинг – це процес кількісної оцінки достовірності позичальника щодо його існуючих фінансових зобов'язань або його частин. Таким чином, кредитний рейтинг може бути застосований до будь-якої фізичної особи, або організації, яка просить кредит або видає будь-яку форму боргу.

У випадку з фізичними особами кредитний рейтинг дуже поширений у англосаксонських країнах і розраховується такими агентствами, як TransUnion, Experian, які відповідають кредитному скорингу Fair Isaac (FICO).

Кредитна оцінка для компаній та урядів – це кредитний рейтинг облігацій, зосереджений на їх потенційній здатності погашати корпоративні або державні облігації. Професіонали та інвестори використовують цю інформацію, опубліковану кредитними рейтинговими агентствами, при оцінці своїх інвестицій. Кредитні рейтинги облігацій, як правило, виконуються Standard & Poor's (S&P), Moody's або Fitch, які також відомі як «Велика трійка». Ці рейтингові агентства оплачуються державним агентством, яке оформило борг. Як ми побачимо у розділі ця специфічна динаміка викликала деякі суперечки через відсутність незалежності, за своєю суттю пов'язаної з цим механізмом.

Індустрія кредитного рейтингу надзвичайно сконцентрована на глобальному рівні, оскільки агентства Великої трійки контролюють ринок. Ці три компанії почали аналізувати кредитні рейтинги на початку XX століття, створюючи своєрідну монополію у своїй сфері. Однак їх роль у фінансовій кризі 2008 року викликала зростаючу стурбованість серед тих, хто ставить під сумнів, чому їх бізнес-модель надає їм достатню незалежність у своїй оцінці, враховуючи, що вони отримують свій прибуток від тих же компаній, які вони судять [40].

Питання незалежності кредитно-рейтингових агентств є дуже складним та стосується інституційних та законодавчих аспектів. Зростаюча увага до того, які саме критерії оцінки використовуються кредитно-рейтинговими агентствами,

виправдовується впливом їх рішень на ринок та суверенні уряди, прибутковість облігацій яких також залежить від рішень агентств. Фінансова криза 2008 року поставила під колективний осуд драматичні наслідки неадекватного кредитного рейтингу, коли Велика трійка не змогла належним чином оцінити фінансові продукти, які були отримані в результаті погіршення кредитів. У кінцевому підсумку це призвело до краху Lehman Brothers, Bear Stearns та інших інвестиційних банків, а також до найбільшої світової економічної кризи за останні 80 років [26].

Після цих подій роль невеликих кредитно-рейтингових агентств зросла за останнє десятиліття [29], частково завдяки зменшенню репутації Великої трійки. Фактично, їх також звинувачують у негативному «упередженому ставленні до сусідства», тобто у більш м'якому підході до американських компаній у порівнянні з європейськими колегами. Цей факт підштовхнув багатьох гравців розглянути ідею Європейського кредитного агентства [16], хоча на даний момент офіційних кроків у цьому напрямку зроблено не було.

У цілому кредитний рейтинг відіграє ключову роль у фінансово-економічній динаміці сучасного суспільства, оскільки істотно впливає на можливості потенційного позичальника отримати кредит або оформити новий борг. При цьому процентна ставка нового кредиту буде залежати від кредитного рейтингу суб'єкта господарювання на момент видачі. Цей аспект лежить в основі раніше згаданого конфлікту інтересів, оскільки кредитно-рейтингові агентства оплачуються тими компаніями, які можуть отримати прибуток або збитки від їх рішень.

У таблиці 1.1 перераховано кредитні бали для Moody's та S&P/Fitch, із їх деталізованим описом. Для S&P та Fitch прогнозована динаміка оцінки виражається знаком +/–. У базі даних ModeFinance, яку буде проаналізовано у основній частині роботи, прогнозована динаміка оцінки не розглядається.

Таблиця 1.1 – Опис класів кредитного рейтингу підприємства

Moody's	**S&P, Fitch**	**Кредитний рейтинг**
Aa	AAA	Позичальник має НАДЗВИЧАЙНО СИЛЬНУ здатність виконувати свої фінансові зобов'язання.
Aa1 Aa2 Aa3	AA+ AA AA-	Позичальник має ДУЖЕ СИЛЬНУ здатність виконувати свої фінансові зобов'язання. Він лише трохи відрізняється від найкращих позичальників.
A1 A2 A3	A+ A-	Позичальник має СИЛЬНУ здатність виконувати свої фінансові зобов'язання, але більш сприйнятливий до несприятливих наслідків змін обставин та економічних умов, ніж категорії з більш високим рейтингом.
Baa1 Baa2 Baa3	BBB+ BBB BBB-	Позичальник має АДЕКВАТНУ здатність виконувати свої фінансові зобов'язання. Несприятливі економічні умови або мінливі обставини, швидше за все, послаблять його здатність виконувати фінансові зобов'язання.
Ba1 Ba2 Ba3	BB+ BB BB-	Позичальник МЕНШ ВРАЗЛИВИЙ у найближчій перспективі, ніж інші зобов'язувачі з нижчим рейтингом. Однак він стикається з серйозною невизначеністю та впливом несприятливих ділових, фінансових або економічних умов.
B1 B2 B3	B+ B B-	Позичальник БІЛЬШ ВРАЗЛИВИЙ, ніж "BB", але має здатність виконувати свої фінансові зобов'язання. Несприятливі ділові, фінансові або економічні розбіжності, ймовірно, погіршать його здатність або готовність виконувати фінансові зобов'язання.
Caa	CCC	Позичальник у даний час ВРАЗЛИВИЙ, і де-лежить на сприятливих ділових, фінансових та еко-номічних умовах для виконання своїх фінансових зобов'язань.
Ca	CC	Позичальник у даний час ДУЖЕ ВРАЗЛИВИЙ.
	C	Позичальник в даний час ДУЖЕ ВРАЗЛИВИЙ до неплатежів. Може бути використаний там, де було подано заяву про банкрутство.
C	D	Позичальник не сплатив одне або кілька своїх фінансових зобов'язань (оцінених або не оцінених), коли він став належним.

1.2 Огляд та аналіз наявних аналогів та публікацій

Існує величезна кількість робіт по визначенню ймовірності банкрутства підприємства та прогнозуванню кредитного рейтингу. Розглянемо деякі з найвпливовіших досліджень.

Спочатку вчені зосереджувалися на моделях лінійної дискримінації «здорових» компаній від тих, що згодом зазнають дефолту. Піонерами у цій галузі можна вважати Альтмана [2] і Олсона [19], які заснували традиційний ймовірнісний економетричний аналіз. Альтман визначив дискримінантну оцінку *Z*, яка залежить від кількох фінансових коефіцієнтів (оборотні кошти/загальні активи, нерозподілений прибуток/загальні активи тощо) для оцінки фінансового стану компанії. Олсон, з іншого боку, використовує лінійну логіт-модель регресії (LR), яка оцінює ймовірність банкрутства компанії. Деякі статті критикують ці методи як нездатні класифікувати компанії як життєздатні чи нежиттєздатні [5]. Однак у більшості літератури обидва підходи використовуються як еталон для оцінки більш складних методів.

Також було створено декілька інших структурних моделей для пояснення корпоративних дефолтів. Мертон [2] розробив міру «відстані до дефолту», використовуючи модель ціноутворення типу Блека-Скоулза. У недавньому дослідженні Джессен і Ландо [3] підтверджують, що модель відстані до дефолту може надійно визначити ризик корпоративного дефолту. Вони також представляють міру відстані до дефолту, коригуючи її за допомогою стохастичної волатильності вартості активів. Гловер [4] пропонує структурну модель для розрахунку очікуваних витрат корпорації у дефолті. Брогаард та ін. [5] використовуйте підхід відстані до дефолту, щоб показати, що підвищена ліквідність акцій знижує ризик корпоративного дефолту. Гіллегейст та ін. [6] стверджують, що ринкова міра, заснована на моделі Блека-Скоулза-Мертона, працює краще, ніж *Z*-оцінка Альтмана [7] та *O*-оцінка Олсона [19] при оцінці моделі дискретної небезпеки. Даффі та ін. [9] пропонують подвійно стохастичну модель для оцінки складних складових структури корпоративного ризику дефолту.

Однак останнім часом статистичні моделі та алгоритми машинного навчання, які передбачають ймовірність корпоративного дефолту, дають більш задовільні результати, ніж структурні моделі в цілому.

За часи дослідження моделей кредитного рейтингу було опубліковано велику кількість наукових робіт, присвячених методам машинного навчання [15, 17, 20]. Найуспішніші були засновані на деревах рішень [11, 14, 16, 23] та нейронних мережах [3, 6, 10, 18, 22]. Як правило, усі ці роботи використовують різні набори параметрів та функцій, залежно від вхідного набору даних. Барбоза та ін. [4] порівнюють такі методики із методом опорних векторів (SVM), показуючи, що методи ансамблю та випадкові ліси (Random Forest) дають найкращий результат.

Узагальнюючи аналіз наукових робіт, можна визначити три покоління основних статистичних моделей для прогнозування корпоративного дефолту (табл. 1.2). Таблиця узагальнює проведені дослідження, враховуючи розвиток кожного методу. Окремо варто відзначити, що кожний із цих методів використовується і донині.

Таблиця 1.2 – Класифікація статистичних моделей прогнозування дефолту

Дискримінантний аналіз	Бівер (1966); Альтман (1968); Меєр та ін. (2017)
Моделі бінарної відповіді	Ольсон (1980); Змієвський (1984); Форман (2003); Кемпбелл та ін. (2008); Кукук і Ренберг (2013); Аретц та ін. (2018)
Моделі небезпеки	Шамвей (2001); Чава і Джарроу (2004); Нам та ін. (2008); Бонфім (2009); Даковіч та ін. (2010); Дуан та ін. (2012); Фіглевський та ін. (2012); Тянь та ін. (2015); Трачинський (2017)

Очікується, що розвиток методологій машинного навчання ще більше прискорить інновації у фінансовому секторі, що призведе до появи нових фінансових послуг і зростання економіки даних на основі розподілу даних. Моделі кредитного рейтингу підтримують рішення про схвалення кредиту, призначаючи

кредитні ставки боржнику, що свідчитиме про його кредитоспроможність. Ці моделі є важливою частиною фінансової системи, оскільки забезпечують необхідну інфраструктуру для визначення кредитних ставок, що є драйвером кредитоспроможності.

1.3 Постановка задачі

Перед початком проектування системи необхідно проаналізувати існуючі аналоги реалізованих в даній сфері систем і сформулювати функціонал майбутньої системи. Далі слід розбити функціонал системи на окремі компоненти та визначити план майбутніх робіт і список реалізації необхідного функціоналу в системі.

Після проведення теоретичних досліджень була сформульована концепція завдання – реалізація підсистеми прийняття рішень для визначення кредитного рейтингу компанії.

Список необхідних кроків при вирішення завдань за допомогою штучних нейронних мереж можна сформулювати наступним чином:

– сформулювати задачу, що буде вирішуватись нейронною мережею: класифікація, прогнозування, кластеризація;

– визначити обмеження задачі (швидкість, точність відповіді);

– встановити вхідні данні (показники фінансової звітності та коефіцієнти) і вихідні дані (кількість класів);

– виконати попередню обробку даних: обробка пропущених значень, стандартизація, нормалізація викидів;

– з’ясувати оптимальну архітектуру нейронної мережі, потрібну для вирішення поставленої задачі.

Перед даною нейронною мережею ставиться задача – визначити кредитний рейтинг підприємства. А саме мова йде про задачу класифікації. Специфіка застосування нейронної мережі потребує певних обмежень для задачі, що розв’язується. Це швидкість відповіді – до 1 секунди, а також ступінь точності розпізнавання. Припускається точність не менше, ніж 90%. У якості вхідних даних

будуть числові значення на шкалі дійсних чисел R^+. У результаті будуть отримані класи кредитного рейтингу, які характеризуватимуть фінансовий стан та платоспроможність підприємства. Створена система потенційно полегшує та автоматизує роботу працівників фінансових установ, наприклад, банків та кредитних агентств, у процесі аналізу фінансової звітності підприємства та оцінки його платоспроможності. Система виконує поставлену задачу, використовуючи методи машинного навчання. Зокрема потрібно сконцентруватись на практичному застосуванні нейронних мереж задля вирішення описаної вище задачі.

Після детального аналізу предметної області, існуючих систем та проблем задачі класифікації підприємств був створений детальний план реалізації даного проекту. План проектування і розробки даної системи відноситься до ітеративних та включає в себе наступні етапи:

– проведення теоретичних досліджень методів машинного навчання у контексті вирішення задачі кредитного скорингу та прогнозування дефолту;

– виявлення та підготовка набору даних, що необхідні для навчання алгоритмів. Важливо також провести дослідження технік та методів попередньої обробки даних, які здатні позитивно вплинути на якість навчання алгоритмів;

– аналіз та відбір найбільш релевантних алгоритмів для класифікації підприємств. Необхідно сфокусуватись на застосуванні штучних нейронних мереж для вирішення зазначеної задачі;

– проведення необхідних експериментів та аналіз отриманих результатів;

– проведення порівняльного аналізу роботи різних алгоритмів.

Висновки до розділу 1

У даному розділі були розглянуті:

– питання актуальності проблеми кредитного скорингу підприємства;

– проаналізовано попередні дослідження та публікації, що використовують моделі машинного навчання для вирішення проблеми;

– поставлена задача кредитного скорингу.

2 МАТЕМАТИЧНІ МОДЕЛІ КРЕДИТНОГО СКОРИНГУ

2.1 Фінансові коефіцієнти

Упродовж десятиліть фінансові коефіцієнти вважались одним із найкращих інструментів оцінки кредитного рейтингу компаній [27, 18]. Зовсім недавно фокус змістився у бік можливостей прогнозування змін кредитного рейтингу, щоб отримати перевагу і використовувати інформацію, перш ніж вона буде доступна на ринку [22, 28].

Беручи до уваги наявну у цій роботі базу даних балансів 10 тисяч німецьких підприємств, розглянемо фінансові коефіцієнти, запропоновані у [22], зосереджуючись на тих, які можна напряму обчислити з балансу підприємства. Ці показники можна умовно розділити на дві групи:

– Фінансові коефіцієнти, які можна безпосередньо обчислити з балансу;
– Фінансові коефіцієнти, які НЕ можна безпосередньо обчислити з балансу.

Як буде показано далі, для розрахунку коефіцієнтів другої групи буде використано деякі наближення.

2.1.1 Коефіцієнт загальної ліквідності (Current Ratio)

Погашення короткострокових зобов'язань є одним з найважливіших аспектів, із якими стикається компанія під час ведення бізнесу. Здатність покриття короткострокових зобов’язань відображається, зокрема, коефіцієнтом загальної ліквідності (Current Ratio, CR), який іноді також називають співвідношенням «оборотного капіталу». Як правило, це співвідношення ліквідності стосується зобов'язань та кредитів, термін дії яких закінчується протягом фінансового періоду (1 рік). Інвестори потребують цієї інформації, оскільки компанія з дефіцитною ліквідністю є дуже ризикованою інвестицією, де капітали можуть бути використані для короткострокових потреб, а не для довгострокових. CR обчислюється наступним чином:

$$CR = \frac{\text{Оборотні активи}}{\text{Поточні зобов}'\text{язання}} \tag{2.1}$$

Оборотні активи та поточні зобов'язання присутні у балансі будь-якого підприємства: перші включають грошові кошти та їх еквіваленти, дебіторську заборгованість та запаси. Дебіторська заборгованість та запаси повинні бути конвертовані у грошові кошти протягом одного фінансового року, або ж вони повинні бути перераховані на інший рахунок. Поточні зобов'язання, з іншого боку, враховують всі короткострокові зобов'язання, такі як кредиторська заборгованість, заробітна плата та податки, а також розстрочка довгострокового боргу, що підлягає сплаті протягом фінансового року.

Як і більшість фінансових коефіцієнтів, CR для певної компанії слід порівнювати з іншими підприємствами тієї ж галузі: значення, подібні або трохи вище середнього, вважаються позитивним показником, тоді як нижчі за середні можуть бути симптомом фінансового розладу. При цьому дуже високі значення CR не обов'язково є позитивним фактором, так як можуть свідчити про те, що керівництво компанії не використовує ресурси ефективно.

Однак, не варто однозначно покладатись на значення CR для оцінки ліквідності компанії. Зокрема, важко порівняти показник для різних галузей промисловості, і це також поширюється на великі групи, що здійснюють кілька різноманітних видів діяльності. Більш того, як і всі інші коефіцієнти, які ми розглянемо, CR слід оцінювати разом з інформацією попередніх років, щоб екстраполювати тенденцію.

2.1.2 Коефіцієнт швидкої ліквідності (Quick Ratio)

Короткострокова ліквідність компанії також оцінюється коефіцієнтом швидкої ліквідності (Quick Ratio, QR), також відомим як "Кислотний тест". Коефіцієнт швидкої ліквідності використовується для оцінки здатності компанії погасити свої фінансові зобов'язання, що підлягають сплаті у фінансовому періоді,

використовуючи лише ліквідні активи, такі як грошові кошти та дебіторська заборгованість, без продажу її запасів або використання інших джерел фінансування.

Що відрізняє QR від CR, так це те, що розглядаються тільки майже грошові активи, тобто активи, які теоретично можуть бути перетворені в готівку на місці. Це робить його більш консервативним показником. Таким чином, у наступному рівнянні запаси не розглядається для його обчислення:

$$QR = \frac{CE + MS + AR}{CL} = \frac{CA - I - PE}{CL} \tag{2.2}$$

де CE – грошові кошти та їх еквіваленти (Cash & Equivalents),

MS – ринкові цінні папери (Marketable Securities),

AR – дебіторська заборгованість (Accounts Receivable),

CL – поточні зобов'язання (Current Liabilities),

CA – оборотні активи (Current Assets),

I – запаси (Inventory),

PE – передплачені витрати (Prepaid Expenses).

Як і CR, QR слід порівнювати з середнім показником по галузі, з низьким рівнем, пов'язаним з фінансовими труднощами, і високим рівнем з високою ліквідністю. Однак ідеальне значення коефіцієнту швидкої ліквідності повинно дати результат $QR = 1$, що означає, що підприємство запланувало свої ліквідні ресурси, щоб відповідати саме короткостроковим зобов'язання, які вони повинні погасити протягом фінансового року. Значення QR нижче такого порогу можуть турбувати керівництво та інвесторів, оскільки вони сигналізують про деяку потенційну проблему ліквідності.

Запаси та інші оборотні активи у QR не враховуються, так як не завжди вдається відразу їх ліквідувати, хоча ступінь складності у цьому значно залежить від бізнесу: у цілому потрібен деякий час, щоб перетворити запаси у гроші.

Нарешті, ще одним упущенням є передплачені витрати, оскільки вони не можуть бути використані для погашення будь-яких фінансових зобов'язань.

2.1.3 Коефіцієнт абсолютної ліквідності (Cash Ratio)

Останнім показником, який вимірює ліквідність підприємства, є коефіцієнт абсолютної ліквідності (Cash Ratio). Він є більш консервативним, ніж QR, оскільки розглядає лише грошові кошти та грошові еквіваленти у чисельнику. До них відносяться кошти на поточних рахунках, ощадні рахунки та казначейські векселі, які також іноді називають ринковими цінними паперами.

Цей показник дуже схожий за своєю природою на CR і QR, і можна розглядати його як найгірший сценарій для CR, у якому запаси та дебіторська заборгованість втрачають усю свою цінність, так, ніби компанія перебуває на межі свого закриття. Формула є наступною (використовуючи абревіатури, визначені вище):

$$Cash\ Ratio = \frac{CE}{CL} \tag{2.3}$$

У ситуації банкрутства коефіцієнт абсолютної ліквідності повідомляє кредиторам, який відсоток від поточних зобов'язань компанії може бути покритий її найбільш ліквідними активами, оскільки передбачається, що продаж інших оборотних активів, таких як запаси та дебіторська заборгованість, навряд чи принесе їх номінальну вартість.

2.1.4 Відношення боргу до власного капіталу (Debt-to-Equity Ratio)

Співвідношення боргу до капіталу (Debt-to-Equity Ratio, D/E) є одним з найвідоміших фінансових інструментів, який виражає баланс між довгостроковими зобов'язаннями та власним капіталом акціонерів, таким чином вимірюючи фінансовий важель компанії. Його формальне визначення відображається наступною формулою:

$$D/E = \frac{\text{Довгострокові борги } + \text{Короткострокові борги}}{\text{Власний капітал}} \quad (2.4)$$

Співвідношення боргу до власного капіталу впливає на вартість капіталу для фірми, і додатково вказує, яка частина капіталу акціонерів повинна бути використана для покриття всіх непогашених боргів, у разі закриття бізнесу, або, як альтернатива, яка частка боргу теоретично може бути стягнута з власного капіталу компанії.

Точні рахунки, що використовуються для побудови чисельника та знаменника у формулі (2.4), можуть відхилятися від традиційного бухгалтерського визначення «боргу» або «власного капіталу», оскільки існують кілька елементів, які можуть призвести до спотворення D/E: тому розрахунок важеля компанії, як правило, не є прямим, і повинен становити кілька елементів, які можуть призвести до спотворення D/E: тому розрахунок важеля компанії, як правило, не є прямим, і повинен становити кілька елементів, які можуть призвести до спотворення D/E: тому розрахунок важеля компанії, як правило, не є прямим, і повинен враховувати кілька відмінностей з точки зору нематеріальних активів, нерозподіленого прибутку/збитків та коригування пенсійного плану.

