Київський національний університет імені Тараса Шевченка

Економічний факультет

Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

ВИКОРИСТАННЯ ІНСТРУМЕНТІВ DATA SCIENCE НА ПІДПРИЄМСТВАХ (НА ПРИКЛАДІ ЛОМБАРДА «СКАРБНИЦЯ»)

Студента 2 курсу магістратури
спеціальності 051 «Економіка»
ОНП «Економічна кібернетика»
денної форми навчання
Іваницького Максима Вадимовича

Науковий керівник:
д.е.н., проф. Затонацька Тетяна Георгіївна

Засвідчую, що в цій роботі
немає запозичень із праць інших авторів
без відповідних посилань

Студент _____
(підпис)

Робота допущена до захисту в ЕК
рішенням кафедри економічної кібернетики
від 4 травня 2022 р., протокол № 13

Завідувач кафедри:
доктор економічних наук, професор
Ляшенко Олена Ігорівна
_____
(підпис)

Київ - 2022

Київський національний університет імені Тараса Шевченка

Економічний факультет

Кафедра економічної кібернетики

ЗАВДАННЯ
на кваліфікаційну роботу магістра
студента 2 курсу спеціальності 051 «Економіка», ОНП
«Економічна кібернетика»
Іваницького Максима Вадимовича

1. Тема роботи: «Використання інструментів Data Science на підприємствах (на прикладі ломбарда «Скарбниця»)
2. Термін завершення роботи: 12.05.2022
3. Предмет дослідження: методи та підходи щодо моделювання рекомендаційної системи на підприємстві.
4. Об'єкт дослідження: використання штучного інтелекту на підприємстві.
5. Мета дослідження: полягає у розкритті методів та інструментів Data Science, які використовуються на підприємстві та дослідження розробки системи DS на прикладі конкретної компанії
6. Завдання дослідження:
   6.1. Розкриття теоретичних засад побудови системи Data Science на підприємстві;
   6.2. Дослідження галузей обробки даних;
   6.3. Аналіз ринку ломбардних послуг;
   6.4. Аналіз бізнес - показників досліджуваного підприємства;
   6.5. Розробка Data Science системи для прогнозу продажів підприємства.

Науковий керівник: д.е.н, професор Затонацька Тетяна Георгіївна _____

Студент: Іваницький Максим Вадимович _____

Затверджено на засіданні кафедри економічної кібернетики
протокол №3 від 12 жовтня 2021 р.

Календарний план виконання кваліфікаційної роботи магістра

№	Етапи роботи	Терміни виконання	Відмітка керівника про виконання
1	Вибір теми кваліфікаційної роботи магістра	01.09.2021 – 01.11.2021	
2	Розробка та затвердження завдання кваліфікаційної роботи магістра	01.11.2021 – 01.12.2021	
3	Збір інформації, її аналіз, обробка, консультації з науковим керівником	01.12.2021 – 02.02.2022	
3	Написання розділу 1	02.02.2022 – 14.03.2022	
4	Написання розділу 2	21.03.2022 – 07.04.2022	
5	Написання розділу 3	16.04.2022 – 30.04.2022	
6	Написання вступу та висновків	До 10.05.2020	
7	Подання роботи до попереднього захисту	До 12.05.2020	
8	Захист магістерської роботи	24.05.2022	

Науковий керівник:……..

Студент:……...

РЕФЕРАТ

Кваліфікаційна робота магістра містить: 44 ст., 15 рис., 12 табл., 43 джерел

Ключові слова: наука про дані, аналіз даних, інтелектуальний аналіз даних, машинне навчання, дані, прогноз, магазин техніки, ломбард.

Об’єкт дослідження: використання штучного інтелекту на підприємстві.

Мета дослідження: полягає у розкритті методів та інструментів Data Science, які використовуються на підприємстві та дослідження розробки системи DS на прикладі конкретної компанії.

Методи дослідження: аналіз та синтез, індукція та дедукція, комплексний та системний підхід, економіко-математичне моделювання, прогнозне моделювання.

Наукова новизна, теоретична значимість дослідження полягає у авторському підході до аналізу та побудови системи Data Science для підприємства.

Практична цінність полягає у тому, що результати робити будуть використовуватися на розглянутому підприємстві та можуть бути використані на інших підприємствах, які пов’язані зі сферою продажу техніки.

RESUME

Kyiv National Taras Shevchenko University,

Faculty of Economics, Department of Economic Cybernetics

Key words: Data Science, Data Analytics, Data Minnig, Machine Learning, data, forecast, tech store, pawnshop.

The graduation research of student use Maksym Ivanytskiy «Using Data Science tools in businesses (the case of Skarbnytsya pawnshop)»
deals with research, which consists in the author's approach to the analysis and construction of the Data Science system for the enterprise.

The work is interesting for the results will be used at the enterprise in question and can be used at other enterprises related to the sale of equipment.

Pages 44, tables 12, bibliog 43, append. 1

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

BD – великі дані (англ. Big Data)

DA – аналіз даних (англ. Data Analytics)

DS – наука про дані (англ. Data Science)

DM – інтелектуальний аналіз даних (англ. Data Minnig)

ML – машинне навчання (англ. Machine Learning)

НКО – небанківські кредитні організації

ВСТУП

Чому Data Science важливо, тому що до 2025 року буде близько 40 зеттабайт даних — це 40 трильйонів гігабайт. Кількість наявних даних зростає в геометричній прогресії. У будь-який час близько 90 відсотків цієї величезної кількості даних генерується протягом останніх років.[37]

Фактично, щодня користувачі Інтернету генерують близько 2,5 квінтильйонів байт даних. До 2025 року кожна людина на Землі буде генерувати близько 146 880 ГБ даних щодня, а до 2030 року це буде 165 зеттабайт щороку.

Це означає, що в науці про дані є величезна робота, і багато чого залишилося розкрити. У 2015 році було проаналізовано лише близько 0,5 відсотка всіх даних.

Простий аналіз даних може інтерпретувати дані з одного джерела або з обмеженої кількості. Однак інструменти DS мають вирішальне значення для розуміння великих даних та даних із кількох джерел у змістовний спосіб. Погляд на деякі конкретні прикладні програми DS в бізнесі ілюструють цю думку і забезпечують переконливий вступ до DS.

У міру розвитку галузі ми можемо очікувати, що побачимо кілька тенденцій, які формують майбутнє DS. По-перше, більше завдань із науки про дані в життєвому циклі, ймовірно, стануть автоматизованими. Ця зміна буде спричинена тиском на збільшення рентабельності інвестицій, оскільки все більше компаній інвестують у машинне навчання та штучний інтелект. З більшою кількістю автоматизованих процесів DS, більше даних стане доступним для більшої кількості людей у більшій кількості вертикалей, а штучний інтелект та машинне навчання також мають розвиватися швидше.

Інше зрушення може відбутися у вигляді ресурсів DS, які будуть більш доступними для більшої кількості людей. Дослідники даних зазвичай мають дуже специфічні навички. Однак як попит на людей, які можуть кваліфіковано виконувати завдання з DS, так і на професіоналів, які керують ініціативами в галузі штучного інтелекту та машинного навчання, зокрема, стрімко зростає. Це зростання, у свою чергу, стимулює тенденцію до громадянської науки у вертикалі.

Це особливо вірогідно в нішевих сферах бізнесу, які вимагають високого рівня знань в області або галузі. Як і в інших наукових дисциплінах, більш складні операції можуть бути зарезервовані для науковців з DS з більш специфічною підготовкою, але менш розріджені завдання будуть рухатися в напрямку доступності. Буде цікаво побачити, скільки ще вертикалей, де використовуватиметься DS, відкриється, оскільки автоматизація прокладає шлях.

Наступна тенденція, яка, ймовірно, сформує майбутнє DS – це напруженість між правом на конфіденційність, необхідністю регулювання та вимогою прозорості. DS здатна зробити алгоритми машинного навчання та процес, за допомогою якого ми навчаємо штучний інтелект, набагато прозорішими, що, у свою чергу, може зробити можливим регуляторний нагляд.

Чому DS в епоху автоматизації? Питання, чи буде DS автоматизованою, є тривалими дискусіями. Хоча багато хто задає питання: «Чи помре наука про дані», краще запитати: «Як DS зміниться з автоматизацією?»

Експерти, вважають, що прогрес у візуалізації даних та обробці природної мови (NLP) означатиме, що дані незабаром оброблятимуться автоматично – по суті, набагато більше людей зможуть збирати інформацію з даних завдяки розширеній аналітиці та іншим технологіям DS.

Оскільки постійно генерується така велика кількість даних, спрощення використання продуктів із DS для вчених, що працюють у сфері даних, лише покращує охоплення компаній, які працюють у космосі. Місце для автоматизації в DS – це інтенсивні, повторювані вручну 101 завдання, які не вимагають глибшого навчання та досвіду.

Наразі розумний погляд на автоматизацію DS полягає в тому, що простіші завдання можуть і будуть автоматизовані. Однак людське управління алгоритмами та аналітикою залишатиметься важливим, оскільки здатність перетворювати людські потреби в бізнес-питання та стратегії ще далека від автоматизації.

Здатність отримувати корисні ідеї зі складних даних, що вимагають автоматизації критичного мислення, що залежить від контексту, – ще попереду. Крім того, фахівці DS з глибоким досвідом ведення бізнесу та помітною галузевою кмітливістю будуть продовжувати бачити високий попит на свої навички.

Незважаючи на те, що ручні завдання, пов’язані з даними, можуть бути автоматизованими, розумні вчені з аналітичними навичками, які володіють даними, будуть більш затребувані в 21 столітті. Кар'єра в галузі DS нікуди не зникне.

Об'єктом дослідження є використання штучного інтелекту на підприємстві.

Предметом дослідження є методи та підходи щодо моделювання рекомендаційної системи на підприємстві.

Мета роботи полягає у розкритті методів та інструментів Data Science, які використовуються на підприємстві та дослідження розробки системи DS на прикладі конкретної компанії.

Завдання полягає у наступному:

- розкриття теоретичних засад побудови системи Data Science на підприємстві
- дослідження галузей обробки даних
- аналіз ринку ломбардних послуг
- аналіз бізнес - показників досліджуваного підприємства
- розробка DS системи для прогнозу продажів підприємства

Відповідно до мети реалізація поставлених завдань зумовила необхідність використання таких загальнонаукових і специфічних методів дослідження: наукової астракції, єдності історичного та логічного, аналізу та синтезу, класифікації та узагальнення, з позицій яких була проаналізована наукова література; емпіричні методи соціологічних досліджень; статистичні – методи математичної статистики для обробки отриманих даних; графічні – для візуалізації результатів дослідження; моделювання – використання мови програмування Phyton.

РОЗДІЛ 1

Теоретичні засади побудови системи Data Science на підприємстві

1.1 Поняття та теоретичні засади Data Science

Data Science дозволяє підприємствам обробляти величезні обсяги структурованих і неструктурованих великих даних для виявлення закономірностей. Це, у свою чергу, дозволяє компаніям підвищувати ефективність, керувати витратами, визначати нові можливості та збільшувати свої переваги на ринку.

Data Science – це процес видобутку великих наборів необроблених даних, як структурованих, так і неструктурованих, для виявлення закономірностей та отримання корисної інформації. Це міждисциплінарна галузь, і основи Data Science включають статистику, висновки, інформатику, прогнозну аналітику, розробку алгоритмів машинного навчання та нові технології для отримання інформації з великих даних.[4]

Щоб визначити DS та покращити управління проектами науки про дані, почніть з його життєвого циклу. Перший етап робочого процесу конвеєра DS включає захоплення: отримання даних, іноді їх вилучення та введення в систему. Наступним етапом є технічне обслуговування, яке включає зберігання даних, очищення даних, обробку даних, постановку даних та архітектуру даних.