Більш того, не завжди існує домовленість між аналітиками у визначенні боргу. Прикладом можуть служити привілейовані акції, які часто класифікується як власний капітал, але містять елементи, які роблять їх ближчими до боргу (привілейовані дивіденди та права на ліквідацію). У певних ситуаціях віднесення їх до одної категорії, а не до іншої, може призвести до значного впливу на значення фінансового важеля.

Доцільність такого низькорівневого аналізу залежить від мети обчислення: у нашому випадку розглядається великий набір даних компаній, так що індивідуальний аналіз балансу кожного підприємства не можливо виконати, тому результат рівняння (2.4) вважається достовірним.

Загалом, хоча більш високе значення важеля може становити ризик як для акціонерів, так і для власників боргів, дуже важко порівняти співвідношення D/E серед підприємств, що працюють у різних галузях промисловості, оскільки капітал потребує змін залежно від типу бізнесу та темпів зростання: галузі із високою капіталомісткістю та повільним зростанням, такі як автомобільний сектор або комунальні послуги, зазвичай демонструють високий D/E, тоді як послуги та високотехнологічні сектори мають набагато нижчий рівень.

Не всі фінансові коефіцієнти можна обчислити безпосередньо з даних, що входять до балансу. Фактично, кілька показників вимагають такої інформації, як доходи, вартість проданих товарів, чистий прибуток та сплачені фінансові відсотки, які містяться у Звіті про прибутки та збитки (Форма 2). Більшість цих рахунків не можна апроксимувати за наявною інформацією, таким чином, корисні коефіцієнти, такі як коефіцієнт покриття відсотків, коефіцієнт обороту запасів, операційний цикл та маржа операційного прибутку не можуть бути використані у цій роботі.

Однак, деяка ключова інформація може бути наближено розрахована шляхом порівняння балансу за два послідовні роки. При сильному припущенні, що всі прибутки/збитки залишаються у компанії, без виплат дивідендів акціонерам, не відбувається збільшення капіталу. Таке припущення, хоча і сильне, не може бути використано для компаній, що котуються на фондовій біржі, через вплив коливань цін на акції. У даній роботі розглядаються тільки підприємства, які не мають законодавчого обов’язку щодо публікації звіту про прибутки та збитки, і тому, за визначенням, не можуть котуватись на фондовій біржі.

Ідея апроксимації полягає у тому, що різниця між загальним акціонерним капіталом (Total Shareholders’ Equity) протягом двох послідовних років дорівнює чистому прибутку компанії за перший рік:

$$NetIncome(0) \approx TotalEquity(1) - TotalEquity(0) \qquad (2.5)$$

де *NetIncome(0)* – апроксимація чистого прибутку підприємства за перший рік,

TotalEquity(0), TotalEquity(1) – власний капітал за перший та другий роки.

Використовуючи апроксимацію (2.5), можна розрахувати наближені значення двох важливих коефіцієнтів: рентабельності активів і рентабельності власного капіталу, які ми проаналізуємо нижче.

2.1.5 Рентабельність активів (Return on Assets)

Метою бізнесу є отримання цінності зі своїх ресурсів, тобто – активів. Таким чином, ключовим показником ефективності діяльності компанії є оцінка того, скільки прибутку створюється її активами, звідси і назва – рентабельність активів (Return on Assets, ROA), визначається наступним чином:

$$ROA = \frac{\text{Чистий прибуток}}{\text{Активи}} \tag{2.6}$$

Інвестори отримують від ROA кількісну міру того, наскільки ефективно інвестиції перетворюються у готівку. На відміну від коефіцієнтів, розглянуто раніше, теоретично ROA не має ніяких недоліків, оскільки збільшення показника просто вказує на те, що бізнес організований ефективним чином, тоді як високий CR або QR може означати, що ресурси використовуються неоптимально.

Як і інші показники, які розглядались до цього, ROA не може бути оцінений як окреме значення. Його слід порівнювати зі значеннями попередніх років, або із середнім показником компаній тієї ж галузі. Фінансові послуги або капіталомісткі обробні галузі вимагають набагато більшої кількості активів у своїх виробничих циклах, порівняно з підприємствами, що працюють у сфері послуг. З цієї причини ROA недоцільно порівнювати у цих галузях промисловості. Однак, зосереджуючись на підвищенні ефективності бізнесу, оцінка річних змін у ROA є одним із найкращих показників динаміки.

ROA не містить ніякої інформації про фінансову структуру компанії, оскільки показник не розрізняє борг і власний капітал. Однак, обидва джерела

використовуються для фінансування бізнес-операцій. Це означає, що рентабельність активів підприємства, яке в основному залежить від боргу, може бути значно меншою, ніж у іншого підприємства, яке використовує лише власний капітал, навіть якщо останнє працює менш ефективно. Це відбувається, оскільки процентні витрати віднімаються від доходів при розрахунку чистого прибутку.

Для врахування цього, у деяких випадках вартість відсотків, що використовуються для придбання активів, додається до чисельника, включеного у формулу ROA, щоб дати більш об'єктивний портрет ефективності бізнесу, що є кінцевою метою індикатора ROA.

2.1.6 Рентабельність власного капіталу (Return on Equity)

Останнім показником є рентабельність власного капіталу (Return on Equity, ROE). Він визначається як співвідношення між чистим прибутком і власним капіталом (як правило, виражається у відсотках):

$$ROE = \frac{\text{Чистий прибуток}}{\text{Власний капітал}} \tag{2.7}$$

Слід зазначити, що, як і ROA, такий показник має сенс, якщо і чисельник, і знаменник є додатними значеннями у (2.7), інакше цей показник може ввести в оману.

Для публічних компаній розглянутий чистий прибуток не повинен включати витрати на виплати дивідендів по звичайним акціям, але має включати витрати на дивіденди по привілейованим акціям та процентні витрати, виплачені кредиторам. У нашому випадку, маючи справу з приватними малими компаніями, цей аспект не представляє проблему, і можна використовувати наближення (2.5).

Усі міркування, висловлені для попередніх коефіцієнтів, з точки зору обмежень цього показника, дійсні і для ROE: співвідношення не містить абсолютної інформації саме по собі, але повинно порівнюватися з попередніми роками, і з іншими компаніями у тій же галузі.

У якості міри ефективності бізнесу, ROE має один істотний недолік, який необхідно розглянути. Реалістичним способом значного підвищення значення ROE є випуск довгострокового боргу, який потім буде використано для створення бази активів та збільшення виручки бізнесу. Це є поширеним способом зростання бізнесу, але в той же час збільшує ризик для акціонерів за рахунок збільшення фінансових важелів. З цієї точки зору, ROE може розглядатись як менш надійний показник для оцінки оперативної ефективності, ніж ROA. Для великих публічних акціонерних компаній, цільовий ROE відповідає ринковому орієнтиру, наприклад, S&P500 визначає ≈14%, та значення ROE нижче 10% вважаються незадовільними для інвесторів.

Прибуток після виплат відсотків та податків збігається з визначенням чистого прибутку протягом фінансового періоду, у той час як власний капітал – це середній власний капітал акціонерів, розрахований шляхом усереднення значень на початку і в кінці того ж періоду. Використання середнього капіталу для розрахунку ROE застосовується для врахування різних часових ліній балансу (фото ситуації на кінець року) та звіту про прибутки та збитки (представлення одного року фінансово-економічних операцій). З огляду на це визначення, з точки зору інвестора, ROE – найкращий показник для оцінки потенційних дій з купівлі/продажу акцій компанії, у випадку публічних компаній.

Незважаючи на ці міркування, ROE може бути небезпечним інструментом, якщо його використовувати поза контекстом. Фактично, у більшості ситуацій перевагу віддають компанії із середньою, але стабільною рентабельністю капіталу, ніж компанії з рентабельністю капіталу, яка демонструє екстремальні стрибки. Причина криється в динаміці між чисельником і знаменником у (2.7): якщо бізнес переживає негативний період, він може накопичувати операційні збитки та утримувати їх, зменшуючи загальний власний капітал з часом. У цих умовах раптове збільшення чистого прибутку, можливо, через інвестиції, сплачені за рахунок нових випущених боргових зобов’язань, може призвести до величезного сплеску ROE, також завдяки зменшеному знаменнику. Однак, з точки зору

інвестора, що грає у довгу, краще було б володіти акціями компанії з постійною позитивною рентабельністю капіталу, хоча і з «меншими максимумами».

Підводячи підсумки, можна сказати, що рентабельність капіталу є одним із найцікавіших показників, але в той же час має ряд недоліків. Можна бути достатньо впевненим, що, якщо ROE є негативним або має надмірно високі значення, слід провести більш глибокий аналіз, щоб визначити причини цієї поведінки. Вони можуть значно відрізнятися: програма зворотного викупу акцій може призвести до негативної рентабельності капіталу навіть у корпораціях з оптимальним управлінням. При будь-яких умовах, слід уникати використання ROE для аналізу бізнесу, коли показники чисельника або знаменника відображають негативні значення. За таких обставин слід враховувати інші показники.

2.2 Дискримінантний аналіз

ВПіонерами цих досліджень були, наприклад, Бевер [1] та Альтман [7]. Ці дослідження будують моделі, які використовують деяку дискримінантну функцію, що розраховує кредитні рейтинги. Відома модель Z-оцінки Альтмана є одним із прикладів, наступні дослідження також використовують цю методологію [14,15].

Дискримінантний аналіз вибирає змінні, які найкраще можуть визначити, чи є компанія банкрутом, і обчислює дискримінантну функцію як лінійну комбінацію цих змінних:

$$D = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \tag{2.8}$$

де D – дискримінантна оцінка, обчислена дискримінантною функцією,

β_0 – константа,

β_m – оцінений коефіцієнт,

X_m – пояснювальна змінна.

Спостереження класифікується як нормальне, якщо оцінка дискримінації нижче певного порогу і класифікується як дефолтне, якщо оцінка вище цього порогу. Однак, дискримінантний аналіз вимагає припущення, що незалежні змінні

слідують багатовимірним нормальним розподілам, матриця коваріації між двома групами визначається як нормальна, а дефолтні значення – ідентичні. Дискримінантний аналіз має перевагу, оскільки дозволяє упорядкувати корпорації відповідно до їх ступеня ризику дефолту.

Найпоширенішими моделями є такі економетричні моделі: двохфакторна модель оцінки ймовірності банкрутства, оцінка ймовірності банкрутства на основі Z-оцінки Альтмана, модель Конана і Гольдера, модель Спрінгейта, модель Фулмера[15, с. 69].

Модель Альтмана є однією з найперших моделей аналізу ймовірності банкрутства підприємства, яка ґрунтувалася на основі дискримінантного аналізу. Дана модель представляє собою алгоритм оцінки загрози банкрутства неплатоспроможного підприємства, заснований на комплексному обліку найважливіших показників, що допомагають виявити кризовий фінансовий стан підприємства: рентабельність, леверидж, ліквідність, платоспроможність.

Розрахунок відбувається таким чином:

$$Z = 1{,}2 \times X_1 + 1{,}4 \times X_2 + 3{,}3 \times X_3 + 0{,}6 \times X_4 + 1 \times X_5 \qquad (2.9)$$

де Z – інтегральний показник рівня загрози банкрутства;

X_1 – власний оборотний капітал/всього активів;

X_2 – нерозподілений прибуток/всього активів;

X_3 – прибуток до виплати відсотків/всього активів;

X_4 – власний капітал/зобов’язання;

X_5 – чиста виручка/всього активів.

Значення ймовірностей банкрутства за моделлю Альтмана зображені у вигляді табл. 2.1.

Таблиця 2.1 – Ймовірність банкрутства за моделлю Альтмана

Значення Z	Приблизна ймовірність банкрутства
$Z < 1{,}81$	80-100%
$1{,}81 \leq Z < 2{,}77$	35-50%
$2{,}77 < Z < 2{,}99$	15-20%
$Z \geq 3$	0

Модель М. Спрінгейта передбачає розвиток інтегрального показника, за допомогою якого здійснюється оцінка ймовірності визнання підприємства банкрутом. Важливим є те, що точність даної моделі є досить високою і становить більше 90%. Формула для розрахунку є наступною:

$$Z = 1{,}03 \times a + 3{,}07 \times b + 0{,}66 \times c + 0{,}4 \times d \qquad (2.10)$$

де Z – інтегральний показник рівня загрози банкрутства;

a – робочий капітал / загальна вартість активів;

b – прибуток до сплати податків та відсотків / загальна вартість активів;

c – прибуток до сплати податків / короткострокові зобов'язання;

d – обсяг продажу / загальна вартість активів.

Таблиця 2.2 – Ймовірність банкрутства згідно моделі Спрінгейта

Значення *Z*	Ймовірність банкрутства
< 0,862	висока
> 0,862	низька

2.3 Моделі бінарної відповіді

Друге покоління моделей корпоративних прогнозів дефолтів використовує моделі бінарної відповіді. Двійкова модель відповіді – це модель, яка визначає стан корпорації як нормальний (= 0), або дефолтний (= 1). Вона оцінює ймовірність дефолту за допомогою пояснювальних змінних і застосовує логістичну або пробітну функцію для більшості випадків. Репрезентативним прикладом є *O*-оцінка з логістичною функцією, введена Олсоном [8], тоді як Змієвський [17] тестує корпоративний ризик дефолту за допомогою моделі пробіт (одиниці ймовірності).

Ці бінарні моделі дозволяють розрахувати ймовірність банкрутства протягом наступного періоду. Як пояснюють Форман [18] та Чарітоу та ін. [16], бінарна модель відповіді визначає ймовірність дефолту компанії *n* наступним чином:

$$P_n(y_n = 1) = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 X_{1,n} + \beta_2 X_{2,n} + \cdots + \beta_m X_{m,n})\right]} \quad (2.11)$$

де $y_n = 1$, якщо корпорація n стане банкрутом і 0 у іншому випадку,

$P_n(y_n = 1)$ – ймовірність дефолту для корпорації n,

$\beta_1, \beta_2, \ldots, \beta_м$ – кутові коефіцієнти,

$X_{1,n}, X_{2,n}, \ldots, X_{m,n}$ – предиктори для корпорації n.

Функція ймовірності задається за допомогою:

$$L = \prod_{i=1}^{N} F(\beta', X_i)^{y_i} \left(1 - F(\beta' X_i)\right)^{1-y_i} \quad (2.12)$$

де $F(\beta', X_i)$ визначається за допомогою рівняння (2.11).

Модель бінарної відповіді має ряд переваг перед дискримінантним аналізом для прогнозування кредитного рейтингу. По-перше, така модель не вимагає будь-яких припущень щодо ймовірності дефолту або розподілів змінних-предикторів. По-друге, вона може перевірити значимість окремих незалежних змінних. Нарешті, таку модель можна використати для розрахунку ймовірності дефолту у наступному періоді. Кемпбелл та інші [19] розширюють бінарну модель відповіді, досліджуючи корпоративні дефолти, використовуючи багатофакторну logit модель. Такий підхід дозволяє розрахувати корпоративні ймовірності дефолту і передбачити ризик дефолту протягом декількох періодів. Еретц та ін. [20] використали методологію Кемпбелла [19] та дані для неамериканських компаній для виявлення корпоративного дефолту. Кукук і Реннберг [21] розширили моделі бінарної відповіді, запропонувавши змішану модель logit, яка дозволяє стохастичні параметри та нелінійність у регресорних змінних. Бонфім [22] перевіряє макроекономічні та фінансові дані, використовуючи пробітну модель та стверджує, що корпоративні дефолти обумовлені кількома специфічними факторами фірми.

Однак результати також показують, що макроекономічні фактори також важливі для оцінки ризику дефолту з плином часу.

2.4 Методи машинного навчання

Самуїл [32] пропонує концепцію машинного навчання і визначає її як «дисципліну, яка дає комп'ютерам можливість вчитися без чіткої програми». Мітчелл [33] далі розвиває це поняття машинного навчання: «Кажуть, що комп’ютерна програма вчиться на досвіді *E* щодо певного класу завдань *T* і показників продуктивності *P*, якщо її якість при виконанні завдань у *T*, виміряних *P*, покращується з досвідом». У зв'язку з цим, алгоритм машинного навчання для прогнозування кредитного рейтингу можна охарактеризувати як серію процесів, що покращують показник кредитного рейтингу (*P*) для виконання завдання прогнозування корпоративного кредитного ризику (*T*) з використанням фактичної корпоративної фінансово-кредитної інформації (*E*).

Дослідження прогнозування кредитного рейтингу з використанням методів машинного навчання проводяться різними способами, особливо в галузі інформатики, і Барбоза та ін. [34] стверджують, що моделі машинного навчання демонструють кращі показники у прогнозуванні корпоративного кредитного рейтингу та ймовірності дефолту. Таблиця 2.3 показує три важливі методології та їх репрезентативні дослідження. У більшості випадків передбачення дефолтів із використанням машинного навчання приймає проблему класифікації, яка класифікує компанію як таку, що знаходиться в одному з двох. або більше станів, визначених як нормальні (= 0) і дефолтні (= 1), і обчислює ймовірність знаходження компанії у конкретному стані. Таким чином, в основному використовуються алгоритми машинного навчання для вирішення задач класифікації. Репрезентативні приклади включають SVM, дерева рішень та алгоритми штучних нейронних мереж.

Таблиця 2.3 – Класифікація методів машинного навчання для визначення кредитного рейтингу підприємства

Метод опорних векторів (SVM)	Шинь та ін. (2005); Чень (2011); Лу та ін. (2014)
Дерева рішень	Олсон та ін. (2012); Цай та ін. (2014)
Штучні нейронні мережі	Ян, Платт і Платт (1999); Фалавінья (2012); Ітурріага і Санс (2015); Азаїт і Аххаб (2016)

2.4.1 Метод опорних векторів

Алгоритм методу опорних векторів (SVM) широко використовується у різних областях, включаючи прогнозування корпоративного дефолту. Цей алгоритм використовує гіперплощину для класифікації n спостережень, кожне з яких має p властивостей. Кожне спостереження можна класифікувати як таке, що має один зі статусів, визначених як $y_i \in \{-1,1\}$, де y_i представляє статус і-го спостереження. Алгоритм SVM повинен визначити гіперплощину, що розділяє спостереження, і це дозволяє деякі неправильні класифікації, щоб уникнути перенавчання. Оптимальна гіперплощина може бути представлена наступними рівняннями:

$$\max_{\beta_0,\beta_1,\dots,\beta_p,\varepsilon_0,\varepsilon_1,\dots,\varepsilon_n} M$$

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) \geq M(1 - \varepsilon_i)$$

$$\varepsilon_i \geq 0$$

$$\sum_{i=1}^{n} \varepsilon_i \leq C$$

де $\beta_0, \beta_1, \dots, \beta_p$ – параметри гіперплощини,

$\varepsilon_0, \varepsilon_1, \dots, \varepsilon_n$ – слабкі змінні, що дозволяють деякі неправильно класифіковані спостереження,

M – відстань між розділюючою гіперплощиною та спостереженнями,

C – параметр, який визначає межі помилок класифікації.

Шин та ін. [36] застосовують SVM до прогнозування корпоративного дефолту та показують, що SVM працює краще, ніж модель нейронної мережі зі зворотнім поширенням помилки. Чень [37] порівнює кілька моделей прогнозування і стверджує, що SVM має високу точність і добре працює для короткострокових і довгострокових прогнозів дефолту. Ліанг та ін. [38] також показують, що SVM дає найкращі результати при прогнозуванні банкрутства з використанням фінансових коефіцієнтів та показників корпоративного управління. Лу та ін. [39] розширюють методологію SVM, використовуючи оптимізацію рою гібридних комутаційних частинок.

2.4.2 Дерева рішень та випадкові ліси

Алгоритм дерева рішень (Decision Tree, DT) – це методологія для вирішення задач регресії або класифікації шляхом побудови правил прийняття рішень у структурі дерева. Алгоритм DT ділить простір ознак, що складається з комбінації p пояснювальних змінних $X_1, X_2, \dots, X_p$, у J областей $R_1, R_2, \dots, R_J$, що не перетинаються, і він надає однаковий прогноз для спостережень, що належать до одної області. Якість такого розподілу вимірюється за допомогою індексу Джині:

$$G = \sum_{j=1}^{J} \hat{p}_{mj}(1 - \hat{p}_{mj}) \tag{2.13}$$

де J – кількість станів, а $\hat{p}_{mj}$ – частка стану j у регіоні R_m.

Процес обрізки, який встановлює окремі дерева прийняття рішень, виконується за допомогою індексу Джині. Алгоритм DT має ту перевагу, що

модель інтуїтивно зрозуміла і проста в інтерпретації, але вона має дуже високу ймовірність перенавчання, що може виникнути у процесі поділу простору функцій або створення гілок, і, в результаті, знижується точність прогнозу.