Далі обробка даних є однією з основ DS. Цей етап включає інтелектуальний аналіз даних, класифікацію та кластеризацію даних, моделювання даних та узагальнення інформації, отриманої з даних, процесів, які створюють ефективні дані.

Далі йде аналіз даних, не менш важливий етап. Тут спеціалісти DS проводять дослідницьку та підтверджувальну роботу, регресію, прогнозний аналіз, якісний аналіз та аналіз тексту.

Під час останнього етапу науковець з даних передає ідеї. Це включає візуалізацію даних, звітування про дані, використання різних інструментів бізнес-аналітики та надання допомоги підприємствам, політикам та іншим особам у прийнятті більш розумних рішень.

Підготовка та аналіз даних є найважливішими навичками DS, але лише підготовка даних зазвичай займає від 60 до 70 відсотків часу. Рідко дані генеруються у виправленій, структурованій формі. На цьому етапі дані трансформуються і готуються для подальшого використання.

Ця частина процесу включає перетворення та вибірку даних, перевірку як характеристик, так і спостережень, а також використання статистичних методів для видалення перешкод. Цей крок також висвітлює, чи функції в базі даних незалежні один від одного, і чи можуть бути відсутні значення в даних.

Цей етап дослідження також є принциповою відмінністю між DS та аналітикою даних. DS використовує макро-погляд, щоб сформулювати кращу аналітику про дані, щоб отримати з них більше інформації та знань.

На етапі моделювання спеціалісти з DS вписують дані в модель за допомогою алгоритмів машинного навчання. Вибір моделі залежить від типу даних і вимог бізнесу.

Далі модель тестується для перевірки її точності та інших характеристик. Це дає змогу спеціалісту з даних коригувати модель для досягнення бажаного результату. Якщо модель не відповідає вимогам, команда може вибрати будь-яку іншу модель DS.

Як тільки правильне тестування з хорошими даними дає бажані результати, модель можна завершити та розгорнути.

1.2. Порівняння Data Science з іншими галузями обробки даних

1. Data Science та Data Analytics

Хоча роботи спеціалістів з DS і DA іноді поєднують, але ці галузі не є однаковими. Термін DS насправді означає те й інше.

Спеціаліст з DS приходить раніше в проєкт, ніж аналітик даних, досліджуючи величезний набір даних, потенціал, виявляючи тенденції та ідеї для візуалізації. Спеціаліст з DA бачить дані на пізнішому етапі. Вони роблять звіти для кращої продуктивності на основі свого аналізу та оптимізують будь-які інструменти, пов’язані з даними.

Аналітик даних, імовірно, аналізує певний набір структурованих або числових даних, використовуючи задане запитання або запитання. Дослідник даних, швидше за все, буде працювати з більшими масивами як структурованих, так і неструктурованих даних. Вони також сформулюють, перевірять та оцінять результативність питань даних у контексті загальної стратегії.[12]

Аналітика даних більше пов’язана з розміщенням історичних даних у контексті та менше з прогнозним моделюванням, машинним навчанням. Аналіз даних не є відкритим пошуком правильного питання; він покладається на наявність правильних запитань на місці з самого початку. Крім того, на відміну від DS, аналітики даних зазвичай не створюють статистичні моделі та не навчають інструментів машинного навчання.

Натомість спеціаліст DA зосереджуються на стратегії для бізнесу, порівнюючи активи даних з різними організаційними гіпотезами чи планами. Аналітики даних також частіше працюють із локалізованими даними, які вже оброблені. Навпаки, як технічні, так і нетехнічні навички вивчення даних є важливими для обробки даних, а також їх аналізу. Звичайно, обидві ролі вимагають математичних, аналітичних і статистичних навичок.

Аналітики даних мають менше потреби в ширшому підході до бізнес-культури у своїй повсякденній роботі. Натомість вони, як правило, використовують більш виміряний, прибитий фокус, коли аналізують фрагменти даних. Їх масштаби та цілі будуть більш обмеженими, ніж у спеціаліста DS.

Підсумовуючи, можна сказати, що спеціалісти DS, швидше за все, дивитимуться вперед, передбачаючи чи прогнозуючи, коли вони переглядають дані. Зв'язок між аналітиком даних і даними є ретроспективним. Спеціаліст з DA, швидше за все, зосередиться на конкретних питаннях, щоб відповісти на них, аналізуючи наявні набори даних, які вже були оброблені для отримання інформації.

2. Big Data та Data Science

Дані надходять з різних джерел, таких як онлайн-покупки, мультимедійні форми, інструменти, фінансові журнали, датчики, текстові файли тощо. Дані можуть бути неструктурованими, напівструктурованими або структурованими.

Неструктуровані дані включають дані з блогів, цифрових аудіо/відео каналів, цифрових зображень, електронної пошти, мобільних пристроїв, датчиків, соціальних мереж і твітів, веб-сторінок та онлайн-джерел. Напівструктуровані дані включають дані з файлів системного журналу, файлів XML і текстових файлів. Структуровані дані, які вже були певним чином оброблені, включають OLTP, RDBMS (бази даних), дані транзакцій та інші формати.[11]

Це все Big Data, і використання цих даних на належному рівні є нагальною роботою 21 століття. Просто неможливо обробити величезні обсяги даних із різних джерел за допомогою простих інструментів бізнес-аналітики чи навіть інструментів аналізу даних. Натомість DS надає підприємствам передові, складні алгоритми та інші інструменти для аналізу, очищення, обробки та вилучення значущих даних.

DS – це не один інструмент, навичка чи метод. Натомість це науковий підхід, який використовує прикладну статистичну та математичну теорію та комп’ютерні засоби для обробки великих даних.

Основи DS поєднують міждисциплінарні переваги очищення даних, інтелектуальних методів збору даних, аналізу даних та програмування. Результатом є здатність дослідника даних отримувати, підтримувати та готувати великі дані для інтелектуального аналізу.

Це один момент, який відрізняє роботу спеціаліста з обробки даних від інженера з даними, хоча іноді ці дві ролі плутають. Інженер з даних готує набори даних для роботи з DS і для отримання інформації. Але робота з інтелектуального аналізу покладається на фахівців DS, а не «інженерів з даних».

Великі дані є сировиною, яка використовується в галузі DS. Відрізняючись швидкістю, різноманітністю та обсягом (3V), великі дані є сировиною для DS, яка надає методи аналізу даних.

3. Data Mining та Data Science

Data Mining – це інтелектуальний аналіз даних, який використовується як у бізнесі, так і в DS, тоді як DS – це фактична галузь наукового дослідження чи дисципліни. Мета DM - зробити дані більш придатними для певних бізнес-цілей. DS, навпаки, спрямована на створення продуктів і результатів на основі даних — як правило, у бізнес-контексті.

DM має справу переважно зі структурованими даними, оскільки дослідження величезної кількості необроблених даних знаходиться в межах DS. Однак інтелектуальний аналіз даних є частиною того, що можуть робити фіхавці з DS, і це вміння, яке є частиною науки.

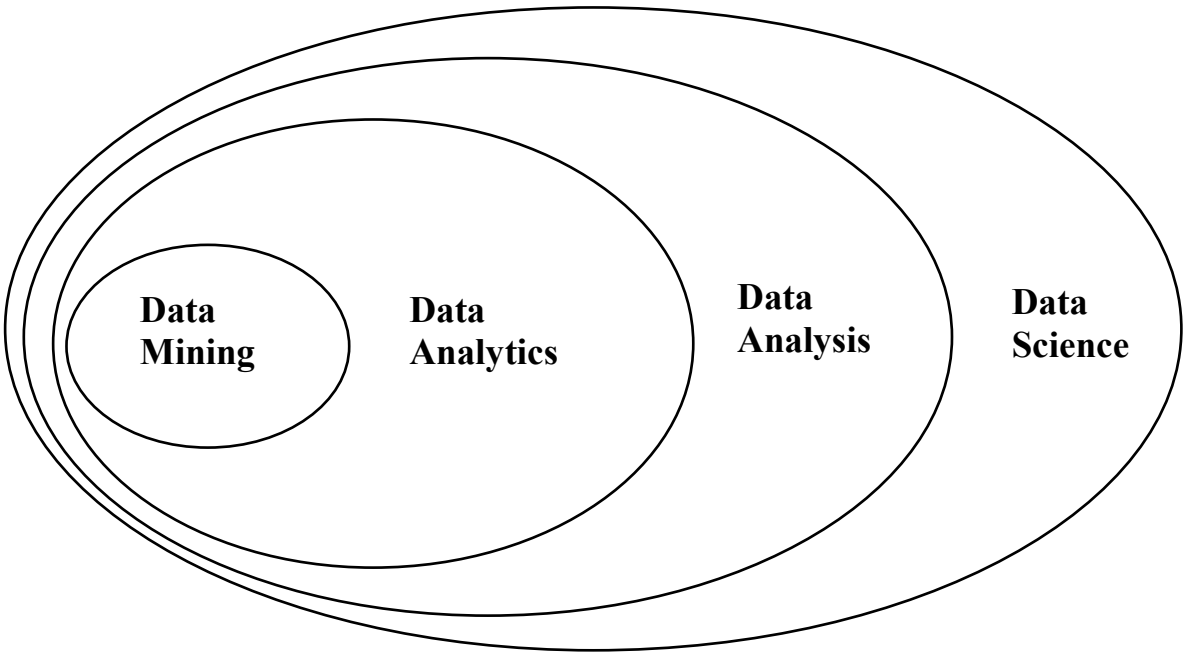

Рис 1.1. Взаємоз’язок між галузями обробки даних

Джерело: розроблено автором

Зв’язок між DS та Machine Learning

DS та ML працюють у тандемі. Machine Learning - це область науки про дані, яка подає комп’ютерам величезні обсяги даних, щоб вони могли навчитися приймати глибокі рішення, подібні до того, як це роблять люди.

Наприклад, більшість людей в дитинстві дізнаються, що таке квітка, не замислюючись про це. Однак людський мозок досягає цього завдяки досвіду — шляхом збору даних – про особливості, пов’язані з квітами.

Машина може зробити те ж саме за допомогою людини. Коли люди подають машині величезну кількість даних, вона може дізнатися, що різні пелюстки, стебла та інші елементи пов’язані з квітами.

Іншими словами, люди передають на машину дані про навчання або вихідні дані, щоб вона могла вивчати всі пов’язані з даними функції. Потім, якщо навчання було успішним, тестування з новими даними повинно виявити, що машина може розрізняти особливості, які вона засвоїла. Якщо ні, йому потрібно більше або краще навчання.

DS є природним розширенням статистики. Він розвивався разом із інформатикою, щоб обробляти величезні обсяги даних за допомогою нових технологій.

Навпаки, ML є частиною науки про дані, але це більше процес. Машинне навчання дозволяє комп’ютерам вчитися – і робити це з часом більш ефективно без явних програм для кожного фрагмента інформації.

У машинному навчанні комп’ютери використовують алгоритми для навчання, але ці алгоритми покладаються на деякі вихідні дані. Машина використовує ці дані як навчальний набір, тому може покращувати свій алгоритм, налаштовуючи та тестуючи його, оптимізуючи по ходу. Таким чином він налаштовує різні параметри своїх алгоритмів DS, використовуючи різні статистичні методи, включаючи регресію та контрольовану кластеризацію.[34]

Однак інші методи, які вимагають участі людини, також є частиною науки про дані, як ми її розуміємо сьогодні. Наприклад, машина може навчити іншу машину виявляти структури даних за допомогою неконтрольованої кластеризації для оптимізації алгоритму класифікації. Але щоб повністю завершити процес, людина все ще повинна класифікувати структури, які ідентифікує комп’ютер – принаймні до тих пір, поки вона не буде повністю навчена.[34]

Сфера DS також виходить далеко за межі машинного навчання, охоплюючи дані, які генеруються не будь-яким механічним процесом, комп'ютером чи машиною. Наприклад, DS також включає дані опитувань, дані клінічних випробувань або будь-які інші наявні дані.