Алгоритм випадкового лісу (RF) – це алгоритм машинного навчання, який використовує декілька дерев рішень. Алгоритм RF вибирає заздалегідь визначену кількість предикторів шляхом рандомізації, коли він створює нове дерево рішень. Як правило, кількість вибраних змінних задається квадратним коренем p від загальної кількості пояснюючих змінних. Якщо позначити це число як k, то алгоритм RF генерує кілька дерев рішень із k випадково вибраними пояснюючими змінними. Для задачі класифікації прогноз цієї моделі базується на найбільш часто прогнозованому результаті у всіх деревах рішень.

Олсон та ін. [10] стверджують, що алгоритм DT є більш зрозумілим і точним, ніж інші алгоритми машинного навчання. Цай та ін. [14] порівнюють ансамблі дерев рішень із SVM і багатошаровим нейронними мережами і стверджують, що ансамблі дерев рішень працюють найкраще. Ціеба та ін. [12] розширюють алгоритми DT за допомогою генерації синтетичних функцій та градієнтного прискорення. Джардін [13], зокрема, пропонує корпоративну модель прогнозування дефолтів із ансамблями карт Кохена і стверджує, що цей підхід є дуже ефективним.

2.4.3 Штучні нейронні мережі

Штучна нейронна мережа – це алгоритм машинного навчання, розроблений відповідно до принципів роботи справжнього людського мозку. Цей алгоритм вирішує складні завдання, імітуючи структуру мозку і з'єднуючи штучні нейрони за допомогою простих структур. Нейрони, які є основними одиницями, що складають головний і спинний мозок людини, відповідають за передачу сигналів, які вони отримують, іншим пов'язаним нейронам. Нейрони сигналізують іншим нейронам тільки тоді, коли інтенсивність отриманого сигналу вище певного порогу.

Штучні нейрони імітують фактичні нейрони за допомогою математичних моделей. Кожен штучний нейрон отримує кілька сигналів, x_1, x_2, …, x_j, що складається з нулів і одиниць, і він обчислює зважену суму сигналів, які він отримує відповідно до їх ваг, w_1, w_2, …, w_j. Залежно від моделі можуть надходити сигнали, що належать до множини $(-\infty, \infty)$ або множини $(0, \infty)$. Потім сигнал передається на наступний штучний нейрон тільки тоді, коли зважена сума отриманих сигналів перевищує певну інтенсивність або поріг.

Серед різних можливих структур штучних нейронних мереж модель з декількома прихованими шарами називається глибокою нейронною мережею. Термін «глибоке навчання» відноситься до використання таких глибоких нейронних мереж для машинного навчання. Такий підхід покликаний вирішити проблему локальних мінімумів, що виникає в існуючій методології штучної нейронної мережі. Він характеризується безперервним використанням отриманих вхідних даних для покращення прогнозної здатності мережі.

Прогнози кредитного рейтингу та дефолтів з використанням штучних нейронних почали використатись після 1990-х років. Ян та інші [14] досліджують кілька алгоритмів і показують, що дискримінантний аналіз Фішера та ймовірнісні нейронні мережі мають найкращу якість прогнозування. Останнім часом, після новаторської роботи Хінтона, Осіндеро і Теха [15], штучні нейронні мережі знову з'являються як методологія прогнозу кредитного рейтингу. Фалавінья [16] прогнозує ризики дефолтів невеликих італійських компаній в умовах невизначенності за допомогою алгоритму штучної нейронної мережі. Лопес Ітурріага і Санс [17] оцінюють і візуалізують ризики дефолту банків, поєднуючи багатошарові перцептрони і самоорганізовані карти Кохена. Азейт і Ачча [18] покращують модель прогнозування нейронної мережі, включаючи дискримінантні змінні. Генг та ін. [19] показують, що підхід нейронної мережі має кращу продуктивність, ніж інші алгоритми класифікації. Крім того, багато досліджень намагалися поліпшити якість прогнозування, змінюючи алгоритми нейронної мережі.

Протягом останніх років глибоке навчання використовувало згорткові нейронної мережі (CNN) і рекурентні нейронні мережі (RNN) для вирішення проблеми перенавчання, яка виникає у процесі тренування моделей. Алгоритми CNN ділять інформацію для навчання на кілька доменів і аналізують найбільш релевантні домени, використовуючи обмежену інформацію. Таким чином, ці алгоритми добре працюють в області розпізнавання зображень і тому широко використовуються. Алгоритми RNN – це спосіб обробки та послідовного аналізу даних, і вони використовуються для прогнозування даних часових рядів та розпізнавання текстових даних. У даний час глибоке навчання використовується для розуміння контексту інформації та виконання різних завдань, таких як розпізнавання образів, обробка природної мови та автономне водіння.

Висновки до розділу 2

У даному розділі було розглянуто математичні методи та моделі кредитного скорингу підприємства. Було охарактеризовано ключові фінансові коефіцієнти, які використовуються у даній роботі для тренування моделей, та визначено їх переваги та недоліки. Додатково було проаналізовано еволюцію наукової думки у галузі прогнозування кредитного рейтингу та визначення ймовірності дефолту компанії. Зокрема, розглянуто ефективність моделей дискримінантного аналізу, бінарної відповіді та методів машинного навчання для вирішення поставленої задачі.

3 ПРОЄКТУВАННЯ СИСТЕМИ КРЕДИТНОГО СКОРИНГУ

У даній роботі розглядається задача визначення кредитного рейтингу підприємств малого та середнього бізнесу в умовах недостатньої інформації, коли пропущені деякі показники балансу. Для визначення найбільш оптимального алгоритму класифікації, на тестовій базі даних 10 тисяч німецьких компаній, було перевірено роботу декількох моделей.

Зокрема, було використано два методи: спочатку було застосовано чистий підхід «чорного ящика», заснований на штучній нейронній мережі (ШНМ). Нейронні мережі – це потужний інструмент, який використовується у різних областях для пояснення складних явищ. Окрім цього, їх архітектура дозволяє навчати ШНМ на базах даних із пропусками інформації.

Другий підхід є «сірим ящиком» і вимагає використання фінансових коефіцієнтів. У випадку, коли немає Звіту про прибутки та збитки підприємства, деякі фінансові коефіцієнти були замінені апроксимаціями, які оцінюються зі Звіту про фінансовий стан (Форма 1). У цьому підході база даних ділиться на дві групи: записи з усіма доступними фінансовими коефіцієнтами та такі, де деякі дані відсутні. Потім використовується техніка випадкового лісу (Random Forest) для класифікації, де класифікатор навчається за допомогою даних із усіма доступними полями. Після цього використовується метод максимізації очікувань (Expectation Maximization, EM) для заповнення відсутніх полів у базі даних. Цей метод використовує припущення, що всі записи є незалежними величинами, отриманими із багатовимірного нормального розподілу. Успіх у процесі вставки пропущених даних за допомогою ітераційного алгоритму EM не гарантується, головним чином через відсутність апріорних знань про розподіл даних. Тим не менш, як буде показано далі, введений набір даних добре працює із класифікатором Random Forest, навченим на повній вибірці даних.

Ідея порівняння між цими двома підходами («чорний ящик» ANN і «сірий ящик» EM/Random Forest) полягає у тому, що ANN використовується як

інструмент порівняльного аналізу для класифікатора на основі фінансових коефіцієнтів. Насправді, для статичної бази даних, яка розглядається у даній роботі, належним чином навчена нейронна мережа із лінійними шарами повинна повністю пояснити зв'язок вихідної змінної (кредитного рейтингу) із вхідними ознаками на основі існуючої кореляції (тобто залишки не повинні містити інформацію). Інакше кажучи, ANN повинна охопити всю інформацію, яка присутня у базі даних. Таким чином, якщо підхід EM/Random Forest має порівнянні результати, можна обґрунтовано стверджувати, що вибір фінансових коефіцієнтів, і, перш за все, наближень, був виправданим.

3.1 Підготовка та аналіз набору даних

3.1.1 Характеристика структури вихідного набору даних

Для успішного створення та використання моделей кредитного скорингу необхідні великі обсяги історичних даних, включаючи такі основи, як фінансові коефіцієнти та вся інформація про дефолти підприємства [22]. Крім того, ці дані, як правило, кластеризуються по галузям, у яких працюють компанії, оскільки динаміка варіюється від одного сектора економіки до іншого. Існують спеціалізовані фірми з екстраполяції такого роду даних і продажу їх зацікавленим сторонам (як правило, зацікавленим сторонам конкретної компанії). Однак, у деяких випадках, великі компанії створюють та розвивають власні моделі прогнозування ймовірності дефолту.

Дані, що використовуються для прогнозування дефолту, випливають із фінансової звітності, яку компанії зобов'язані регулярно публікувати. За допомогою цих даних обчислюються фінансові коефіцієнти, що описують обсяг, фінансові важелі, ліквідність, прибутковість та різні показники ефективності роботи підприємства. Такі коефіцієнти вважаються ключовими факторами для визначення ймовірності дефолту компанії. Доповнюючи цю інформацію історією компаній, які досягли дефолту, можна побудувати статистичні регресійні моделі та

використати їх для визначення компаній, які ймовірно також стануть дефолтними. Такі регресійні моделі дають відповідь з точки зору ймовірності, яка потім може бути перетворена на шкалу кредитного рейтингу, яку використовують рейтингові агентства [8].

Як і всі моделі на основі даних, PD моделі надійні на стільки, наскільки якісними є історичні данні, використані для їх створення. Теоретично, фірми слід порівнювати тільки в одній галузі, та протягом нетривалого періоду часу (через зміну динаміки ринку), однак таким чином інформаційний пул скоротиться, що ускладнить екстраполяцію моделі, або ж навіть унеможливіть її. Якість даних у цілому є проблемою, оскільки наявні фінансові звіти містять неправильні записи, а також відсутні поля. У більшості випадків це спричинено різними правилами звітності у різних країнах, що дозволяє компаніям не розкривати частину своєї фінансової звітності. Крім того, навіть інформація про дефолтні компанії може бути зміщеною у часі або не надаватись взагалі через тривалі судові позови або поглинання [42].

У даній роботі було застосовано набір даних, наданий Modefinance, який включає три роки балансових рахунків 10 000 німецьких фірм. Modefinance – це молода компанія, яка була розпочата у Трієсті як університетський проект, що працює у секторі FinTech. Вона використовує технології штучного інтелекту (зокрема машинного навчання) для розрахунку кредитного рейтингу компаній. Починаючи з 2015 року ця компанія також розпізнається ESMA (Європейським управлінням із цінних паперів та ринків) [4], як ліцензоване рейтингове агентство для фірм та фінансових установ [11]. Modefinance розробляє додаток для професіоналів і малих середніх підприємств (МСП), що дозволяє їм отримати доступ до фінансової інформації усіх фірм Європи. Їхні послуги використовують новітні технології штучного інтелекту та великих даних у поєднанні з фінансовими технологіями, щоб спростити фонові дослідження своїх покупців під час роботи з клієнтами, постачальниками та джерелами фінансування.

Виконуючи прогнози, ми припускаємо, що оцінка кредитного рейтингу Modefinance є «правильною», і ми розглянемо можливі рішення для її прогнозування, використовуючи інформацію балансу, зокрема, розглядаючи ситуації відсутності даних.

Ієрархічна структура наявної бази даних показана на рис. 3.1. У цій структурі визначено чотири рівні:

– Рівень 0: Активи (Total Assets) та Зобов'язання та власний капітал (Total shareholders’ funds & liabilities);

– Рівень 1: Розмежування між основними та оборотними активами, довгостроковими та короткостроковими зобов'язаннями та власним капіталом;

– Рівень 2: Більш детальний опис щодо характеру рахунків (наприклад, кредити та кредиторська заборгованість);

– Рівень 3: Додаткові відомості про певні рахунки (наприклад, «грошові кошти та їх еквіваленти», які відносяться до рахунку «інші оборотні активи»).

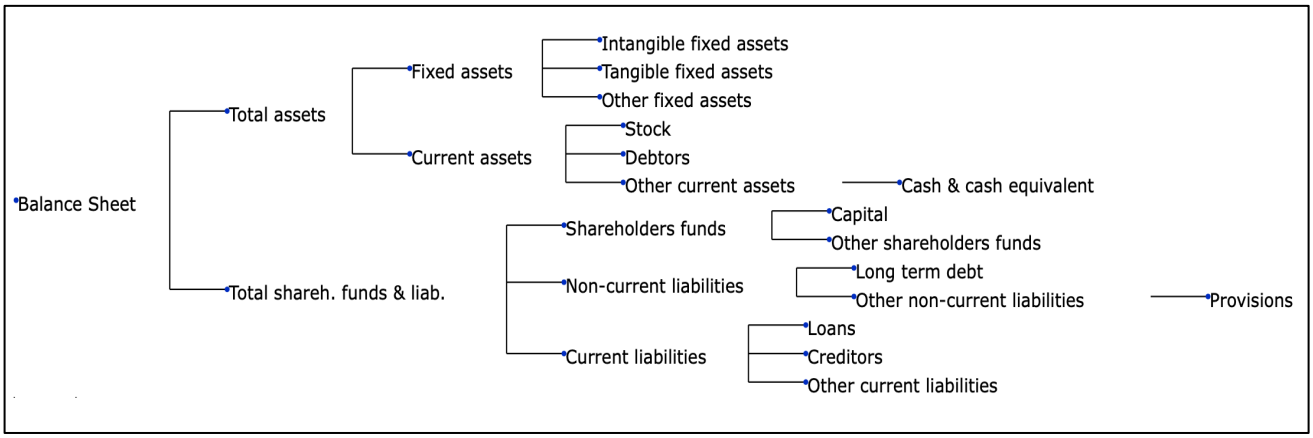

Рисунок 3.1 – Деревовидна діаграма ієрархії балансових рахунків у базі даних

Під час аналізу відсутньої інформації, зосередимося на рівнях 1 та 2, які надають вичерпне та симетричне пояснення показників балансу, не розглядаючи резервні фонди (Provisions) та припускаючи, що інші оборотні активи повністю складаються з грошових коштів та їх еквівалентів. Таким чином, ми повністю нехтуємо третім рівнем.

Також була виконана деяка попередня обробка вихідного набору даних, а саме:

– база даних розширена кодом класифікації NACE, який характеризує галузь діяльності кожної компанії. Ця категоріальна змінна закодована методом швидкого кодування (one-hot encoding) [14] для навчання нейронної мережі та класифікатора Random Forest;

– створена додаткова база даних фінансовх коефіцієнтів, отриманих із балансу (див. розділ 2.1), з метою збереження ключової фінансової інформації зі зменшеною кількістю змінних;

– у обох базах даних розглядаються тільки МСП, вибираючи компанії з загальними активами, які менше або дорівнюють 43 млн. євро. Навіть при такій фільтрації, розмір бази даних істотно не змінюється, оскільки на МСП припадає понад 98% усіх даних;

– кожен окремий рік розглядається як самостійний запис, тому база даних включатиме 30 тисяч окремих рядків, які містять по 22 змінні та код діяльності NACE, закодований методом швидкого кодування.

Для здійснення навчання нейронної мережі масив даних був розділений на дві частини: навчальну (training set) та тестову (test set). Співвідношення вибірок – 80:20 відповідно. Наступним кроком було виділення з навчального набору даних 10% на перевірочну вибірку (validation set) – масив даних, що буде застосовуватись в процесі навчання для підбору гіперпараметрів мережі.

3.1.2 Аналіз пропущених даних

База даних спочатку аналізується щодо наявних у ній пропусків значень. Для розрахунку відсотків пропущених даних, виконаємо наступний код Python:

```
dh_0 = DataHandler()
dh_0.complete_only = False
dh_0.incomplete_only = False
dh_0.full_db = True
dh_0.fill_nans_as_zeros = False
dh_0.ratios_db = False
dh_0.SME_only = True
```

```
dh_0.NACE_classification = False
df_0_X,df0_y = dh_0.load_as_pandas()

level1_incmplete_percentage =
round(df_0_X[level_1_mask].isnull().sum()/df_0_X[level_1_mask].
shape[0]*100,2)
print('Percentage of incomplete Level 1 entries:')
print(level1_incmplete_percentage)
```

Таблиця 3.1 – Відсотки пропущених даних по категоріям першого рівня

Назва змінної	Пропусків даних, %
Fixed assets	0.00
Current assets	0.00
Shareholders' funds	0.01
Non-current liabilities	0.00
Current liabilities	0.00

Результат виконання коду відображено у таблиці 3.1. Як можна побачити, рахунки «Рівня 1» (основні засоби, оборотні активи, власний капітал, довгострокові зобов'язання, поточні зобов'язання) не демонструють систематичних тенденцій відсутності даних, оскільки їх публікація обов'язкова. Розглянемо рахунки другого рівня (таблиця 3.2).

Таблиця 3.2 – Відсотки пропущених даних (показники другого рівня)

Назва змінної	Пропусків даних, %	Класифікація
Intangible fixed assets	2.13	A
Tangible fixed assets	2.13	A
Other fixed assets	2.13	A
Stock	0.81	B
Debtors	0.81	B
Other current assets	0.20	C
Capital	1.92	D
Other shareholders funds	2.01	D
Long term debt	0.11	E
Other non-current liabilities	0.11	E
Loans	43.01	F
Creditors	42.94	F
Other current liabilities	37.09	F

Зі схеми на рис. 3.1 та таблиці 3.2 видно, що існує шість основних груп другого рівня, у яких потенційно відсутня інформація. Ці групи відображено у таблиці 3.2 літерами A,B,…F.

Було обрано саме таку класифікацію, оскільки у більшості випадків показники однієї групи мають пропуски даних одночасно (наприклад, Запаси та Дебіторська заборгованість). Кореляція пропусків даних між цими шести групами (рис. 3.2) показує, що, коли у деякому записі відсутня інформація на стороні активів (оборотні або необоротні), цілком ймовірно, що у цьому ж записі також відсутня інформація про інші складові активів. У той самий час, ммаіж складовими активів та зобов'язаннями кореляція майже відсутня.

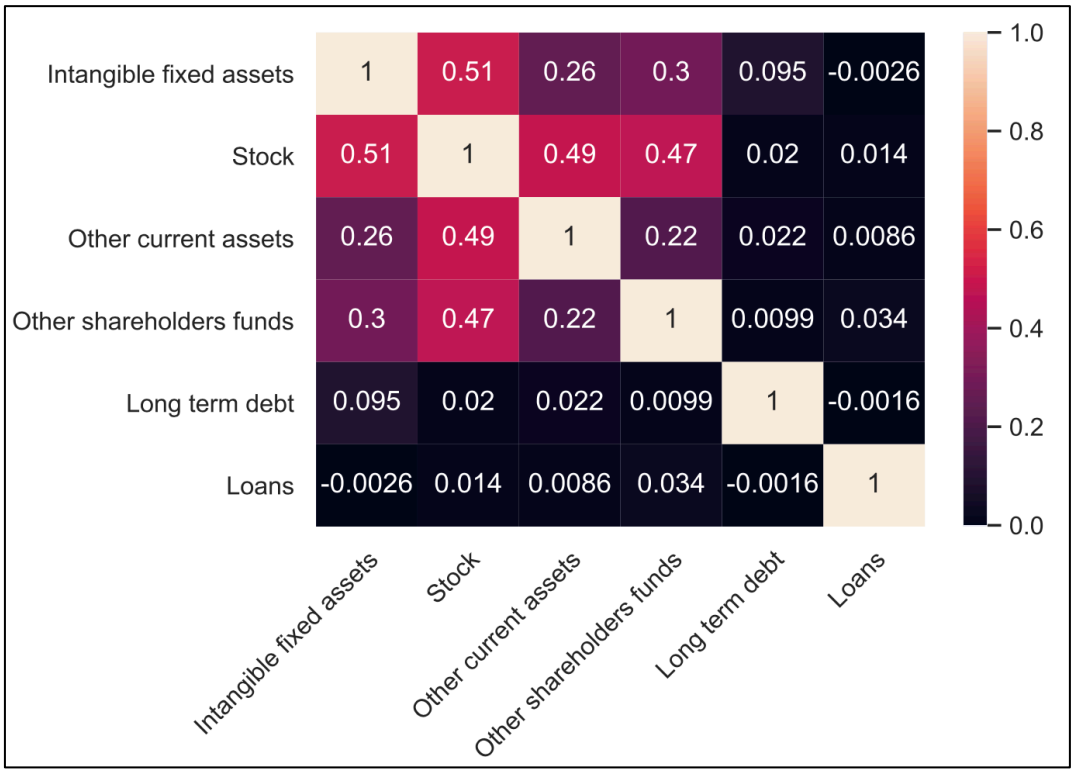

Рисунок 3.2 – Кореляційна матриця пропусків даних між показниками другого рівня

Виходячи з цих міркувань, можна виділити 3 кластери пропусків даних, які потім буде використано для оцінки ефективності моделей класифікації/регресії для класів кредитного рейтингу:

– Пропуски даних по "Активам" (нематеріальні/матеріальні основні засоби, запаси/дебіторська заборгованість, інші оборотні активи, капітал/інші фонди акціонерів);

– Пропуски даних по Довгостроковим боргам;

– Пропуски даних по Кредитам.

Аналіз кореляційної матриці пропусків даних між трьома обраними групами та результуючим кредитним рейтингом (рис. 3.3) показує, що кореляція майже відсутня, отже, вибір кластерів даних є обґрунтованим.

Рисунок 3.3 – Кореляційна матриця пропусків даних між обраними кластерами та кредитним рейтингом

3.2 Вибір бібліотек машинного навчання

Коли перед розробниками ставиться задача із застосування статистичних методів або аналізу даних у своїй роботі, вибір нерідко робиться на користь мови Python. Вона також широко застосовується вченими, чиї завдання мають бути інтегрованими із веб додатками або виробничими середовищами.

Для вирішення описаної у попередніх розділах задачі було обрано мову програмування Python. Вона є оптимальною у даній ситуації, оскільки широко застосовується для організації машинного навчання. Це обумовлено тим, що кодувати даною мовою відносно просто легко та швидко. У той час, як для

реалізації ідентичної функціональності на мові C++ або Java необхідно було б написати набагато значно більшу кількість строк коду. Крім того, мова Python відрізняється простим синтаксисом та наявністю динамічної типізації. При застосуванні нейронних мереж необхідно проводити велику кількість експериментів. Серед них: визначення архітектури нейронної мережі, яка буде підходити для поставленої задачі, оцінка параметрів навчання тощо. Для описаних цілей мова Python є актуальною, оскільки дає можливість легко модифікувати програму для проведення різноманітних експериментів, При цьому програма при таких модифікаціях залишає свої властивості, а саме: легкість читання та високу якість підтримки.

Одну перевагу Python потрібно виділити окремо. Вона є головною причиною застосування мови для машинного навчання. Перевага полягає в наявності великої кількості фреймворків і бібліотек. Вони суттєво полегшують процес написання коду і зменшують кількість часу, що витрачається на розробку. До основних фреймворків і бібліотек, що застосовуються у машинному навчанні, а також будуть використані у даній роботі належать наступні:

– NumPy;
– SciPy;
– Scikit-learn;
– Matplotlib;
– Pandas;
– PyTorch.

NumPy – це одна із ключових бібліотек аналізу даних у Python. Вона полегшує роботу із векторами та матрицями. Бібліотека містить готові методи для операцій різних типів. Завдяки NumPy розробники проводять створення, зміни форми, множення і розрахунок детермінанту матриць, вирішують лінійні рівняння, проводять сингулярне розкладання, а також виконують безліч інших завдань.

SciPy – це бібліотека, призначена спеціально для наукових розрахунків. Вона базується на NumPy та значно розширює її можливості. До неї входять методи

лінійної алгебри разом із методами, призначеними для роботи з ймовірнісними розподілами, інтегральним обчисленнями, а також із перетвореннями Фур'є.

Scikit-learn – бібліотека, що базується на NumPy і SciPy. Вона забезпечує програміста алгоритмами для машинного навчання, а також інтелектуального аналізу наступний категорій даних: кластеризації, регресії, а також класифікації.

Бібліотека Matplotlib відноситься до низькорівневих, вона призначена для створення двовимірних діаграм і графіків. Одна із важливих задач машинного навчання – це візуалізація даних. Використовуючи бібліотеку Matplotlib, можна побудувати графік будь-якого типу.

Pandas – бібліотека, що забезпечує розробників структурами даних і інструментами для аналізу. Може застосовуватися для обробки неповних, невпорядкованих і немаркованих даних (які найбільш часто можуть зустрітися в житті). Pandas дає можливість замінити досить складні операції з даними однією чи двома командами. До її складу входить велика кількість готових методів групування, фільтрації, об'єднання даних, а також можливість розпізнавання різноманітних джерел. У Pandas є можливість об'єднувати таблиці аналогічно до SQL JOIN. При цьому дані беруться безпосередньо з файлів. Завдяки цій системі зникає необхідність організовувати бази даних. Додаткова перевага бібліотеки полягає у швидкості роботи.

PyTorch – це фреймворк машинного навчання з відкритим вихідним кодом, заснований на бібліотеці Torch, що використовується для таких технологій як комп’ютерний зір та обробка природної мови, в основному розроблений компанією Meta AI. Це безкоштовне програмне забезпечення з відкритим вихідним кодом, випущене під модифікованою ліцензією BSD. Хоча інтерфейс Python є більш відшліфованим і основним напрямом розробки, PyTorch також має інтерфейс C++.

На основі PyTorch створено ряд програм глибокого навчання, зокрема Tesla Autopilot, Uber's Pyro, Hugging Face's Transformers, PyTorch Lightning і Catalyst. PyTorch надає дві високорівневі функції:

– Тензорні обчислення (наприклад, NumPy) із сильним прискоренням за допомогою графічних процесорів (GPU);

– Глибокі нейронні мережі, побудовані на основі автоматичної стрічкової системи диференціювання.

PyTorch визначає клас під назвою Tensor (torch.Tensor) для зберігання й роботи з однорідними багатовимірними прямокутними масивами чисел. Тензори PyTorch подібні до масивів NumPy, але також можуть працювати на графічному процесорі NVIDIA з підтримкою CUDA. PyTorch також розробляє підтримку для інших платформ графічних процесорів, наприклад, AMD ROCm і Apple Metal Framework.

Традиційні програми, написані мовою Python працюють досить повільно, проте бібліотеки глибокого навчання використовують різноманітні методи для підвищення швидкості роботи програм. Так, бібліотеки використовують оптимізовані математичні бібліотеки, створені на C. Деякі з бібліотек, засновані на коді мовою Python автоматично генерують код мовою C. У подальшому виконується вже цей код, який працює значно швидше. Для ще більш суттєвого збільшення рівня продуктивності припускається використання графічних прискорювачів (GPU) із використанням NVIDIA cuDNN.

Для створення коду було застосовано інтерактивну оболонку, призначену для мови Python – Jupyter Notebook. Вона забезпечує розширену інтроспекцію, наявність додаткового командного синтаксису, наявність доступу до оболонки системи, взаємодію із pdb відладчиком і Python профайлером, а також забезпечує збереження історії вводу даних, у всіх сеансах, підсвічує код здійснює автоматичне його доповнення. Jupyter Notebook відрізняється можливістю розробляти, документувати та запускати програми, написані мовою Python. До його складу входять 2 компоненти: веб-додаток, що відкривається у браузері, і ноутбуки. Notebooks називаються файли, у яких передбачена можливість роботи із вихідним кодом програми, запускати цей код, здійснювати введення та виведення даних тощо. Веб додаток забезпечує розробника можливістю для редагування Python

коду у браузері. При цьому відбувається підсвічування синтаксису, автовідступ та автодоповненням. Також є можливість відкривати код в браузері, виводити результати обчислень з медіа поданням (схеми, графіки). Ноутбуки – це файли, які містять вихідний код, вхідні і вихідні дані, що були отримані в процесі сесії. Фактично, він являє собою запис роботи. При цьому є можливість заново виконати код, записаний на ньому. Є можливість експортувати ноутбуки у формати PDF і HTML. Jupyter Notebook входить до складу Anaconda. Це дистрибутив для мови Python.

Крім того, у процесі експериментального моделювання було застосовано Google Colaboratory. Це безкоштовна платформа для машинного навчання і нейронних мереж . На ній вже встановлено усі необхідні бібліотеки для навчання нейронних мереж: Keras, TensorFlow, PyTorch та інші. На даній платформі можна використовувати ноутбуки, що мають багато спільного із ноутбуками Jupyter. Головна перевага Google Colaboratory в тому, що навчання нейронних мереж вимагає великих об’ємів ресурсів для обчислення, а Google Colaboratory безкоштовно забезпечує користувачів досить потужним GPU. Таким чином, використовуючи платформу Google Colaboratory, вдалося суттєво підвищити швидкість навчання нейронних мереж.

3.3 Характеристика штучних нейронних мереж

3.3.1 Принцип роботи

Основні типи завдань, які вирішуються з використанням штучних нейронів мереж (НМ) – це класифікація та регресія. Порівняно з деякими іншими методами машинного навчання, НМ мають низку переваг: можливість самостійно визначати важливі для вирішення завдання фактори, здатність до навчання та донавчання, стійкість до «шуму» даних, універсальність. До недоліків НМ відносять високу вимогливість до обчислювальних ресурсів, до обсягів та різноманітності

навчальних даних, також ле водночас існує небезпека перенавчання. Крім того, результати, одержувані НМ, важко піддаються поясненню із людської точки зору. Історично першою широко застосованою моделлю НМ став перцептрон, в якому сигнали, що надходять від датчиків, передаються асоціативним елементам, а потім елементам, що реагують. Вихід перцептрону – це лінійна комбінація значень на його входах (xi) та відповідних вагів у «зв'язках» (wi) – рис. 3.4. Використовувані нині НМ, частіше за все, мають вхідний шар, вихідний шар та довільну кількість проміжних (прихованих) шарів.

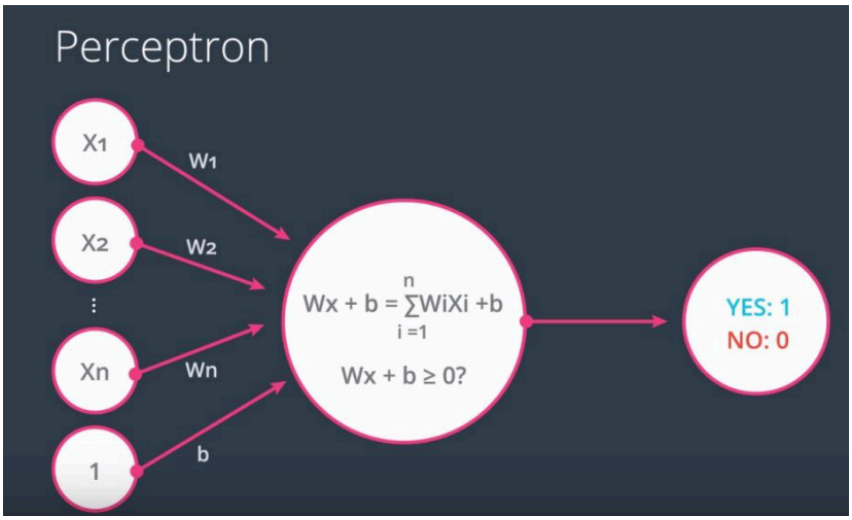

Рисунок 3.4 – Принципова схема класичного перцептрону

У сучасних НМ у якості вузлів виступають штучні нейрони, розподілені за шарами. Нейрони на практиці реалізують певну математичну функцію (функцію активації) від лінійної комбінації всіх вхідних сигналів Отриманий результат пересилається на єдиний вихід у нейрона. Нейрони класифікують на основі їх положення топоології мережі: вхідні, вихідні (часто наявний один) та нейрони в проміжних шарах. У більшості класичних архітектур НМ зв'язки наявні тільки між нейронами попереднього і наступного шару, але не всередині шару.

3.3.2 Налаштування моделі

На відміну від інших алгоритмів машинного навчання, де головною складовою успіху є вибір вхідних параметрів, при створенні нейромережевої

моделі основним завданням інженера є підбір характеристик НМ. Хоча останнім часом з'явилися алгоритми, здатні до автоматизованого підбору таких характеристик, вони мають високу обчислювальну складність, тому досвід інженера, як і раніше, затребуваний. До характеристик НР відносять:

Власне підсумкові значення коефіцієнтів моделі (параметрів) встановлюються алгоритмом навчання НМ, який оптимізує їх так, щоб підсумкова помилка моделі була мінімальною. Тим не менш, важливим моментом може бути ініціалізація. Визначення початкових значень параметрів, які б дозволили швидше виконання алгоритму (зменшуючи час навчання). Однією із стратегій ініціалізації є «перенесення знань» (transfer learning) із існуючих моделей. Наприклад, існуюча НМ, навчена пророкувати вік людей за фотографіями осіб, можливо доопрацьована для розпізнавання емоцій, що виражаються людьми. Ще одна з можливих стратегій при побудові НМ – нормалізація даних, тобто. приведення вхідних значень, наприклад, до однакової шкали та діапазону.

Гіперпараметри – на відміну від параметрів моделі, вони не змінюються у ході роботи алгоритму машинного навчання і отже, мають бути обрані інженером.

3.3.3 Основні гіперпараметри

1. Кількість прихованих шарів. Чим шарів менше, тим швидше працює НМ, проте чим воно більше, тим якісніше НМ може вирішувати складні завдання. При його виборі інженер може або керуватися теоретичними міркуваннями (наприклад, про структури головного мозку людини, які відповідають за розпізнавання візуальної інформації), або почати з невеликої кількості шарів, поступово збільшуючи їх, доки якість моделі не перестане збільшуватися.

2. Функції активації або активаційні функції (англ. activation functions) застосовуються у якості нелінійного фактору після згорткових, агрегаційних, а інколи й інших шарів згорткових нейронних мереж за допомогою формування їх остаточних вихідних даних. Вибір та використання функцій активації сильно впливає на роботу мережі. Це пояснюється тим, що без їх застосування мережа не

могла би навчатися розпізнавати нелінійні закономірності та розподіли. Також слід зауважити, що у випадку, коли рівні нейронної мережі не мають функції активації між собою, буде створюватись залежність лінійного типу між початковим та кінцевим рівнями одного проміжку. Описаний процес фактично робить значення кількості рівнів між ними невпливовим, тобто ці рівні стають еквівалентними одному. Крім того, за допомогою правильного вибору активаційної функції можливий контроль того, який вплив на результати буде мати елемент (нейрон) в залежності від його вхідного значення. Припускається рівень пливу навіть до заборони впливу деяких із них.

У більшості ситуацій функції активації знижують значення або мають меншу за діапазон вхідних даних, з великою ймовірністю обмежену область значень. Це пов'язано з тим, що в інакшому випадку існує ризик нескінченного зростання параметрів. Також вони мають бути диференційовними. Це потрібно для уможливлення алгоритму зворотнього поширення помилки. Відповідно, стають можливими градієнтні методи навчання мережі. Вибір функції активації має значний вплив на спосіб, у який потрібно робити вибір початкових значеннь для нейронної мережі перед початком її навчання. Прикладом може слугувати ситуація, коли функція активації не наближує тотожне перетворення поблизу нульової точки, то у результаті задання значень мережа може гірше навчатися в майбутньому.

Випрямлена лінійна одиниця (англ. rectified linear unit, скорочення ReLU) – це функція активації, що дедалі частіше викристовується в сучасних нейронних мережах. Розповсюдженість саме такої функції пояснюється тим, що вона оброблює дані та навчає мережу у декілька разів швидше, ніж це роблять інші функції активації. У той час також створюється нелінійність. Цей фактор, окрім пришвидшення навчання мережі, ще зменшує вірогідність її перенавчання. Це зумовлено тим, що мережа навчається більш складним зв'язкам за однаковий період часу, у порівнянні з іншими функціями активації, що в свою чергу дозволяє

навчати складніші моделі, використовуючи ті самі дані. Формула знаходження значення виглядає наступним чином:

$$y = f(x) = \max(0, x) \tag{3.1}$$

де y – елемент поточного рівня, x – елемент з вхідних даних.

Як можна зрозуміти із формули функції, вона усуває негативні значення з мережі, надаючи таким чином їм нульового значення. Натомість вона виконує тотожнє перетворення для інших значень. Швидкість роботи цієї функції з апаратної точки зору підвищується також завдяки тому, що вона є звичайним порівнянням значення з нулем. Похідна ReLU дорівнює 0 для від'ємних значень і 1 для всіх інших, тому процес роботи алгоритму зворотнього поширення помилки також стає спрощеним. Тим не менше, ReLU не наближує тотожного перетворення поблизу нульової точки. Саме тому вона потребує наявності хоча б одного додатного початкового значення для ефективного навчання. Проте слід зауважити, що активація не всіх початкових значень позитивно впливає на роботу мережі через ефект розрідження. Впровадження ReLU було пов'язано зі знайденням аналогії з її роботою в обробці візуальної інформації корою головного мозку.

Ще двома функціями, що традиційно застосовуються для функцій активації в нейронних мережах, є гіперболічний тангенс ($f(x) = \tanh(x)$) та функція сігмоїди, що також використовується в якості функції активації на останньому рівні мереж в задачах бінарної класифікації. Це зумовлено тим, що її область значень дорівнює [0;1]. Це дає можливість апроксимувати імовірності належності вхідних даних до двох класів:

$$y = f(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

В задачах багатокласової класифікації на останньому рівні мережі у більшості випадків застосовується функція Softmax завдяки тому, що вона

враховує всі значення з попереднього рівня мережі та може обрахувати відповідну імовірність для кожного з класів:

$$y_i = f(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_i}}, \forall i, j \in (1, N) \quad (3.3)$$

де yi – елемент поточного шару з індексом i,

xi – елемент попереднього шару з індексом i,

N – кількість елементів на попередньому шарі.

Фактично, функція Softmax нормалізує екпоненційні значення даних з попереднього рівня, тим самим переводячи вхідний масив довільних значень в область [0;1]. Цей процес використовується для обчислення імовірностей. Функція може бути модифікована за допомогою заміни бази степені з е на інше число. Це дає можливість збільшити вагу великих значень на попередньому рівні (при збільшенні бази), або навпаки зрівнюючи вплив різних значень (при зменшенні бази). Також функція стійка до збільшення всіх показників з попереднього рівня на стале значення за рахунок збільшення стійкості результуючих імовірностей належності до класів.

3. Алгоритм оптимізації, що задає рух до оптимуму. Класичним варіантом є Stochastic Gradient Descent (SGD), який мінімізує втрати на основі градієнта, вибираючи на кожній ітерації повчальні дані випадковим чином. На сьогодення більш прогресивними вважаються алгоритми RMSProp (відрізняється від SGD тим, що у кожного параметра швидкість навчання, що адаптується) і Adam Optimizer (дозволяє встановлювати гіперпараметри НМ).

4. Швидкість навчання. Розмір кроку на кожній ітерації алгоритму, що прагне мінімізації функції втрат. Занадто великий крок веде до підвищення ризику, що алгоритм «промахнеться» повз оптимуму, а занадто маленький може призвести до збільшення до часу навчання НМ. Сьогодні існують алгоритми, здатні з певним успіхом варіювати цей крок залежно від кривизни функції втрат у поточній точці.

5. Епохи навчання. Відображає задану тривалість навчання НМ. Може задаватися інженером виходячи з обчислювальної складності моделі, але може і визначатися алгоритмом. Наприклад, навчання може автоматично зупинятися у випадку, коли помилка (значення функції втрат) перестає зменшуватися протягом будь-якої кількості епох.

6. Архітектури – це те, яким чином організовані шари у НМ і встановлені зв'язки між нейронами. Схема, представлена в Додатку А, дає деяке уявлення про складність та різноманітність сучасних архітектур НМ. В даний час для вирішення реальних завдань найбільш популярні дві архітектури:

Згорткові НМ (CNN) – набули популярності для багатьох задач, пов'язаних з розпізнаванням зображень (з тих пір, як у 2012 р. на конкурсі ImageNet згорткова НМ значно перевершила конкурентів). CNN є чергуванням згорткових шарів (обробляє попередній шар за фрагментами – наприклад, «вікнами» 8*8 клітин) і субдискретезуючих шарів («ущільнює» фрагменти попереднього шару, зменшуючи їх розмір). Такий підхід є надзвичайно прогресивною альтернативою попіксельному розпізнаванню зображень. До недоліків CNN відносять значну кількість гіперпараметрів мережі (кількість шарів, крок зсуву ядра при обробці шару, необхідність шарів субдискретизації, ступінь зменшення ними розмірності, функція зменшення розмірності тощо), значення яких в основному доводиться підбирати «творчо».

Рекурентні НМ (RNN) – найбільш широко підходять для вирішення задач, пов'язаних із розпізнаванням будь-яких послідовностей: голосових, текстових, музичних тощо. RNN відрізняються наявністю пам'яті: можна сказати, що нейрони запам'ятовують свої попередні відповіді і при наступному запуску використовують їх як додаткові вхідні дані. Для обмеження зростання кількості зв'язків застосовуються алгоритми, що здійснюють скидання запам'ятовуваних значень (забування). Тим не менш, для навчання RNN потрібні дуже значні обчислювальні ресурси (крім значної кількості навчальних даних).

Також, новинкою в галузі архітектури НМ (запропонованою у 2014 р.) є генеративно-змагальні мережі (GAN), здатні до навчання без вчителя. Основним принципом у такому алгоритмі є наявність двох НМ, одна з яких генерує зразки, а інша (дискримінативна) намагається відрізнити правильні зразки від неправильних. Генеративна мережа працює на основі змішування декількох вихідних зразків (існуючих даних) та прагне підвищити відсоток помилок другої мережі. Дискримінаційна мережа навчається розрізняти справжні та підроблені зразки, а результати розрізнення подаються на вхід генеративної мережі, щоб вона могла підправити набір латентних параметрів. З використанням цієї технології були зокрема згенеровані фотографії людей, які не відрізняються від реальних.

3.3.4 Навчання та якість моделі

У машинному навчанні виділяють навчання з учителем (коли відомі пари «стимул-реакція»), навчання без учителя (знаходження раніше невідомих закономірностей у даних) та навчання з підкріпленням (під час взаємодії з деяким середовищем). Як було зазначено вище, основна перевага НМ – здатність навчатися на доступних прикладах даних, тобто адаптувати свою внутрішню структуру (насамперед, значення параметрів) для досягнення найбільшої точності розв'язання задачі (мінімізації помилки, що спостерігається). Найбільш ефективним алгоритмом навчання для сучасних НМ вважається зворотне поширення помилки. Фактично, отриманий на виході НМ результат порівнюється з правильним, а потім алгоритм, рухаючись від вихідного шару по внутрішніх шарах, змінює значення ваги у зв'язках нейронів, щоб мінімізувати помилку.