DS також передбачає розгортання даних не лише для навчання машин. Далеко не обмежуючись питаннями статистичних даних, область DS, безумовно, включає в себе автоматизацію машинного навчання та прийняття рішень на основі даних. Однак він також охоплює інтеграцію даних, розробку даних та візуалізацію даних, а також розподілену архітектуру та створення інформаційних панелей та інших інструментів бізнес-аналітики. Фактично, будь-яке розгортання даних у виробничому режимі також входить до сфери DS.

Отже, коли фахівці з DS створюють ідеї, які вони отримують з даних, машина навчається на основі тих уявлень, які вже були сприйняті науковцем із даних. І хоча машина може будувати власне розуміння існуючої алгоритмічної структури, відправна точка спирається на якісь структуровані дані.

Можна зробити висновок, що фахівець з DS повинен розуміти ML, яке використовує багато методів науки про дані. Але «дані» для вченого можуть включати або не включати дані механічного процесу.

1.3. Поняття життєвого циклу Data Science – проєкту

Data Science швидко розвивається і стає однією з найпопулярніших галузей в технологічній індустрії. Завдяки швидкому прогресу в обчислювальній продуктивності, який тепер дозволяє аналізувати масивні набори даних, ми можемо виявити закономірності та уявлення про поведінку користувачів і світові тенденції.

Для кращого розуміння процесів роботи у сфері DS на рис. 1.2 представлено сім кроків, які складають життєвий цикл DS: розуміння бізнесу, інтелектуальний аналіз даних, очищення даних, дослідження даних, конструювання ознак, прогнозне моделювання та візуалізація даних. [43]

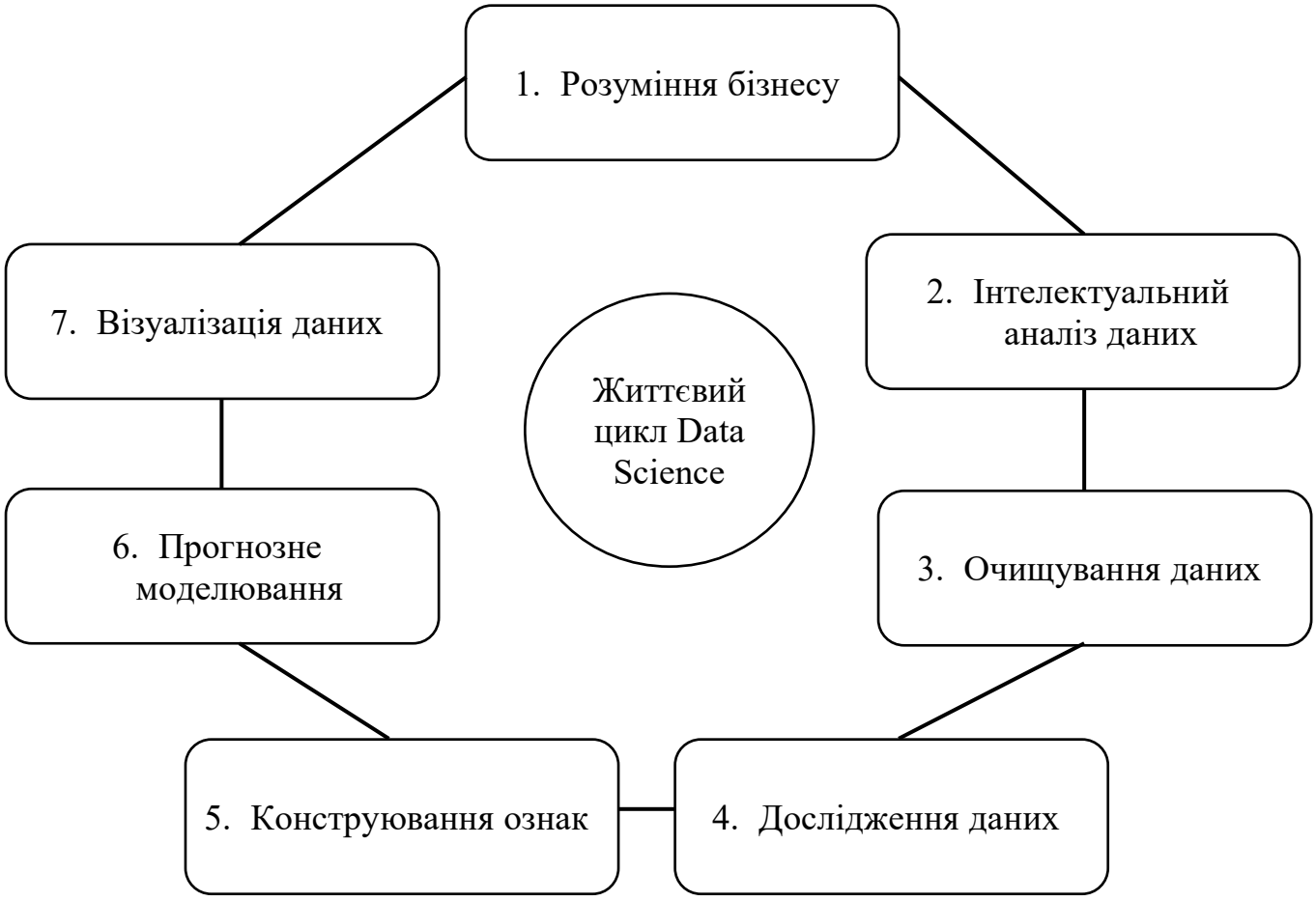

Рис 1.2. Життєвий цикл Data Science

Джерело: розроблено автором

1. Розуміння бізнесу

Фахівці з DS – це люди, які постійно запитують, чому. Це ті люди, які хочуть переконатися, що кожне рішення, прийняте в компанії, підкріплюється конкретними даними, і що воно гарантовано (з високою ймовірністю) досягне результатів. Перш ніж розпочати проект із DS, дуже важливо зрозуміти проблему, яку намагаєтесь вирішити.