Для проведення навчання, як правило, весь доступний набір даних поділяється в деяких пропорціях (наприклад, 0.7:0.1:0.2) на три піднабори:

Навчальні дані (training set) – використовуються власне налаштування параметрів моделі у вигляді мінімізації помилки ними.

Перевірочні дані (validation set) – періодично використовуються для контролю процесу навчання, щоб запобігти перенавченню НР.

Тестові дані (testing set) – застосовуються перевірки можливості навченої моделі до генералізації (здатності працювати з новими, раніше не «баченими» моделлю даними). Якість моделі, як правило, розуміється саме як точність її роботи на тестових даних.

Одним із найпоширеніших показників підсумкової якості нейромережевої моделі є середньоквадратична помилка (MSE), яка відображає усереднену відстань між передбаченими моделлю значеннями та реальними значеннями. Значення MSE позитивне, і чим ближче до 0 – тим краща якість моделі.

Навіть в ідеальному випадку значення помилки в НМ не повинно точно дорівнювати 0; інакше це свідчить про явище «перенавчання», тобто. встановлення функціональної залежності від навчальних даних. Така модель, скоріше за все, покаже низьку точність на тестових даних, оскільки вони відрізнятимуться від даних, до яких «пристосувалася» модель. Зазвичай перенавчання виникає, коли обсяг навчальних даних малий, а кількість вхідних параметрів та нейронів у НМ відносно велика. Крім збільшення обсягу навчальних даних та зниження кількості епох навчання, для того, щоб уникнути перенавчання, використовуються також такі стратегії:

Регуляризація – додає штраф у модель, якщо ваги надто великі або їх забагато.

Відсів (dropout) – видалення (зазвичай із деякою заданою ймовірністю) деяких вузлів із НМ, щоб її структура не переускладнювалася.

Поділ даних на пакети (batch size) – надання алгоритму не всіх навчальних даних одночасно, а поділ їх у підмножини.

Рівень втрат (англ. loss layer) зазвичай стає останнім рівнем мережі. Він відповідає за оновлення параметрів у процесі її тренування. На даному рівні за допомогою заздалегідь визначеної функції втрат (англ. loss function) розраховується різниця між справжніми значеннями даних (справжніми даними) та результатами роботи мережі (прогнозом). Після закінчення процесу навчання

мережі при наступному етапі тренування мережі ці дані використовуються для подальшого тренування алгоритма.

Вибір та задання функції втрат – це один із найбільш важливих елементів побудови нейронної мережі, оскільки від цього значною мірою залежить швидкість та якість тренування мережі, а також здатність мережі до оцінювання точності результатів, що отримуються. Значення функції втрат – це критерій оцінки ефективності роботи мережі. Це обумовлено тією обставиною, що за її допомогою переваги, недоліки та неточності мережі мають бути приведені до одного значення. Саме це значення має характеризувати мережу так, щоб результати роботи для різних варіантів моделі можна було впорядкувати за таким параметром як точність. Слід також зауважити, що функція втрат має враховувати усі особливості конкретної задачі, для якої застосовується мережа.

Два головні типи функцій втрат для нейронних мереж – це функції перехресної ентропії (англ. cross-entropy functions) та функції середньоквадратичної похибки (англ. mean-squared error functions). Однак для згорткових мереж в задачах класифікації з метою зіставлення результатів роботи мережі з одним з визначених класів застосовуються імовірності того, що вхідні дані належать кожному з визначених класів. У задачі такого типу знаходження параметрів відбувається за принципом максимальної правдоподібності. Це пов’язано із тим, що нейрона мережа за допомогою наближення до еталоних параметрів намагається наблизити розподіл результатів до розподілу фактичних даних. У результаті цього, шуканим критерієм помилки роботи мережі стає різниця між імовірностями належності вхідних даних до всіх належних класів, що були отримані мережею та заданих у вибірці. Саме тому для задач класифікації зі згортковими мережами застосовується переважно тільки функції перехресної ентропії. Крім того, функція сереньоквадратичного відхилення для задачі класифікації може бути різновидом функції перехресної ентропії, у разі, коли дані розподіляються за нормальним законом.

Функція перехресної ентропії, що застосовується як функція втрат для задачі класифікації, називається функцією логарифмічних втрат (англ. logloss function). Вона застосовується паралельно із застосуванням функцій активації Softmax (для багатокласової класифікації) та сігмоїди (для бінарної класифікації) на останньому рівні мережі для обчислення прогнозів.

Її можна бути знайти за формулою:

$$L(p, y) = -\sum_{n} y_n ln p_n, n \in [1, N] \tag{3.4}$$

де p – отримані за допомогою мережі ймовірності належності даних до вибраних класів,

y – справжні значення для вхідних даних,

N – кількість класів.

3.3.5 Використання нейронних мереж у фінансах

У контексті фінансів та економіки штучні нейронні мережі в основному використовуються з метою прогнозування руху фондових ринків, на основі великої серії історичних даних про ціноутворення. Наукової літератури в цій галузі досить багато, хоча прогнозування фондового ринку насправді є одним із найскладніших завдань у області прогнозів часових рядів через складність кореляції між сигналами, значний шум та волатильність цін.

Інформація є ключовою особливістю будь-якої діяльності у цій галузі. Автоматизація та цифровізація, що мали місце у фінансовому світі, породили надзвичайно великі та диференційовані набори даних, доступні операторам, хоча у той же час вони все більше ускладнюють аналіз даних, у тому числі, через непередбачуваність та помилки у таких наборах даних.

Огляд різних методологій можна знайти у [20], де висвітлюються різні аспекти аналізу фінансових часових рядів: інтелектуальний аналіз даних

фінансових новин, попередня обробка та кластеризація наявних даних, прогнози щодо руху ринків та багато інших аспектів.

Ще одна цікава робота у цій галузі – [21], де аналізується застосування декількох методів машинного навчання для аналізу та прогнозування фінансових даних компаній, що котуються на фондовому ринку. Із [21] з'ясовується, що здатність машинного навчання класифікувати набори даних робить його особливо придатним для високочастотних прогнозів на фондовому ринку, зокрема при використанні глибокого навчання.

Хоча спектр алгоритмів глибокого навчання надзвичайно широкий, з точки зору структури ШНМ і різних гіпер-параметрів, найважчим є представлення даних, тому попередня обробка є одним із найважливіших етапів. У [21] аналізуються 3 різні методи навчання без учителя, використовуючи як вхідні дані ціни на фондовому ринку, для оцінки здатності ШНМ до прогнозування ринку.

Як можна побачити, застосування ШНМ у області прогнозування фондового ринку знаходиться на досить високому рівні, оскільки трейдери почали шукати більш технологічні підходи для максимізації свого прибутку. Однак, інші аспекти світу фінансового аналізу не розглядаються так сильно, принаймні з точки зору публікацій. Це зокрема стосується оцінки кредитного рейтингу на основі даних.

У наступному розділі розглянемо застосування нейронної мережі, що складається з лінійних шарів, з метою пошуку закономірностей у базі даних Mode Finance.

3.4 Характеристика моделі випадкового лісу

Класифікатор Random Forest є одним із найбільш використовуваних у машинному навчанні, та зокрема був застосований для оцінки кредитного рейтингу компаній, наприклад, у [41]. Розглянемо загальні принципи роботи випадкового лісу. Більш вичерпне пояснення цього алгоритму можна знайти у [12].

Випадковий ліс, як випливає з назви, заснований на деревах рішень, які є основним інструментом для визначення того, як властивості впливають на

приналежність елементів до певного класу. Як показано на прикладі, наведеному на рис. 4.3, розглядаючи класифіковані вхідні дані, дерева рішень надають функціональні критерії для розділення цих даних, щоб записи одного класу були настільки ж "схожими", наскільки це можливо між собою, і максимально відрізнялись від записів, що належать до інших класів.

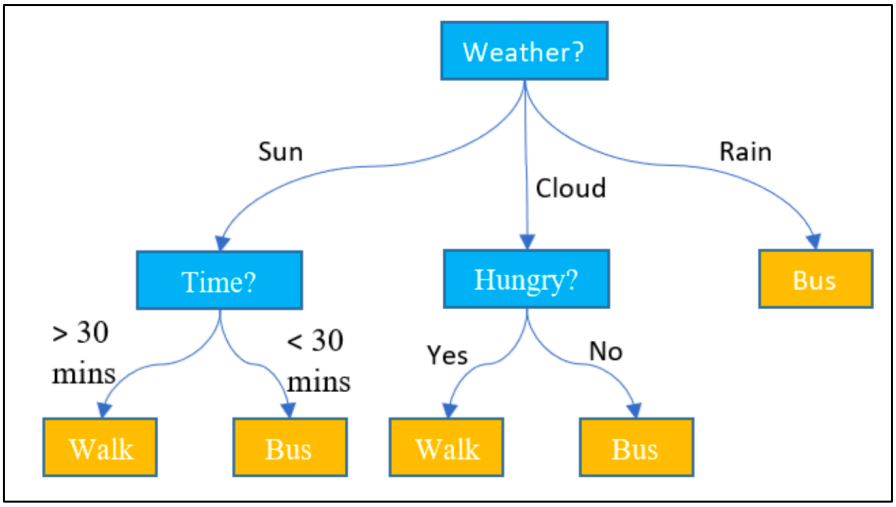

Рисунок 3.5 – Приклад дерева рішень: йти пішки чи поїхати на автобусі, залежно від трьох вхідних змінних

Загалом, класифікатори, засновані на окремих деревах рішень, як правило, мають погані можливості узагальнення, оскільки це інструмент, який за своєю природою перенавчається (точність прогнозу дерева рішень на навчальній вибірці завжди становить 100%). Із цієї причини використовується так звана «мудрість натовпу», для того, щоб використати перевагу простоти реалізації дерев рішень та одночасно подолати проблему перенавчання.

Випадковий ліс складається з декількох окремих дерев рішень, які працюють як "ансамбль". Прогноз класифікатора Random Forest є результатом процесу голосування, в якому оцінюються всі прогнози окремих дерев рішень. Ефективність алгоритму полягає в тому, що внутрішні моделі дерев рішень відносно некорельовані між собою, оскільки вони навчаються на різних партіях вхідного набору даних.

Принцип, що лежить в основі цього результату, полягає в тому, що різні дерева, як правило, компенсують індивідуальні помилки, припускаючи, що немає загального ухилу в один бік, що може бути проблемою набору даних, і впливає на більшість методів класифікації. У той час, як окреме дерево може помилятися, більшість, швидше за все, буде правими, і тому загалом «ансамбль» виведе правильну відповідь.

Висновки до розділу 3

У цьому розділі було проаналізовано вхідні дані системи, розглянуто проблему їх відсутності та досліджено кореляцію між групами пропущених даних. Також було обрано засоби для програмної реалізації системи. Для вирішення описаної у попередніх розділах задачі було обрано мову програмування Python. Вона є оптимальною у даній ситуації, оскільки широко застосовується для організації машинного навчання.

Досліджено бібліотеки Python для роботи з нейронними мережами. Перелік основних особливостей наведено у пункті вибору технологій. Більшість бібліотек мають повну та скорочену версії, що дозволяє випробувати бібліотеки та обрати ту, яка буде найкраще вирішувати поставлені задачі перед розробниками.

Також було проведено детальний аналіз математичних моделей нейронних мереж та випадкового лісу для оцінки кредитного рейтингу підприємства.

4 МОДЕЛЮВАННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ СИСТЕМИ

4.1 Реалізація нейронної мережі

У даній роботі використовується модель глибокої нейронної мережі. Структура мережі показана на рис. 4.1 (для візуалізації фактична кількість нейронів зменшена).

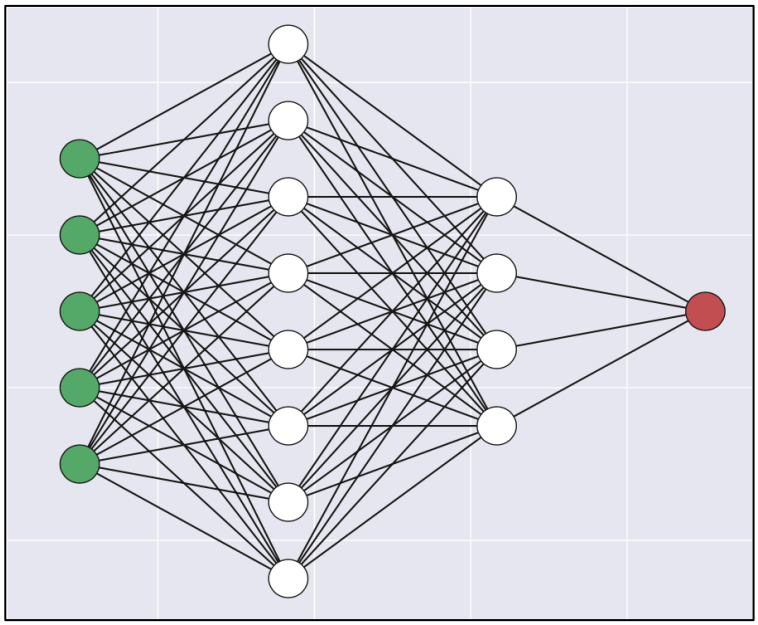

Рисунок 4.1 – Структура глибокої нейронної мережі

У даній структурі вхідні дані подаються до першого шару нейронів, які обчислюють їх лінійні комбінації, а потім передають вихід через функцію активації до наступного шару. У якості функції активації застосовується ReLU, описана у підрозділі 3.4.3.

Проблема визначення класу кредитного рейтингу компаній трансформується у задачу регресії, розглядаючи прогресію від найнижчого класу D до найвищого AAA як лінійну шкалу від 1 до 10 (ModeFinance приймає стандарт, що використовується S&P та Fitch у таблиці 1.1).

Як можна помітити на рис. 4.2, розподіл компаній по класам кредитного рейтингу сильно незбалансований. Це є проблемою для алгоритмів машинного

навчання, оскільки для них навчальна вибірка має бути максимально вичерпною та збалансованою. Для усунення цієї проблеми рейтингові класи на кінцях розподілу (AAA/AA/A та CC/C/D) було згруповано.

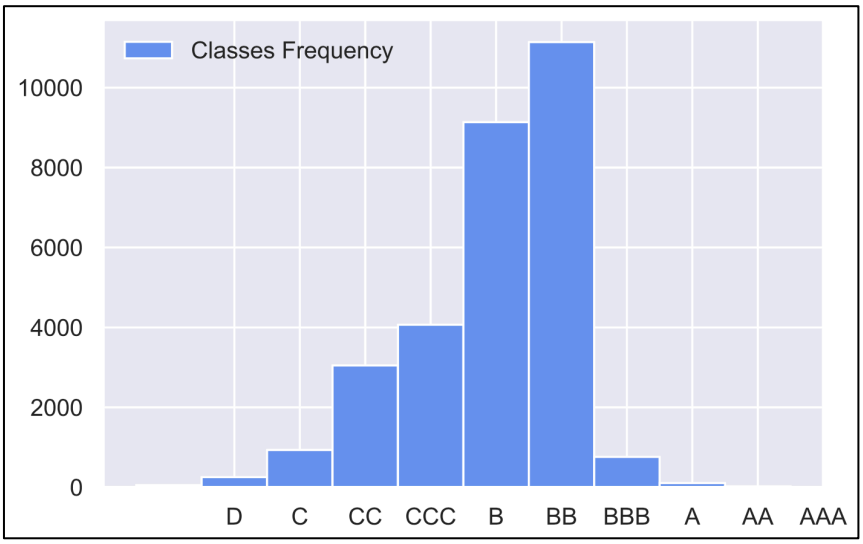

Рисунок 4.2 – Розподіл класів кредитного рейтингу у доступній базі даних компаній

За допомогою модуля PyTorch було створено динамічну нейронну мережу, кількість шарів якої можна змінювати (не більше 5). Моделювання показало, що для наявної бази даних структура із 2 прихованими шарами є оптимальною. Лістинг коду нейронної мережі наведено у додатку А.

Розглянемо найважливіші аспекти попередньої обробки та навчання мережі:

– нейронна мережа навчається і тестується на повній базі даних, яка включає усі елементи;

– пропущені значення заповнюються «0» (що ідентично відсутності інформації для нейронної мережі);

– вхідні параметри Min-Max нормалізовані по стовпцях у діапазоні 0-1;

– для компенсації проблеми незбалансованого набору даних, функція втрат обчислюється за допомогою зваженої середньої квадратної помилки (MSE), тобто для класів із меншою кількістю навчальних даних (AAA/AA/A та CC/C/D) помилки мають більшу вагу:

$$L(x_b, y_b) = \frac{1}{n_{batch}} \sum_{i \le n_{batch}} \left(w(y_i) \cdot (f_\theta(x_i) - y_i)\right)^2$$

де $f_\theta(\cdot)$ – вихідне значення прогнозу нейронної мережі, $w(\cdot)$ – вага класу y_i у функції втрат.

– на противагу тренувальній функції втрат, перевірочна функція втрат – це звичайна середня квадратична похибка (MSE), оскільки кінцевою метою роботи системи є правильне визначення усіх класів, без пріоритету щодо розподілу даних.

Еволюція нормалізованих тренувальних та тестових функцій втрат показана на рис. 4.3. Навчання переривається, як тільки досягається чітка стабілізація можливості прогнозування на тестовому наборі даних, щоб уникнути перенавчання на перевірочному наборі. Це відбувається приблизно на 450 епосі.

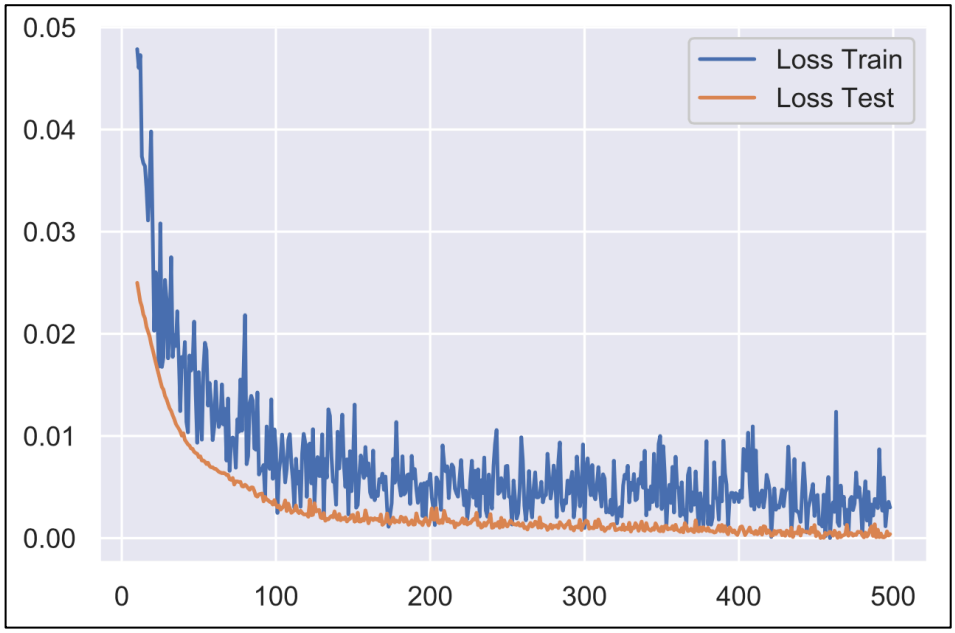

Рисунок 4.3 – Історія навчання мережі: функції втрат на навчальній та тестовій вибірках. Обидві функції нормовані щодо їх максимального значення

4.2 Аналіз результатів нейронної мережі

Для оцінки створеного класифікатора, ми визначаємо наступні два види точності, які потім буде використано для оцінки інших моделей:

– «Точність 0» (*Acc0*): відсоток компаній, для яких модель точно прогнозує клас кредитного рейтингу;

– «Точність 1» (*Acc1*): відсоток компаній, для яких модель визначає точний або сусідній клас.

Регресор нейронної мережі, навчання якого було описано у попередньому підрозділі, оцінюється на тестовій вибірці, яка становить 20% від загального обсягу доступного набору даних.

Для узагальнення результатів, нейронна мережа була навчена та протестована декілька разів на різних партіях навчальних/перевірочних даних. Для всіх симуляцій, *Acc*0 знаходилася у діапазоні 72%-74%, тоді як *Acc*1 завжди становила 99%, що є відмінним результатом. На рис. 4.4 показано нормалізовану матрицю помилок для однієї із симуляцій.

Рисунок 4.4 – Нормалізована матриця помилок на тестовій вибірці.

Acc0 = 72.98%, *Acc1* = 98.8%, RMSE: 0.63

Відхилення RMSE для декількох симуляцій знаходиться у межах діапазону ±0.2, тому можна стверджувати, що нейронна мережа у середньому «навчилася»

визначати правильний клас без зсуву (у статистичному плані, модель $f_\theta(\cdot)$ є адекватною).