Зазвичай використовують DS, щоб відповісти на п’ять типів запитань, які представлені у таблиці 1.1.

Таблиця 1.1

Типи запитань, які використовують для відповіді у Data Science

Питання	Метод який використовується
Скільки?	Регресія

Яка категорія?	Класифікація
Яка група?	Кластеризація
Це дивно?	Виявлення аномалій
Який варіант варто прийняти?	Рекомендації

Джерело: розроблено автором

На цьому етапі потрібно визначити головні цілі вашого проекту, визначивши змінні, які необхідно передбачити. Якщо це регресія, це може бути щось на зразок прогнозу продажів. Якщо це кластеризація, це може бути профіль клієнта. Розуміння сили даних і того, як ви можете використовувати їх для отримання результатів для свого бізнесу, ставлячи правильні запитання, – це більше мистецтво, ніж наука, і для того, щоб зробити це добре, необхідно мати великий досвід.

2. Інтелектуальний аналіз даних

Коли визначили цілі свого проекту, настав час почати збирати дані. Інтелектуальний аналіз даних – це процес збору ваших даних з різних джерел. Можна групувати пошук і очищення даних разом, але кожен із цих процесів є настільки значущим кроком, що краще розділити їх на частини. На цьому етапі варто розглянути деякі з питань:

- Які дані мені потрібні для цього проекту?
- Де їх зібрати?
- Як я можу їх отримати?
- Який найефективніший спосіб зберігати й отримувати доступ до всього цього?

Якщо всі дані, необхідні для проекту, запаковані та передані вам, ви виграли в лотерею. Найчастіше пошук потрібних даних вимагає як часу, так і зусиль. Якщо дані зберігаються в базах даних, ваша робота відносно проста – ви можете знаходити відповідні дані за допомогою SQL-запитів або маніпулювати ними за допомогою інструмента фрейму даних, наприклад Pandas. Однак, якщо ваші дані насправді не існують у наборі даних, вам потрібно буде їх очистити. Beautiful Soup – популярна бібліотека, яка використовується для пошуку даних на веб-

сторінках. Якщо ви працюєте з мобільним додатком і хочете відстежувати взаємодію та взаємодію користувачів, у програму можна інтегрувати незліченну кількість інструментів, щоб ви могли почати отримувати цінні дані від клієнтів. Google Analytics, наприклад, дозволяє визначати спеціальні події в програмі, які допоможуть зрозуміти, як поводяться ваші користувачі, і збирати відповідні дані.

3. Очищування даних

Після того як отримали всі свої дані, ми переходимо до найбільш трудомісткого кроку з усіх – очищення та підготовка даних. Особливо це стосується проєктів Big Data, які часто передбачають терабайти даних для роботи. Робота з даними часто займає від 50 до 80 відсотків часу.

Причина, чому це такий трудомісткий процес, полягає в тому, що існує дуже багато можливих сценаріїв, які можуть вимагати очищення. Наприклад, дані також можуть мати невідповідності в одному стовпці, тому що деякі рядки можуть мати мітку 0 або 1, а інші – ні або так. Типи даних також можуть бути неузгодженими - деякі з 0 можуть бути цілими числами, тоді як інші можуть бути рядками. Якщо ми маємо справу з категоріальним типом даних із кількома категоріями, деякі категорії можуть бути написані з помилкою. Це лише підмножина прикладів, де ви можете побачити невідповідності, і на цьому етапі важливо знайти та виправити їх.

Один із кроків, про який на цьому етапі часто забувають, що згодом спричиняє багато проблем, – це наявність відсутніх даних. Відсутні дані можуть викликати багато помилок у створенні та навчанні моделі. Один із варіантів – ігнорувати екземпляри, які мають відсутні значення. Залежно від вашого набору даних це може бути нереальним, якщо у вас багато відсутніх даних. Іншим поширеним підходом є використання того, що називається імпутація, яка замінює пропущені значення середнім значенням для всіх інших випадків. Це не завжди рекомендується, оскільки це може зменшити мінливість ваших даних, але в деяких випадках це має сенс.

4. Дослідження даних

На наступному етапі, коли у вас є блискучий чистий набір даних, можливо розпочати свій аналіз. Етап дослідження даних схожий на мозковий штурм аналізу даних. Саме тут ви розумієте закономірності та упередження ваших даних. Це може включати витяг і аналіз випадкової підмножини даних за допомогою Pandas, побудову гістограми або кривої розподілу, щоб побачити загальну тенденцію, або навіть створення інтерактивної візуалізації, яка дозволить вам зануритися в кожну точку даних і досліджувати історію, яка стоїть за викидами.[12]

Використовуючи всю цю інформацію, можна формувати гіпотези про свої дані та проблему, яку ви вирішуєте. Наприклад, якщо ви передбачали результати учнів, ви можете спробувати уявити зв'язок між балами та сном. Якби ви передбачали ціни на нерухомість, ви могли б нанести ціни як теплову карту на просторовому графіку, щоб побачити, чи зможете ви вловити якісь тенденції.

5. Конструювання ознак

У машинному навчанні ознака – це вимірна властивість або атрибут явища, що спостерігається. Якби ми передбачали бали учня, можливою ознакою є кількість сну, яку він отримує. У складніших завданнях прогнозування, таких як розпізнавання символів, ознаками можуть бути гістограми, які підраховують кількість чорних пікселів.