Чисельна оцінка точності для різних класів пропущених даних подана у таблиці 4.1. Ці результати показують, що точність моделі трохи погіршується у разі відсутності даних щодо довгострокових боргів (Accuracy 0 = 62.5%, RMSE = 0.83). Ці результати буде використано як еталон для оцінки другого методу, заснованого на випадковому лісі.

Таблиця 4.1 – Оцінка точності моделі для різних класів пропущених даних

	Активи	Довгострокові борги	Кредити
Accuracy 0	71.05%	62.5%	79.2%
Accuracy 1	98.5%	100%	99.65%
RMSE	0.72	0.83	0.59

4.3 Аналіз результатів випадкового лісу

4.3.1 Застосування моделі до повної бази даних

Застосуємо випадковий ліс до аналогічного набору вхідних даних, замінюючи пропуски нульовими значеннями. Для цього імпортуємо клас *RandomForestClassifier* із бібліотеки Scikit-Learn та виконаємо тренування моделі:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split

X_train, X_validation, Y_train, Y_validation = train_test_split(X, y,
test_size=0.2, random_state=1)

m1 = RandomForestClassifier()
m1.fit(X_train, Y_train)

print(m1)

predictions = m1.predict(X_validation)

print(accuracy_score(Y_validation, predictions))

fig = plot_confusion_matrix(Y_validation, predictions,
condensed_classes=True)
fig.savefig('./img/CM_RF.png', dpi=300)
```

Нормалізована матриця помилок для даної моделі показана на рис. 4.5.

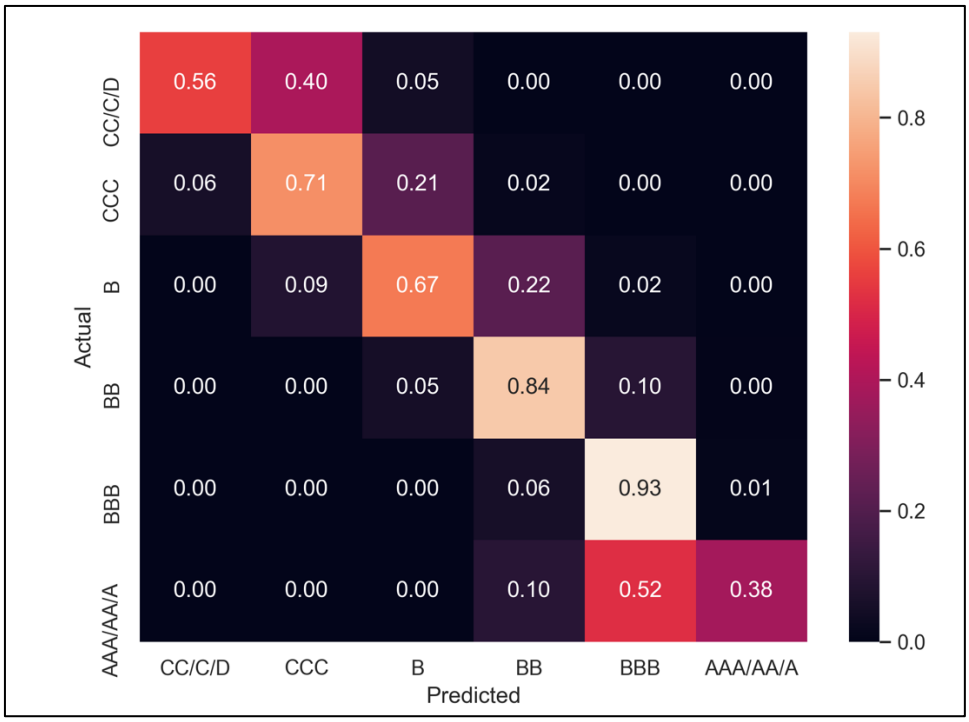

Рисунок 4.5 – Нормалізована матриця помилок на тестовій вибірці.

Acc0 = 81.3%, *Acc1* = 98.6%, RMSE: 0.48

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Рисунок 4.6 – Параметри моделі випадкового лісу

Як видно із результатів моделювання, випадковий ліс дає кращі результати точності *Acc0* та має менший корінь середньоквадратичної помилки, ніж нейронна мережа у попередньому підрозділі.

Найнижча точність прогнозу демонструється у класі AAA/AA/A, що є досить закономірним для дерев рішень в умовах незбалансованості навчальної вибірки.

4.3.2 Модифікація моделі випадкового лісу

Було також вирішено перевірити роботу класифікатора випадкового лісу на модифікованій базі даних, що складається частково із фінансових коефіцієнтів, представлених у розділі 2.1. Існує дві причини такого вибору:

1. Класифікатор Random Forest зазвичай працює особливо добре, коли порівнює вхідні змінні із певними пороговими значеннями, оскільки внутрішні дерева рішень у таких випадках залишаються досить короткими. Використання фінансових коефіцієнтів сприяє цьому;
2. Заповнення пропущених даних буде виконано за допомогою алгоритму максимізації очікувань (Expectation Maximization), який передбачає, що всі спостереження взяті з багатовимірного гаусівського розподілу. Рівень складності та обчислювальні вимоги різко зростають зі збільшенням розмірності вхідних даних, тому використання фінансових коефіцієнтів допомагає знизити кількість вхідних параметрів.

Із цієї причини всі представлені у розділі 2.1 фінансові коефіцієнти включаються у базу даних, при необхідності використовуючи апроксимацію чистого прибутку:

– Коефіцієнт загальної ліквідності (Current Ratio)
– Коефіцієнт швидкої ліквідності (Quick Ratio)
– Коефіцієнт абсолютної ліквідності (Cash Ratio)
– Відношення боргу до власного капіталу (Debt-to-Equity Ratio)
– Рентабельність активів (Return on Assets)
– Рентабельність власного капіталу (Return on Equity)

Крім цих показників, розглядаються також:

– Сукупні активи, з метою врахування розмірності компанії;
– Співвідношення між матеріальними та загальними активами, щоб врахувати обсяг нематеріальних необоротних активів.

Як і у випадку ШНМ, одним із найважливіших питань у навчанні класифікатора випадкового лісу є перебалансування навчального набору даних. У даному випадку було встановлено, що групування класів кредитного рейтингу на кінцях розподілу буде недостатнім. Як було показано раніше на рис. 4.1, у вибірці домінують компанії із середніми оцінками кредитного рейтингу (B, B, BB).

Більш того, на відміну від нейронної мережі, для класифікатора випадкового лісу можуть використовуватися тільки поля без пропусків даних. У попередньому підрозділі пропущені дані було замінено нульовими значеннями, проте таке рішення за своєю сутністю є некоректним по відношенню до значення фінансових коефіцієнтів.

Проблема пропусків даних значно зменшує пул доступних для навчання записів. Ситуація також ускладнюється вимогами щодо розрахунків фінансових коефіцієнти ROA та ROE, які потребують інформації про власний капітал за звітний та попередній роки. Таким чином, за 2016 рік не буде жодного запису, який можна було би використати для навчання моделі випадкового лісу.

Для вирішення розглянутих проблем навчальний набір було додатково перебалансовано. Класи було розділені за частотою: "висока" (BB, BBB), "середня" (B, CCC) і "низька" (AAA/AA/A, CC/C/D), і відсоток записів, що належать до кожної з цих категорій, було обрано методом проб і помилок. У даних результатах було використано наступні пропорції: 10 (висока частота), 8 (середня частота), 3 (низька частота), використовуючи 5250 випадково обраних записів із повної частини бази даних. Усього було використано 53% доступних «повних» записів.

Розподіл «повних» записів по класам, що складає приблизно 1⁄3 усієї бази даних, показано на рис. 4.7, де частоти зображені для:

– вихідної бази даних (тільки записи із усією доступною інформацією)
– згрупованих класів (де об'єднані AAA/AA/A та CC/C/D)
– навчальний набір випадкового лісу.

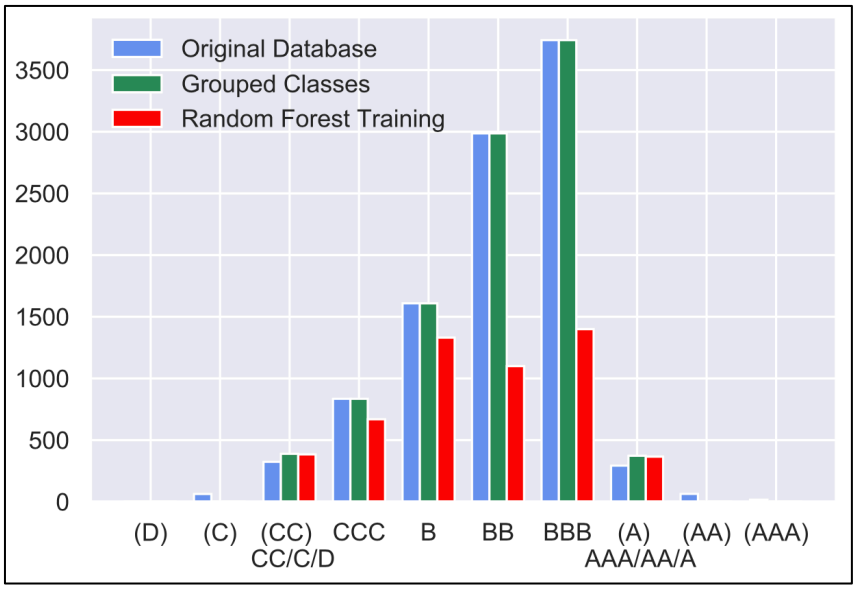

Рисунок 4.7 – Розподіл записів по класах кредитного рейтингу. Зображено лише «повні» записи із бази даних фінансових коефіцієнтів

Модифікована модель випадкового лісу тестується на всіх записах, які не входять до навчальної вибірки. Цей тестовий набір вкрай незбалансований, оскільки значну частину записів не було використано для навчання. Тому ці результати не є дуже інформативними і не повинні використовуватися для прямого порівняння моделі із прогнозами ШНМ, представленими у підрозділі 4.2.

Тим не менш, у залежності від розподілу на навчальну-тестову вибірки, модель випадкового лісу правильно прогнозує:

– 100% записів, що використовуються для навчання;
– 75%-80% повних записів перевірочної вибірки ($Acc0$);
– 98%-99% із одним класом позибки ($Acc1$).

Ці результати знову таки краще, ніж у нейронної мережі, та мають порівняно низьку похибку прогнозу (RMSE ≈ 0.6). Нормалізовану матрицю помилок для одного із прогонів моделі показано на рис. 4.8.

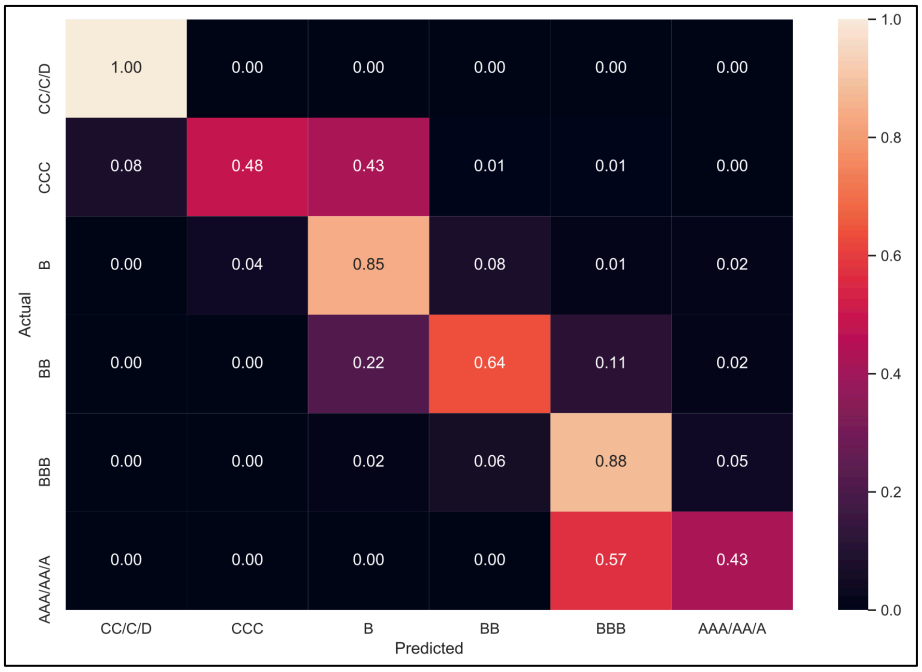

Рисунок 4.8 – Нормалізована матриця помилок на тестовій вибірці для модифікованої моделі випадкового лісу (*Acc0* = 76.2%, *Acc1* = 98.1%, RMSE: 0.58)

Додатково роботу моделі випадкового лісу було перевірено на модифікованому наборі даних, у якому пропущені значення було визначено за допомогою алгоритму максимізації очікувань (Expectation Maximization).

Загалом існує кілька методів визначення пропущених даних: найбільш частими (і найпростішими) з них є ті, що використовують середнє, медіану або найбільш часте значення [38]. Однак, коли проблема стає складнішою, може бути більш ефективним використання складнішого алгоритму для цієї мети. Один з них, відомий як алгоритм максимізації очікувань EM [33], буде використано у даній роботі.

Процедура перевірки модифікованої бази даних на навченому класифікаторі випадкового лісу складається з двох етапів:

1. для роботи алгоритму EM викорситовується повна база даних;
2. записи, які було використано для навчання випадкового лісу, відкидаються (таким чином, фаза тестування не включає жодного запису, який вже було використано для навчання).

Нормалізована матриця помилок для одного із прогонів моделі показана на рис. 4.9.

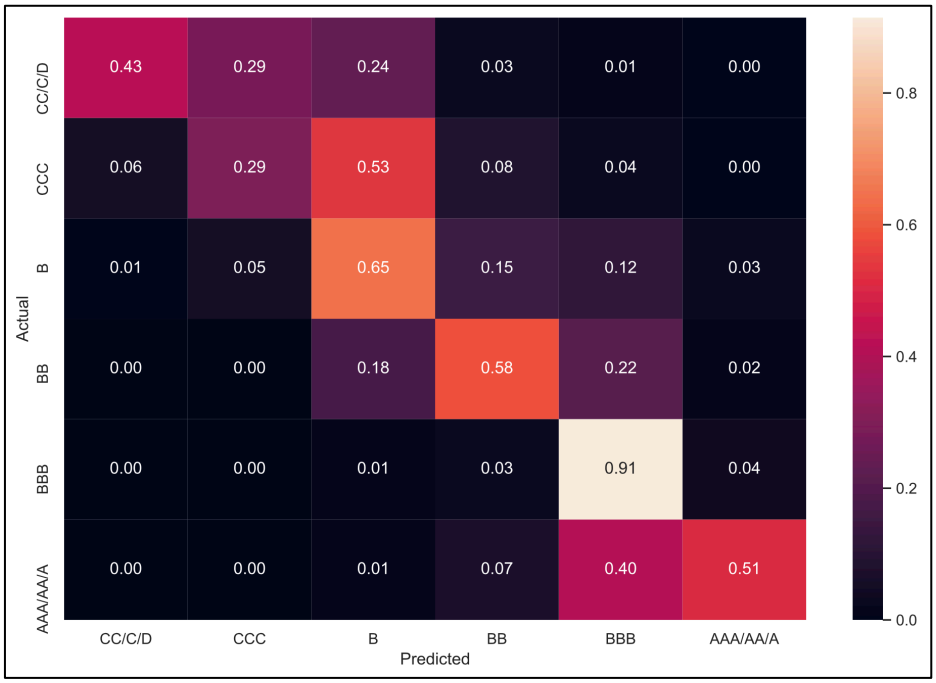

Рисунок 4.9 – Нормалізована матриця помилок моделі випадкового лісу на модифікованій тестовій вибірці (*Acc0* = 68.6%, *Acc1* = 94.5%, RMSE: 0.97)

Як бачимо, показники точності не набагато гірше, ніж у нейронної мережі, але корінь середньоквадратичної помилки має значно більше значення (0,97). Такі результати є очікуваними, враховуючи, що нейронна мережа навчалася на повній вибірці даних.

Висновки до розділу 4

Порівнюючи результати розглянутих моделей, можна відразу помітити, що нейронна мережа працює краще, ніж комбінація алгоритму EM та випадкового лісу. За своєю суттю, правильно навчена нейронна мережа здатна охопити всю інформацію та кореляції у базі даних, хоча пояснити це може бути непросто через її природу «чорного ящику».

Метод EM-RF значно складніший та складається із 4 основних етапів:

1. визначення фінансових коефіцієнтів та апроксимацій для зниження розмірності бази даних;

2. навчання класифікатора випадкового лісу (простіший, ніж нейронна мережа, але дуже ефективний на низькорозмірних наборах даних) на повній частині бази даних, для фіксації взаємозв'язків між фінансовими коефіцієнтами та кредитним рейтингом;

3. приписування відсутніх значень за допомогою алгоритму EM;

4. прогнозування за допомогою випадкового лісу на модифікованому наборі даних, після видалення навчальних даних.

Хоча передбачення нейронної мережі є досить точними ($Acc0$ = 75%, $Acc1$ = 98%), метод EM-RF не сильно відступає та демонструє досить гарний результат ($Acc0$ = 69%, $Acc1$ = 95%). Крім того, розглядаючи точність прогнозів обох моделей в умовах відсутності даних, було виявлено, що прогноз значно погіршується, коли інформація про довгострокові борги компанії невідома.

Спеціальний розділ

ОХОРОНА ПРАЦІ

до кваліфікаційної роботи

на тему:

«ВИКОРИСТАННЯ НЕЙРОННИХ МЕРЕЖ ДЛЯ КРЕДИТНОГО СКОРИНГУ ПІДПРИЄМСТВА»

Спеціальність 122 «Комп’ютерні науки»

122 – БКР – 401.21810304

Виконав студент 4 курсу, групи 401

_____***О. І. Бородін***

«___» _____ 2022 р.

Консультант: старший викладач

_____ ***О. В. Макарова***

«___» _____ 2022 р.

Миколаїв – 2022

5 ОХОРОНА ПРАЦІ

Охорона праці – це система правових, соціально-економічних, організаційно-технічних, санітарно-гігієнічних і лікувально-профілактичних заходів та засобів, спрямованих на збереження життя, здоров'я і працездатності людини у процесі трудової діяльності.

Важливим соціально-економічним завданням є охорона праці. Охорона праці передбачає систему правових, технічних економічних та санітарно-гігієнічних заходів, спрямованих на забезпечення не лише безпечних, а й здорових умов праці. Охорона праці досліджує трудовий процес з позиції забезпечення життя та здоров'я трудящих.

Раціональне поєднання можливостей людини та характеристик машин, відповідний розподіл функцій у середині системи, суттєво підвищують її ефективність. Взаємодія людини та техніки в системі виробництва повинна розглядатися при проектуванні та створенні безпечних умов праці, при вирішенні завдань оптимізації праці.

Метою даного розділу є створення безпечних і здорових умов праці на робочому місці або у виробничому приміщенні. На основі аналізу робочого приміщення визначаються небезпечні ділянки, співвідношення умов праці на робочому місці до нормативних даних, можливі аварійні ситуації і розробляються заходи щодо їх усунення або обмеження наслідків.

Задачами, що розглядаються в даному розділі, є:

– аналіз нормативно-правових актів про охорону праці;

– визначення оптимальних умов праці;

– аналіз робочого середовища відповідно до нормативів.

5.1 Аналіз умов роботи

Основним робочим процесом при виконанні роботи є робота на персональному комп'ютері. Довготривала робота на комп'ютері може вплинути поганим чином на організм людини та призвести до небезпечних захворювань.

Серед небезпечних факторів слід виділити наступні:

а) санітарно-гігієнічні фактори;

б) ергономічні фактори;

в) психофізіологічні фактори.

Аналіз умов роботи допоможе визначити, які дії необхідно провести для покращення умов роботи до нормативних, таких, що відповідають закону про охорону праці.

5.1.1 Мікроклімат

Мікроклімат виробничих приміщень – умови внутрішнього середовища виробничих приміщень, що впливають на тепловий обмін працюючих з оточенням шляхом конвекції, кондукції, теплового випромінювання та випаровування вологи. Ці умови визначаються поєднанням температури, відносної вологості та швидкості руху повітря, температури оточуючих людину поверхонь та інтенсивністю теплового (інфрачервоного) опромінення.

Упродовж робочого часу в робочих приміщеннях забезпечується мікроклімат, що відповідає фізіологічним потребам організму працюючих, із врахуванням енергетичних витрат на виконувану роботу згідно з ДСН 3.3.6.042-99 «Санітарні норми мікроклімату виробничих приміщень».

У приміщеннях для відпочинку, для чергового персоналу, їдальнях, приміщеннях санітарно-побутового призначення і для надання медичної допомоги температура повинна відповідати призначенню цих приміщень відповідно до СНіП 2.09.04-87.

Мікрокліматичні умови виробничих приміщень характеризуються такими показниками:

– температура повітря;

– відносна вологість повітря;

– швидкість руху повітря;

– інтенсивність теплового (інфрачервоного) опромінення;

– температура поверхні.

За ступенем впливу на тепловий стан людини мікрокліматичні умови поділяють на оптимальні та допустимі.

Для робочої зони виробничих приміщень встановлюються оптимальні та допустимі мікрокліматичні умови з урахуванням важкості виконуваної роботи та періоду року.

Оптимальні мікрокліматичні умови – поєднання параметрів мікроклімату, які при тривалому та систематичному впливі на людину забезпечують зберігання нормального теплового стану організму без активізації механізмів терморегуляції.

Допустимі мікрокліматичні умови – поєднання параметрів мікроклімату, які при тривалому та систематичному впливі на людину можуть викликати зміни теплового стану організму, що швидко минають і нормалізуються та супроводжуються напруженням механізмів терморегуляції в межах фізіологічної адаптації.

Категорія робіт – розмежування робіт за важкістю на основі загальних енерговитрат організму.

Легкі фізичні роботи (категорія I) охоплюють види діяльності, при яких витрата енергії дорівнює 105-140 Вт (90-120 ккал/год.) – категорія Іа та 141-175 Вт (121-150 ккал/год.) – категорія Іб. До категорії Іа належать роботи, що виконуються сидячи і не потребують фізичного напруження. До категорії Іб

належать роботи, що виконуються сидячи, стоячи або пов'язані з ходінням та супроводжуються деяким фізичним напруженням.

Фізичні роботи середньої важкості (категорія II) охоплюють види діяльності, при яких витрата енергії дорівнює 176-232 Вт (151-200 ккал/год.) – категорія IIа та 233-290 Вт (201-250 ккал/год.) – категорія IIб. До категорії IIа належать роботи, пов'язані з ходінням, переміщенням дрібних (до 1 кг) виробів або предметів в положенні стоячи або сидячи і потребують певного фізичного напруження. До категорії IIб належать роботи, що виконуються стоячи, пов'язані з ходінням, переміщенням невеликих (до 10 кг) вантажів та супроводжуються помірним фізичним напруженням.

Важкі фізичні роботи (категорія III) охоплюють види діяльності, при яких витрати енергії становлять 291-349 Вт (251-300 ккал/год.). До категорії III належать роботи, пов'язані з постійним переміщенням, перенесенням значних (понад 10 кг) вантажів, які потребують великих фізичних зусиль.

Робота над дипломним проектом відноситься до категорії робіт Іа (легкі роботи), так як виконується сидячи і потребує невеликих витрат енергії: 75 ккал/год.

Для виявлення інформації про те, чи відповідає робоче приміщення вказаним нормам, наводиться таблиця 5.1.

Таблиця 5.1 – Допустимі величини температури, відносної вологості та швидкості руху повітря в робочій зоні виробничих приміщень для категорії робіт Іа на постійних робочих місцях

Період року	Температура, град.С		Відносна вологість (%) на робочих місцях	Швидкість руху (м/сек) на робочих місцях
	Верхня межа	Нижня межа		
Холодний період року	25	21	75	не більше 0,1
Теплий період року	28	22	55 – при 28 град. С	0,2-0,1

Згідно даних таблиці 5.1 можна зробити висновок, що робоче приміщення повністю задовольняє оптимальним умовам мікроклімату.

5.1.2 Освітленість

Робота з комп’ютером супроводжується тривалими зоровими навантаженнями і негативно впливає на здоров’я очей. Правильно виконане освітлення робочого місця здійснює позитивний вплив на психофізіологічний стан людини, сприяє підвищенню ефективності та високій працездатності. Досягнення

оптимальних умов роботи проводиться шляхом забезпечення природнього освітлення в світлий час доби та сприятливого штучного освітлення в темний період доби.

Освітлення робочих зон повинно відповідати вимогам ДБН В.2.5-28-2006 «Природне і штучне освітлення», затверджених наказом Міністерства будівництва, архітектури та житлово-комунального господарства України від 15.05.2006 № 168.

Робочі зони, в яких працівникам у випадку відключення штучного освітлення може загрожувати небезпека, обладнуються достатнім аварійним освітленням.

Для забезпечення умов, що необхідні для зорового комфорту, в системі освітлення повинні бути реалізовані наступні вимоги:

– рівномірне освітлення;
– оптимальна яскравість;
– відсутні відблиски;
– правильний контраст;
– правильна кольорова гамма;
– відсутній стробоскопічний ефект або пульсація світла.

Робота над дипломним проектом відноситься до IV розряду зорових робіт (мінімальний розмір об’єкту, що розглядається – товщина напису букви – 0,3 мм; розряд зорової роботи – робота високої точності) при великому контрасті та світлому фоні (підрозряд зорової роботи «г»).

Робоче місце освітлюється у світлий період доби через вікно (природнє бокове освітлення), яке виходить на південно–західний бік, і сонячне світло не перекривається посторонніми об’єктами. У темний час доби штучне освітлення забезпечується люмінесцентними лампами. Часта переадаптація ока до різних значень яскравості і дистанцій є одним з головних негативних факторів при роботі з дисплеєм. Несприятливим фактором світлового середовища є невідповідність нормативним значенням рівнів освітленості робочих поверхностей стола, екрану, клавіатури. Часто на екрані помітне дзеркальне відображення джерел світла і

оточуючих предметів. Враховуючи всі дані, можна зробити висновок, що робота ускладнюється та може призвести до порушень основних функцій зорової системи.

Розрахунок освітленості. Нормованим параметром природного освітлення являється коефіцієнт природного освітлення (КПО). КПО встановлюється в залежності від розряду виконуваних зорових робіт.

Для штучного освітлення нормованим параметром виступає $E_{мін}$ – мінімальний рівень освітленості, та $K_п$ – коефіцієнт пульсації світлового потоку, який не повинний бути більшим ніж 20 %.

Мінімальна освітленість встановлюється в залежності від розряду виконуваних зорових робіт. Для IV розряду зорових робіт вона складає 400-700 лк.

Для штучного освітлення у приміщенні використовуються люмінесцентні лампи. Розрахунок штучного освітлення проведемо для кімнати площею 16 м^2, ширина якої складає 4 м, довжина – 4 м, висота – 3 м.

Скористаємося методом використання світлового потоку. Для визначення потрібної кількості світильників, які повинні забезпечити нормований рівень освітленості, визначимо світловий потік, що падає на робочу поверхню за формулою [43]:

$$F = \frac{E * K * S * Z}{\eta} \tag{5.1}$$

де F – світловий потік, що розраховується, лм;

E – нормована мінімальна освітленість, лк; E = 400 лк;

S – площа освітлюваного приміщення (у нашому випадку S = 16 м^2);

Z – відношення середньої освітленості до мінімальної (зазвичай приймається рівним 1,1... 1,2, в нашому випадку Z =1,1);

K – коефіцієнт запасу, що враховує зменшення світлового потоку лампи в результаті забруднення світильників в процесі експлуатації (його значення

залежить від типу приміщення і характеру робіт, що проводяться в ньому, в нашому випадку К = 1,5);

η – коефіцієнт використання світлового потоку, (виражається відношенням світлового потоку, що падає на розрахункову поверхню, до сумарного потоку всіх ламп, і обчислюється в долях одиниці; залежить від характеристик світильника, розмірів приміщення, забарвлення стін і стелі, що характеризуються коефіцієнтами відбиття від стін ($\rho_{стін}$) і стелі ($\rho_{стелі}$)), значення коефіцієнтів дорівнюють $\rho_{стін}$ = 40 % і $\rho_{стелі}$= 60 %.

Обчислимо індекс приміщення за формулою [43]:

$$I = \frac{S}{h(A + B)} \tag{5.2}$$

де S – площа приміщення, S = 16 м^2;

h – розрахункова висота підвісу, h = 2,9 м;

A – ширина приміщення, A = 4 м;

B – довжина приміщення, B = 4 м.

Підставивши значення отримаємо значення індексу рівне 0,69. Знаючи індекс приміщення I, η =0,24.

Підставимо всі значення у формулу для визначення світлового потоку F отримаємо його значення рівне 44000 лм.

Для освітлення використані люмінесцентні лампи типу ЛБ-40-4, світловий потік яких F = 3000 лм. Розрахуємо необхідну кількість ламп у світильниках за формулою [44]:

$$N = \frac{F}{F_л} \tag{5.3}$$

де N – кількість ламп, що визначається;

F - світловий потік, F = 44000 лм;

$F_л$- світловий потік лампи, $F_л$ = 3000 лм.

Кількість ламп дорівнює 14. На момент виконання роботи у приміщенні було 8 люмінесцентних ламп, тому рівень штучного освітлення не задовольняв санітарним нормам.

5.1.3 Електробезпека

Згідно з правилами улаштування електроустановок існує три класи приміщень, що розрізняються за ступенем ризику ураження електричним струмом: приміщення без підвищеної небезпеки, приміщення з підвищеною небезпекою і приміщення особливо небезпечні.

Робоче приміщення, в якому пишеться дипломна робота, відноситься до приміщень без підвищеної небезпеки, так як воно не сире (вологість повітря не перевищує 75%), температура не перевищує +28 С (середнє значення +22 С), і регулярно проводиться прибирання приміщення, що не дозволяє утворюватися струмопровідній пилюці. Робоча напруга в мережі 220 В. Проводи ізольовані і розташовані таким чином, що ймовірність випадкового контакту людини з проводами значно знижена.

5.1.4 Шум

Шум визначають як сукупність аперіодичних звуків різної інтенсивності і частоти.

Під час облаштування робочого місця працівника з екранними пристроями необхідно обирати таке устаткування, яке не створює зайвого шуму та не виділяє надлишкового тепла. Рівні шуму на робочих місцях осіб, які працюють з екранними пристроями, мають відповідати вимогам Санітарних норм виробничого шуму, ультразвуку та інфразвуку ДСН 3.3.6.037-99, затверджених постановою Головного державного санітарного лікаря України від 01 грудня 1999 року № 37.

Шуми різняться за різними параметрами, бувають такі, як:

– низько-, середньо-, і високочастотні;

– постійні і непостійні;

– тривалі і короткочасні.

Велике значення надається амплітудно-тимчасовим, спектральним і імовірнісним параметрами непостійних шумів, які характерні для сучасного виробництва. Інтенсивний шум сприяє зниженню працездатності, знижує концентрацію і швидкість роботи, є причиною накопичення втоми [48].

5.1.5 Вібрація

Вібрація – малі механічні коливання, що виникають в тілах під впливом фізичного поля. Сильна вібрація негативно позначається на здоров'ї людини, і слід обладнати своє робоче місце таким чином, щоб уникнути її впливу. В іншому випадку при впливі вібрації погіршується зір, координація, робота внутрішніх органів.

5.1.6 Електромагнітні випромінювання

Комп'ютер випускає електромагнітне випромінювання, яке має більший вплив зі зменшенням відстані від джерела до об'єкта. Вважається, що електромагнітне випромінювання сприяє розладу нервової системи і негативно впливає на серцево-судинну систему. Розрізняють чотири види опромінення:

– професійне;

– непрофесійне;

– опромінення в звичайному середовищі;

– опромінення в лікувальних цілях.

Ступінь впливу електромагнітних випромінювань визначається частотою випромінювання, інтенсивністю і тривалістю впливу.

5.1.7 Ергономіка робочого місця

Правильна організація робочого місця може мати значний вплив на продуктивність праці і концентрацію і дозволити більш ефективно вирішувати поставлені завдання.

Ергономічні вимоги до робочого місця, робочу позу працівників під час використання виробничого обладнання роботодавець повинен враховувати під час застосування мінімальних вимог до виробничого обладнання, зазначених у розділі III Вимог безпеки та захисту здоров'я під час використання виробничого обладнання працівниками.

Раціональне планування робочого місця має забезпечувати: найкраще розміщення знарядь і предметів праці, не допускати загального дискомфорту, зменшувати втомлюваність працівника, підвищувати його продуктивність праці. Площа робочого місця має бути такою, щоб працівник не робив зайвих рухів і не відчував незручності під час виконання роботи. Важливо мати також можливість змінити робочу позу. Проте доцільно виключати або мінімізувати всі фізіологічно неприродні і незручні положення тіла.

5.2 Безпека у надзвичайних ситуаціях

Медичні працівники можуть наражатися на професійні небезпеки, що становлять ризик захворювань, травм і навіть смерті. Ці професійні ризики включають:

(а) професійне інфікування;

(б) шкірні розлади та тепловий стрес від тривалого використання ЗІЗ;

(в) вплив токсинів через інтенсивне використання дезінфікуючих засобів;

(г) психологічний дистрес;

(д) хронічну втому;

(е) стигматизацію, дискримінацію, фізичне та психологічне насильство та переслідування.

Зменшення цих загроз та захист здоров'я, безпеки та добробуту медичних працівників вимагає узгоджених та комплексних заходів щодо профілактики інфекцій та інфекційного контролю, безпеки і здоров'я на роботі, управління трудовими ресурсами в галузі охорони здоров'я та забезпечення психічного здоров'я та психосоціальної підтримки. Недостатні заходи з безпеки і здоров'я на

роботі можуть призвести до збільшення серед медичних працівників частоти захворювань, пов'язаних з професійною діяльністю, високих показників відсутності на роботі, зниження продуктивності праці та зниження якості медичної допомоги.

Професійний вплив хвороб на медичних працівників може відбуватися в будь-який час у закладах охорони здоров'я та в позалікарняному середовищі, під час пов'язаних із роботою подорожей до району циркуляції вірусу серед місцевого населення, а також на шляху до робочого місця та з нього. Систематичний аналіз свідчить про те, що професійний ризик для медичних працівників може зростати в певних клінічних умовах або при неоптимальній гігієні рук, тривалому робочому часі, неправильному або неоптимальному використанні чи відсутності ЗІЗ.

Вірогідність професійного впливу хвороб на медичних працівників може визначатися ймовірністю прямого, непрямого або тісного контакту з людиною, зараженою вірусом. Це включає безпосередній фізичний контакт або догляд за нею, контакт із забрудненими поверхнями та предметами, проведення аерозоль-генеруючих процедур у випадку хворих пацієнтів без належного особистого захисту або роботу з інфікованими людьми в закритих людних місцях з недостатньою вентиляцією. Ризик професійного впливу зростає відповідно до рівня передачі захворювань серед населення. Роботодавці, за результатами консультацій з медичними працівниками та їхніми представниками, а також за підтримки експертів з питань профілактики інфекцій та інфекційного контролю (ПІІК) та безпеки і здоров'я на роботі, повинні проводити та регулярно оновлювати оцінку ризиків щодо хвороб на робочому місці. Мета полягає у визначенні рівня ризиків потенційного професійного впливу стосовно різних робочих місць, робочих завдань та умов роботи, а також плануванні та впровадженні адекватних заходів щодо запобігання та пом'якшення ризиків, а також оцінки придатності медичних працівників до роботи і їх повернення до роботи.

Наступні рівні ризику на робочому місці можуть бути корисними для роботодавців та служб з безпеки і здоров'я на роботі при проведенні швидких оцінок ризику потенційного професійного впливу для різних робіт і завдань.

1. Низький ризик – роботи або завдання без частого тісного контакту з громадськістю чи іншими особами, які водночас не потребують контакту з людьми із підтвердженим або підозрюваним випадком інфікування.

2. Середній ризик – роботи або завдання з тісним частим контактом з пацієнтами, відвідувачами, постачальниками та колегами, але які не потребують контакту з людьми із підтвердженим або підозрюваним випадком інфікування.

3. Високий ризик – роботи або завдання з високою вірогідністю тісного контакту з людьми із підтвердженим або підозрюваним випадком інфікування або контакту з предметами та поверхнями, можливо зараженими вірусом.

4. Дуже високий ризик – роботи або завдання з ризиком контакту з аерозолями, в умовах регулярного проведення аерозоль-генеруючих процедур у випадках пацієнтів з COVID-19 або робота з інфікованими людьми в закритих людних місцях без належної вентиляції.

Рівні ризику на робочому місці, навіть в одній і тій самій робочій обстановці, можуть відрізнятися залежно від завдань та ролей медичних працівників. Отже, оцінка ризиків на робочому місці повинна проводитися для кожного конкретного середовища, а також для кожної ролі, завдання чи комплексу завдань.

Під час пандемії медичні працівники можуть працювати довший час із більшими навантаженнями та недостатнім часом для відпочинку та відновлення сил. Такі вимоги можуть призвести до хронічної втоми та нестачі енергії в поєднанні зі зниженням пильності, координації та ефективності роботи; збільшеного часу реакції; порушення когнітивної ефективності та емоційного притуплення або зміни настрою.

У випадку оголошення надзвичайної ситуації, такої як пандемія, винятки із положень про звичайний робочий час повинні дозволятись лише тимчасово та відповідно до рекомендацій МОП. Слід вживати заходів для оптимальної

організації робочого часу, чергувань та відпочинку, наскільки це практично можливо, з урахуванням місцевої ситуації.

Рекомендації ВООЗ та МОП щодо запобігання виснаженню під час надзвичайних ситуацій:

Тривалість зміни. П'ять 8-годинних змін або чотири 10-годинних зміни на тиждень зазвичай терпимі. Більш довгі зміни є фактором ризику виснаження. Залежно від навантаження, 12-годинні робочі дні можуть вимагати частіших днів проміжного відпочинку. Ввечері та вночі коротші зміни (наприклад, вісім годин) переносяться краще, ніж довші зміни.

Втома посилюється при нічній роботі через нічну сонливість та недостатній денний сон. Перевагу слід віддавати зміщенню змін вперед (з ранішньої – до денної – до нічної), враховуючи вподобання працівників та місцеві умови.

Робоче навантаження. Збалансувати зміни з більш легкими та важкими робочими завданнями. Вивчити робочі вимоги щодо тривалості зміни. Дванадцятигодинні зміни більш терпимі для «легших» завдань (наприклад, робота з документами). Коротші робочі зміни допомагають протистояти виснаженню від дуже інтенсивної роботи, фізичних навантажень, екстремальних умов або впливу інших небезпек для здоров'я або безпеки.

Відпочинок та відновлення. Запровадити політику щодо тривалості робочого часу та перерв на відпочинок (наприклад, щонайменше 10 годин поспіль на день гарантованого вільного часу для отримання 7-8 годин сну та 48 годин перерви після 14 днів роботи поспіль). Забезпечення частих коротких перерв на відпочинок (наприклад, кожні 1-2 години) під час важкої роботи є більш ефективним проти втоми, ніж менша кількість довших перерв. Передбачити довші перерви для харчування. Запланувати один-два повних дні відпочинку, щоб виконувати п'ять послідовних 8-годинних змін або чотири 10-годинних зміни. Розглянути можливість двох днів відпочинку після трьох послідовних 12-годинних змін.

Міжнародні рекомендації щодо протидії насильству та домаганням на робочому місці у галузі охорони здоров'я:

- залучити працівників і їхніх представників до розроблення, впровадження та контролю реалізації політики щодо запобігання та протидії насильству, домаганням, стигмі та дискримінації на робочому місці, а також забезпечити обізнаність і дотримання усім персоналом, включно з керівництвом;
- запровадити процедури заборони дискримінації та домагань та сприяти чесному поводженню з працівниками;
- забезпечити інструктажі з питань безпеки та охоронний персонал у зонах високого ризику;
- проводити ініціативи з підвищення обізнаності та проведення тренінгів щодо насильства та домагань;
- надавати своєчасну та точну інформацію персоналу та пацієнтам, щоб зменшити невизначеність та тривогу;
- впорядкувати маршрут пацієнтів та уникати скупчення людей та часу очікування;
- забезпечити безпечний доступ до робочого місця та з нього, а також чітко визначені аварійні виходи;
- забезпечити сигналізацію (наприклад, тривожні кнопки, телефон, звуковий сигнал, короткохвильове радіо) у зонах передбачуваного ризику;
- розробити протоколи для звітування, розслідування та реагування на випадки насильства, стигматизації та дискримінації в неосудливий спосіб;
- запровадити заходи для захисту заявників, жертв, свідків та викривачів від віктимізації або помсти та забезпечити захист конфіденційності;
- надавати конфіденційну допомогу, консультування та підтримку жертвам насильства, домагань та стигматизації;
- регулярно оцінювати ризик насильства та домагань, консультуючись із працівниками та їхніми представниками;

- уникати носіння форми чи іншого медичного одягу під час поїздки на роботу та під час відвідування громадських місць, домогосподарств або громади з приводів, які не пов'язані з виконанням професійних обов'язків.

Кожен заклад охорони здоров'я повинен мати програму з безпеки і здоров'я на роботі, як це сформульованов концепції ВООЗ/МОП щодо програм безпеки і здоров'я медичних працівників на роботі, а також призначеного та відповідним чином підготовленого координатора з питань безпеки і здоров'я на роботі.
У великих закладах охорони здоров'я повинен бути комітет з питань безпеки та здоров'я на роботі, до складу якого входять працівники і керівництво, та службу безпеки і здоров'я на роботі з основними профілактичними функціями.

Медичні працівники повинні проходити конфіденційну оцінку медичними фахівцями щодо придатності до здійснення певних робіт і завдань та щодо будь-якого порушення здоров'я, яке може становити підвищений ризик хвороби або травми від професійного впливу. Деякі працівники потребують особливої уваги. До них належать:

- медичні працівники старшого віку та ті, хто мають хронічнізахворювання, або вагітні працівниці. Вони можуть мати підвищений ризик розвитку серйозних захворювань;
- працівники зі станами психічного здоров'я, які можуть погіршитися внаслідок додаткових переживань, таких як піклування про велику кількість важкохворих та помираючих пацієнтів;
- працівники з підвищеним робочим навантаженням, тривалими робочими змінами або занепокоєнням про власне здоров'я чи здоров'я членів сім'ї.

Якщо інфікування мало місце в результаті професійної діяльності, такий випадок може розглядатися як нещасний випадок на роботі. Такі випадки необхідно розслідувати та повідомляти про це державному органу, відповідальному за здійснення виплат, пов'язаних із нещасним випадком на роботі, відповідно до національних нормативних актів. Країни повинні розглянути

можливість оновлення своїх переліків професійних захворювань, критеріїв впливу та звітування.

Висновки до розділу 5

Під час роботи на людину впливають різні параметри виробничої обстановки. Фізіологічне забезпечення трудової діяльності на належному рівні й у проміжок даного відрізка робочого часу знаходить вираження у дієздатності. У зв'язку з цим можна сказати, що дієздатність людини є фізіологічною основою продуктивності живої праці. Тому при рівності всіх інших умов, продуктивність праці буде тим вище, чим вище працездатність людини, яка бере участь у трудовому процесі.

Поліпшення умов і охорони праці дуже вагомо впливають як на економічні так і на соціальні результати виробництва. Комфортні умови праці сприяють високій працездатності. Механізація і автоматизація виробничих процесів, удосконалення технології, скорочення і ліквідація важкої ручної праці є важливим фактором високої працездатності і продуктивності праці.

Результат розрахунків показав, що рівень температурний режим та рівень освітленості у приміщені не відповідають санітарним нормам. Зокрема, було виявлено, що за санітарними нормами має бути встановлено 14 ламп зі світловим потоком 3000 лм, тоді як фактично їх було лише 8. Ця ситуація може негативно вплинути на роботу працівників.

Проте, при дотриманні вказаних норм та рекомендацій можна забезпечити безпечні умови праці, за яких рівні шкідливих впливів зводяться до мінімуму. Це дозволяє зберегти здоров'я та високу працездатність при регулярній довготривалій роботі.

ВИСНОВКИ

Швидкий доступ до достовірної інформації кредитоспроможності компаній є важливим аспектом діяльності суб'єктів економічної та фінансової діяльності на будь-якому рівні. Доказом цього є рекомендації щодо банківського законодавства та нормативних актів, викладені у Базелі II [23]. Враховуючи це, моделі ймовірності дефолту набирають все більшу актуальність, враховуючи підвищення прозорості та поліпшення гармонізації стандартів бухгалтерського обліку.

У даній роботі було виконано проектування інформаційної системи визначення кредитного рейтингу компаній, яка ґрунтується на методах машинного навчання, та здатна надавати статистично адекватні прогнози навіть в умовах часткової відсутності вхідних даних. Було використано базу даних ModeFinance, припускаючи, що вона містить достовірну оцінку кредитного рейтингу підприємств.

Спочатку у якості класифікаційної моделі було застосовано глибоку нейронну мережу (яка виступила інструментом «чорного ящика») та показала досить гарні результати виявлення закономірностей між балансовими рахунками компаній та кредитним рейтингом. Зокрема, ШНМ має невелике значення кореня середньоквадратичної помилки, що характеризує її як статистично адекватну модель. Розглянутий тип нейронної мережі також добре підходить для обробки записів із пропущеними значеннями, хоча у той самий час модель розглядає поля із нульовими значеннями як такі, де значення відсутні. Цей аспект є проблемним, оскільки компанії, які не розкривають інформацію про свій борг розглядаються так само, як підприємства без боргів.

Другий підхід, заснований на моделі випадкового лісу та алгоритмі максимізації очікувань, показує, що доцільно припустити про багатовимірний нормальний розподіл вибірки даних із фінансовими коефіцієнтами компаній, яка має значно меншу розмірність у порівнянні зі звичайною базою балансових рахунків. Однак методика заповнення пропущених даних, заснована на алгоритмі

максимізації очікувань, має високу обчислювальну складність і погано масштабується у випадку наборів даних високої розмірності. Цей другий підхід, заснований на випадковому лісі, також демонструє досить високу точність, хоча відсутність повноти інформації для навчання класифікатора призводить до більшої похибки прогнозу.

Використовуючи результати роботи можна зробити деякі загальні висновки. Нейронні мережі є потужним інструментом, який також має перевагу у використанні набагато більшого навчального набору даних: у даній роботі нейронна мережа використовує 80% усієї бази даних для навчання, тоді як випадковий ліс навчається лише на 17% даних. Це пов'язано з наступними причинами: а) неможливо навчити випадковий ліс на даних із пропущеними значеннями; б) навчальний набір має бути хоча б частково збалансованим. Враховуючи таку незбалансованість вхідної інформації, метод EM-RF демонструє досить високу точність.

Крім цих висновків, пропонується ряд можливих напрямків продовження дослідження:

– аналіз альтернатив заміни пропусків даних на «0» під час навчання нейронної мережі: у даній роботі різні категорії відсутніх даних (NaN та «0») трактуються однаково. Однак, у деяких ситуаціях «0» повинен нести деяку інформацію, наприклад, той факт, що компанія не має довгострокового боргу.

– у даній роботі баланси компаній за різні роки вважаються самостійними записами, що нехтує історією конкретної компанії, яка включена у базу даних. Тому слід розглянути різні моделі навчання нейронної мережі, у яких еволюція кредитного рейтингу протягом декількох років буде врахована за допомогою 2D вхідних шарів, або, якщо доступне велике часове вікно (10+ років), згорткові та/або рекурентні нейронні мережі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Bao W., Yue J., and Rao Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PloS one, 2017.

2. Bartels B., Weder di Mauro, B. A rating agency for Europe a good idea? 2013.

3. Bassemir M. Why do private firms adopt IFRS? Accounting and Business Research, 2018. 237–263 p.

4. Beaver W. H. Financial ratios as predictors of failure. Journal of accounting research. 1966, 71 – 111 p.

5. Bishop C. M. Pattern recognition and machine learning. Springer, 2006.

6. Bond credit rating. *Wikipedia* : вебсайт. URL: https://en.wikipedia.org/wiki/Bond_credit_rating (дата звернення: 25.05.2022).

7. Cavalcante R. C., Brasileiro R. C., Souza V. L., Nobrega J. P., and Oliveira Chong E., Han C., and Park F. C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 2017, 83:187 – 205.

8. Ciampi F. and Gordini N. Using economic-financial ratios for small enterprise default prediction modeling: An empirical analysis. In 2008 Oxford Business & Economics Conference Proceedings, Association for Business and Economics Research (ABER), 2008. pages 1–21.

9. Decamps J.-P., Rochet J.-C., and Roger B. The three pillars of Basel ii: optimizing the mix. Journal of Financial Intermediation, 2004. 13(2):132 – 155.

10. Decision trees. *SQL Shack* : вебсайт. URL: https://www.sqlshack.com/microsoft-decision-trees-in-sql-server/ (дата звернення: 12.05.2022).

11. EM imputation algorithm. *Github* : вебсайт. URL: https://joon3216.github.io/research_materials/2019/em_imputation.html. Accessed: 2020-06-29 (дата звернення: 23.04.2022).

12. *Esma* : вебсайт. URL: https://www.esma.europa.eu/ (дата звернення: 19.03.2022).

13. Eurostat, N. Rev. 2–statistical classification of economic activities in the European community. Office for Official Publications of the European Communities. Luxemburg, 2008.

14. Haller A., Ernstberger J., Froschhammer M. Implications of the mandatory transition from national GAAP to IFRS – empirical evidence from Germany. Advances in Accounting, 2009. 25(2):226 – 236.

15. Harper S. Credit rating agencies deserve credit for the 2007-2008 financial crisis: An analysis of liability following the enactment of the dodd-frank act. Wash. & Lee L. Rev. 2011. 68:1925.

16. Horrigan J. O. The determination of long-term credit standing with financial ratios. Journal of Accounting Research, 1966. pages 44–62.

17. *IFRS foundation* : вебсайт. URL: https://www.ifrs.org/ (дата звернення: 29.03.2022).

18. IFRS. *Investopedia* : вебсайт. URL: https://www.investopedia.com/terms/i/ifrs.asp (дата звернення: 29.03.2022).

19. Lu H.-M., Tsai F.-T., Chen H., Hung M.-W., and Li S.-H. Credit rating change modeling using news and financial ratios. ACM Transactions on Management Information Systems (TMIS), 2012. 3(3):1 – 30.

20. Mattarocci G. The independence of credit rating agencies: How business models and regulators interact. Academic Press, 2013.

21. Meier, L., Van De Geer, S., and Bühlmann, P. The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008. 70(1):53 – 71.

22. Michalkova, L. and Kramarova, K. Capm model, beta and relationship with credit rating. In Advances in Applied Economic Research, 2017. Springer. pages 645–652.

23. Mitchell, T. M. et al. Machine learning. 1997. Burr Ridge, IL: McGraw Hill, 1997. 45(37):870 – 877.

24. ModeFinance. *Topvendors* : вебсайт. URL: https://www.topvendors.it /en/fintech/ modefinance (дата звернення: 12.05.2022).

25. Moon, T. K. The expectation-maximization algorithm. IEEE Signal processing magazine, 1996. 13(6):47 – 60.

26. Moya, S., Perramon Costa, J., and Constans, A. Ifrs adoption in europe: the case of Germany. 2005.

27. Nehrebecka, N. and Polski, N. B. Probability-of-default curve calibration and validation of internal rating systems.

28. Nelwamondo, F. V., Mohamed, S., and Marwala, T. Missing data: A comparison of neural network and expectation maximization techniques. Current Science, 2007. pages 1514–1521.

29. One-hot encoding. *Wikipedia* : вебсайт. URL: https://en.wikipedia.org/wiki/ One-hot.

30. PWC. Doing Business and Investing in Germany. PricewaterhouseCooper. 2018.

31. Segment reporting. *Accounting tools* : вебсайт. URL: https://www.accounting-tools.com/articles/what-is-segment-reporting.html (дата звернення: 12.04.2022).

32. SME default probabilities. *Investopedia* : вебсайт. URL: https:// www.investopedia.com/articles/financial-theory/11/default-probabilities-small-medium-corporations.asp (дата звернення: 12.04.2022).

33. Structural business statistics, SME. *Eurostat* : вебсайт. URL: https://ec.europa.eu/eurostat/web/structural-business-statistics/structural-business-statistics/sme (дата звернення: 19.03.2022).

34. Understanding random forest. *Towards Data Science* : вебсайт. URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (дата звернення: 22.05.2022).

35. VanderPlas, J. Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc., 2016.

36. Wahba G. and Wang Y. Representer theorem. Wiley StatsRef: Statistics Reference Online, 2014. pages 1–11.

37. White L. J. Markets: The credit rating agencies. Journal of Economic Perspectives, 2010. 24(2):211 – 26.

38. Yeh, C.-C., Lin, F., and Hsu, C.-Y. A hybrid KMV model, random forests and rough set theory approach for credit rating. Knowledge-Based Systems, 2012. 33:166 – 172.

39. Yeh, I.-C. and Lien, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 2019. 36(2):2473 – 2480.

40. Дії роботодавця, якщо у працівника виявлено коронавірус. *Охорона праці* : вебсайт. URL: https://oppb.com.ua/news/diyi-robotodavcya-yakshcho-u-pracivnyka-vyyavleno-koronavirus (дата звернення: 12.05.2022).

41. Постанови Кабінету міністрів України від 22.07.2020 р. №641. URL: https://zakon.rada.gov.ua/laws/show/641-2020-%D0%BF#Text (дата звернення: 01.06.2022).

42. Про загальнообов’язкове державне соціальне страхування : Закон України від 23.09.1999 р. №1105-XIV. URL: https://zakon.rada.gov.ua/laws/show/1105-14#Text (дата звернення: 01.06.2022).

43. Левченко, О. Г. Безпека життєдіяльності та цивільний захист : навч. посіб. Вид. 2-ге, переробл. і допов. Київ, 2019. 308 с.

44. Січко В. М., Тубальцев А.М. Безпека життєдіяльності : довідник. Житомир, 2017. 121 с.

ДОДАТОК А

Програмний код класу нейронної мережі