Конструювання ознак – це процес використання предметних знань для перетворення ваших необроблених даних в інформаційні характеристики, які представляють бізнес-проблеми, які ви намагаєтеся вирішити. Цей етап безпосередньо вплине на точність прогнозної моделі, яку ви побудуєте на наступному етапі.

Зазвичай ми виконуємо два типи завдань у інженерії об'єктів – вибір об'єктів та конструювання. Вибір функцій – це процес скорочення ознак, які додають більше шуму, ніж інформації. Зазвичай це робиться, щоб уникнути розмірності, яке відноситься до підвищеної складності, яке виникає через простори високої розмірності (тобто занадто багато ознак).

Наприклад, якщо у вас є показник віку, але ваша модель цікавиться лише тим, чи є особа дорослою чи неповнолітньою, ви можете встановити її на рівні 18 і призначити різні категорії екземплярам вище або нижче цього порогу. Також можете об’єднати кілька функцій, щоб зробити їх більш інформативними, взявши їх суму, різницю або добуток. Наприклад, якщо ви передбачали результати учнів і мали функції для кількості годин сну щоночі, ви можете створити функцію, яка позначатиме середній сон, який спав студент.

6. Прогнозне моделювання

Прогнозне моделювання – це те, де машинне навчання нарешті входить у ваш проект DS. На основі запитань, які ви задавали на етапі розуміння бізнесу, ви вирішуєте, яку модель вибрати для своєї проблеми. Модель (або моделі), які в кінцевому підсумку навчаєте, залежатиме від розміру, типу та якості ваших даних, скільки часу та обчислювальних ресурсів ви готові інвестувати, а також від типу результату, який маєте намір отримати.[19]

Після того, як ви навчили свою модель, дуже важливо оцінити її успіх. Для вимірювання точності моделі зазвичай використовується процес, який називається k-кратною перехресною валідацією. Це включає в себе розділення набору даних на k груп екземплярів однакового розміру, навчання всіх груп, крім однієї, і повторення процесу з різними групами, які не враховуються. Це дає змогу навчати модель на всіх даних замість використання типового розбиття тесту.

Для моделей класифікації часто перевіряється точність за допомогою PCC (percent correct classification) разом із матрицею плутанини, яка розбиває помилки на хибнопозитивні та хибнонегативні результати. Для порівняння успіху моделі також використовуються такі графіки, як криві ROC, що є істинним позитивним показником, наведеним на графік проти хибнопозитивного показника. Для регресійної моделі загальні показники включають коефіцієнт детермінації (який дає інформацію про відповідність моделі), середню квадратичну помилку (MSE) і середню абсолютну помилку.[40]

7. Візуалізація даних

Візуалізація даних є складною областю, можливо, це одна з найважчих речей, які можна зробити добре. Це тому, що дані, поєднують галузі комунікації, психології, статистики та мистецтва, з кінцевою метою передачі даних простим, але ефективним і візуально доступним способом. Після того, як ви отримали передбачувані результати своєї моделі, ви повинні представити їх таким чином, щоб різні ключові зацікавлені сторони проекту могли зрозуміти.

Дослідники даних представляють дані у вигляді графіків, діаграм та інших візуалізацій. Ці візуалізації даних дозволяють користувачам «бачити» статистику, невидиму в аркушах даних Excel. Наприклад, ви можете зобразити, як певні тенденції в даних пов’язані один з одним або як збігаються кілька факторів.

Середовища візуалізації даних – це звичайний спосіб розгортання результатів DS для широкої аудиторії, наприклад, за допомогою веб-інструментів, які дозволяють досліджувати отримані дані та взаємодіяти з ними. Щоб підтримувати ефективну візуалізацію даних, система повинна мати доступ до відповідних результатів DS та мати інтуїтивно зрозумілі можливості взаємодії.

Візуалізація даних на діаграмі розсіювання або іншому графіку може виявити закономірності та зв’язки, які неможливо спостерігати інакше. Це також може запропонувати подальші шляхи для досліджень та нові бізнес-стратегії.

Отже, можна зробити висновок, DS – це процес видобутку великих наборів необроблених даних, як структурованих, так і неструктурованих, для виявлення закономірностей та отримання корисної інформації. Це міждисциплінарна галузь, і основи Data Science включають статистику, висновки, інформатику, прогнозну аналітику, розробку алгоритмів машинного навчання та нові технології для отримання інформації з великих даних.

Також DS напряму взаємопов’язано з другими галузями обробки та моделювання даних, таких як Big Data, Data Analytics, Data Science, Data Minnig Machine Learning.

Для кращого розуміння процесів роботи у сфері DS представлено сім кроків, які складають життєвий цикл DS: розуміння бізнесу, інтелектуальний аналіз даних, очищення даних, дослідження даних, конструювання ознак, прогнозне моделювання та візуалізація даних.

РОЗДІЛ 2

Аналіз ринку ломбардних послуг

2.1. Теоретичне обґрунтування ломбарду як фінансово-кредитної організації

У сучасній фінансово-кредитній системі існує значна кількість кредитних організацій, серед яких особливо виділяються небанківські кредитні організації (НКО), які є важливим елементом фінансово-кредитної системи, що повноцінно функціонує. НКО мають право здійснювати окремі банківські операції.[5]

Однією з цих структур є ломбарди. На жаль, в Україні роль і значення ломбардів у фінансово-кредитній сфері часто недооцінюється, а деякі автори взагалі не зараховують ломбарди до кредитних інститутів. Крім того, розгляд цього питання ускладнюється відсутністю достовірної статистичної інформації. Ломбарди є найстарішим видом фінансових організацій, послуги яких і досі користуються високим попитом.