```
class Net(nn.Module):
    def __init__(self,n_inputs=10,n_outputs=1,*argv,**kwargs):
        super().__init__()

        self.net_depth = np.maximum(1,len(argv))
        self.n_inputs = n_inputs

        if not argv:
            self.n_layer_1 = 1
        else:
            for i,arg in zip(range(1,self.net_depth+1),argv):
                setattr(self, 'n_layer_'+str(i), arg)
        self.n_outputs = n_outputs

        self.lr = 0.001
        self.log_df = pd.DataFrame(columns=['Epoch','Loss Train','Loss Test'])
        self.reshuffle_in_epoch = True

        self.update_NN_structure()

        @property
        def net_depth(self):
            return self._net_depth
        @net_depth.setter
        def net_depth(self,value):
            if value > 5:
                raise AttributeError(f'NN is too deep')
            else:
                self._net_depth = value

        @property
        def n_inputs(self):
            return self._n_inputs
        @n_inputs.setter
        def n_inputs(self,value):
            self._n_inputs = value

        @property
        def n_layer_1(self):
            return self._n_layer_1
        @n_layer_1.setter
        def n_layer_1(self,value):
            self._n_layer_1 = value

        @property
        def n_layer_2(self):
            return self._n_layer_2
        @n_layer_2.setter
        def n_layer_2(self,value):
            self._n_layer_2 = value

        @property
        def n_layer_3(self):
            return self._n_layer_3
        @n_layer_3.setter
        def n_layer_3(self,value):
```

```
        self._n_layer_3 = value

    @property
    def n_layer_4(self):
        return self._n_layer_4
    @n_layer_4.setter
    def n_layer_4(self,value):
        self._n_layer_4 = value

    @property
    def n_layer_5(self):
        return self._n_layer_5
    @n_layer_5.setter
    def n_layer_5(self,value):
        self._n_layer_5 = value

    @property
    def n_outputs(self):
        return self._n_outputs
    @n_outputs.setter
    def n_outputs(self,value):
        self._n_outputs = value

    @property
    def fc1(self):
        return self._fc1
    @fc1.setter
    def fc1(self,value):
        self._fc1 = value

    @property
    def fc2(self):
        return self._fc2
    @fc2.setter
    def fc2(self,value):
        self._fc2 = value

    @property
    def fc3(self):
        return self._fc3
    @fc3.setter
    def fc3(self,value):
        self._fc3 = value

    @property
    def fc4(self):
        return self._fc4
    @fc4.setter
    def fc4(self,value):
        self._fc4 = value

    @property
    def fc5(self):
        return self._fc5
    @fc5.setter
    def fc5(self,value):
        self._fc5 = value

    @property
    def fc_output(self):
        return self._fc_output
```

```
        @fc_output.setter
        def fc_output(self,value):
            self._fc_output = value

        # lr = optimizer learning rate
        default_optimizer = optim.Adam(self.parameters(),self.lr)
        default_loss_function = nn.MSELoss()
        default_weights_dict = {new_list: .1 for new_list in range(1,11)}

        allowed_keys = {'EPOCHS':6,'BATCH_SIZE':200,'device': torch.device("cpu"),
\
                         'optimizer' : default_optimizer , 'loss_function' :
default_loss_function, \
                         'weights_dict' : default_weights_dict }#, 'VAL_PCT':0.25 ,
\

        # initialize all allowed keys
        self.__dict__.update((key, allowed_keys[key]) for key in allowed_keys)
        # and update the given keys by their given values
        self.__dict__.update((key, value) for key, value in kwargs.items() if key
in allowed_keys)


    def update_NN_structure(self):
        # defining first layer
        self.fc1 = nn.Linear(self.n_inputs, self.n_layer_1)
        # defining layers 2,3,etc.
        for i in range(1,self.net_depth):
            setattr(self, 'fc'+str(i+1), nn.Linear(getattr(self,
'n_layer_'+str(i)),getattr(self, 'n_layer_'+str(i+1) ) ) )
        # define last layer
        last_layer_width = getattr(self, 'n_layer_'+str(self.net_depth)  )
        self.fc_output = nn.Linear(last_layer_width, self.n_outputs)


    def forward(self,x):
  #      x = F.relu(self.fc1(x))
        iterate_idx = 1
        for attr in self._modules:
            if 'fc'+str(iterate_idx) in attr:
#                set_trace()
                x = F.relu(self._modules[attr](x))
                iterate_idx += 1
                if iterate_idx > self.net_depth:
                    break
        x = self.fc_output(x)
        return x #F.log_softmax(x,dim=1)  # we return the log softmax (sum of
probabilities across the classes = 1)

    # train the network
    def train_my_nn(self,train_X, train_y, test_X, test_y):

        test_X, test_y = test_X.to(self.device), test_y.to(self.device)

        for epoch in tqdm(range(self.EPOCHS)):

            if self.reshuffle_in_epoch:
               train_data = torch.cat((train_X, train_y),1)
               indexes = torch.randperm(train_data.shape[0])
               train_data = train_data[indexes]
               train_X, train_y = torch.split(train_data, [train_X.shape[1],1], 1)
```

```
            for i in range(0, len(train_X), self.BATCH_SIZE): # from 0, to the len
of x, stepping BATCH_SIZE at a time. [:50] ..for now just to dev
                #print(f"{i}:{i+BATCH_SIZE}")
                batch_X = train_X[i:i+self.BATCH_SIZE] #.view(-1, 1,
train_X.shape[1])
                batch_y = train_y[i:i+self.BATCH_SIZE]

                weights_torch =
torch.from_numpy(np.vectorize(self.weights_dict.get)(batch_y.numpy()))
                #set_trace()

                weights_torch = weights_torch.to(self.device)
                batch_X, batch_y = batch_X.to(self.device),
batch_y.to(self.device)

                self.zero_grad()

                outputs = self(batch_X)
              #  loss = self.loss_function(outputs, batch_y)
                loss = weighted_mse_loss(outputs, batch_y,weights_torch)
                loss.backward()
                self.optimizer.step()     # Does the update

            with torch.no_grad():
                eval_test = self(test_X)
                loss_test = self.loss_function(eval_test, test_y)

            new_row = [epoch, loss.item(), loss_test.item()]
            self.log_df.loc[len(self.log_df)] = new_row

            #print(f"Epoch: {epoch}. LossTrain: {loss}\n")
            #print(f"Epoch: {epoch}. LossTest : {loss_test}")

    def save_net_params(self,path_log,net_name):
        filename_pt = os.path.join(path_log, net_name + '.pt')
        filename_log = os.path.join(path_log, 'TrainingLog_'+ net_name + '.pkl')
        torch.save(self.state_dict(),filename_pt)
        self.log_df.to_pickle(filename_log)

    def load_net_params(self,path_log,net_name):
        filename_pt = os.path.join(path_log, net_name + '.pt')
        filename_log = os.path.join(path_log, 'TrainingLog_'+ net_name + '.pkl')
        self.log_df = pd.read_pickle(filename_log)
        # update net parameters based on state_dict() (which is loaded afterwards)
        self.n_inputs =  (torch.load(filename_pt)['fc1.weight']).shape[1]
        self.net_depth = int(len(torch.load(filename_pt))/2)-1
        for i in range(1,self.net_depth+1):
            setattr(self, 'n_layer_'+str(i), (torch.load(filename_pt)['fc' +
str(i) + '.weight']).shape[0])
        self.n_outputs =  (torch.load(filename_pt)['fc_output.weight']).shape[0]
        self.update_NN_structure()
        self.load_state_dict(torch.load(filename_pt))
        self.to(self.device)  # required step when loading a model on a GPU (even
if it was also trained on a GPU)

    def plot_training_log(self, init_epoch=0, final_epoch =-1):
        log_norm_df=(self.log_df-self.log_df.min())/(self.log_df.max()-
self.log_df.min())
        #fig, ax = plt.subplots()
        fig = plt.figure(figsize=(12, 12))
```

```
        #ax = fig.gca()
        ax = log_norm_df[init_epoch:final_epoch][['Loss Train','Loss
Test']].plot()
        fig = ax.get_figure()
        return fig


    def run_test_prediction(self, test_X, test_y, condensed_classes = False,
print_acc = True):
        if not torch.is_tensor(test_X):
            test_X = torch.tensor(test_X.astype(np.float32))
            test_y = torch.tensor(test_y.astype(np.float32))

        test_X = test_X.to(self.device)
        y_pred = self(test_X).to(self.device)[:,0].view(test_X.shape[0],-1)

        device_cpu = torch.device("cpu")
        y_pred.to(device_cpu)
        pred_y = np.round(y_pred.data.cpu().numpy()).astype(float)

        y_min = 1
        y_max = 10
        if condensed_classes:
            y_min = 3
            y_max = 8

        pred_y = np.minimum(np.maximum(y_min,pred_y), y_max)

        correct_total = [pred_y[i] == test_y.numpy()[i] for i in
range(len(pred_y))]
        accuracy = np.sum(correct_total)/len(test_y)

        if print_acc:
            print(f'Accuracy = ' + str(round(accuracy*100,2)) +'%')

        return pred_y, accuracy
```