Ломбард – це фінансова установа, яка надає зазвичай дуже дрібні позички на короткий термін. Ці позики видаються під заставу особистої власності позичальника (такі як годинник, ювелірні вироби, електроніка, автомобілі). Ломбардні позики надаються під відсоток, значно вищий, ніж той, що стягують основні фінансові установи. У різних країнах річний процент у ломбардах має діапазон від 6-12% (Франція, Бельгія) до 36-300% (США) і навіть до 120-360% (Україна).[3]

Соціальна значимість ломбардів обумовлена специфікою послуг. Якщо громадянин не зміг вчасно повернути гроші або прийняв рішення їх не повертати, то позику в будь-якому разі вважатиметься погашеною незалежно від суми, яку виручить ломбард при реалізації майна, що перебуває у нього в заставі. Невикуплені клієнтами вироби проходять ретельну передпродажну підготовку на найсучаснішому обладнанні.

Про український ринок ломбардних послуг відомо дуже мало. Аналітики та вчені не вивчають і не пишуть про ломбардну індустрію, тому що вона досить закрита та непрозора. Класичних підручників та інших об'єктивних джерел

інформації поки що дуже мало і основні дані про особливості цього бізнесу накопичені, в основному, співробітниками та власниками ломбардів.

Стати клієнтом ломбарду може будь-який громадянин України, при собі потрібно мати паспорт та річ, яку споживач має закласти для отримання позики. До того ж, ломбардом завжди можуть скористатися піддані інших країн, які тимчасово проживають на території України і тому не викликають довіри до банків: іммігранти, іноземці, туристи.

Клієнти ломбардних послуг, власне, є зріз нашого суспільства. Основними клієнтами є представники кількох соціальних груп: працівники бюджетної сфери, які мають стабільні, але дуже скромні зарплати, а також люди з низькими доходами. Найчастішими клієнтами ломбардів стають представники малого підприємництва. А для пенсіонерів це, по суті, єдина нагода отримати позику. Проведені опитування показують, що частина з них часто закладають сімейні коштовності та обручки, щоб звести кінці з кінцями до наступної пенсії, але щоразу викуповують застави назад.

Особливістю ломбардів, що визначає інтерес до них з боку держави та суспільства, є те, що вони виконують функцію своєрідних соціально-економічних «стабілізаторів» у будь-яких економічних формаціях. Видаючи короткострокові кредити громадянам, ломбарди задовольняють потреби населення в грошах, зменшуючи цим соціальну напруженість, і сприяють суто економічним шляхом підвищення платоспроможного попиту на товари. Це завдання особливо актуальне в умовах економічної кризи.

Існує ще окремий клас споживачів, які використовують ломбарди як швидку фінансову допомогу. Здебільшого, це жінки середнього віку, які здають коштовності для того, щоб через місяць викупити їх назад, коли закінчується їхня особиста «фінансова криза».

Ломбарди є суперниками комерційних банків над ринком споживчого кредитування. Сучасні ломбарди – це високотехнологічні установи, обладнані різноманітною технікою, що практично ні в чому не поступаються банкам у сфері обслуговування клієнтів, а за рядом параметрів переважають їх,

насамперед, через відсутність черг та мінімальний час на отримання кредиту: на оформлення заставного квитка витрачається від 15хв. до 2 годин. Крім того, банки майже не займаються мікрокредитуванням, а в ломбарді можна зайняти будь-яку суму. Завдяки роботі ломбардів, позикові кошти стали доступні тим споживачам, кому важко отримати кредити в банках через малі обсяги коштів, відсутності кредитної історії та інших причин.

Особливості ломбардного кредитування, його переваги та недоліки наведено в табл. 2.1.

Таблиця 2.1

Переваги та недоліки ломбардного кредитування

Переваги	**Недоліки**
Надання ломбардних кредитів українським громадянам, а також нерезидентам.	Високі процентні ставки користування кредитом: від 7 до 23 % на місяць (до 250 % на рік)
Надання кредитів клієнтам з низьким рівнем доходів, які не мають доступу до банківських кредитів	Низька оцінка майна, що закладається, скуповування за заниженими цінами ювелірних виробів та прикрас з дорогоцінних металів
Надання кредитів незалежно від мети їх використання	Незаконне збільшення ставки кредитування
Широкий спектр майна, що закладається	Продаж закладених речей під час пільгового терміну
Гнучкі терміни надання кредиту – від одного місяця до року	Продаж застав без проведення аукціонних торгів
Надання роздрібних фінансових послуг: видача мікрокредитів – невеликих сум від 1000 до 15000 грн.	Низька якість послуг в цілому
Мінімальні терміни оформлення кредитної угоди – від 15 хв. до 2 год.	Участь у ломбардному бізнесі несумлінних учасників ринку – псевдоломбардів
Платоспроможність клієнтів, їх доходи, місце роботи, соціальний статус не з'ясовуються	Прийом під заставу вкраденого майна
Прийняті під заставу речі страхуються на користь заставника на повну суму їх оцінки	Високі ризики для користувачів ломбардних послуг зберігаються
Ломбарди несуть відповідальність за втрату та пошкодження закладених речей	Незаконне переведення в готівку та «відмивання» грошей, отриманих кримінальним шляхом
Можливість викупити заставу в будь-який час	
Допомога в розвитку малого бізнесу за допомогою кредитування індивідуальних підприємців	
Пом'якшення соціальної напруженості в суспільстві при затримках зарплат, виникненні	

тимчасових фінансових труднощів у кризовий час	

Джерело: розроблено автором

В умовах фінансової кризи все виразніше стали виявлятися переваги ломбардів як кредитних інститутів ринку споживчого кредитування, табл. 2.2. Однією з основних переваг ломбардів є те, що вони можуть видавати кредити без тривалих і складних процедур. Крім того, інформація про платоспроможність клієнтів та подальший контроль над ними не потрібна, тому що ціна застави перевищує суму виданого кредиту. Соціальний статус клієнтів, їх прибутки, місце роботи не з'ясовуються. У ломбардів довгострокова заборгованість із боку клієнтів виникнути не може. Якщо боржник не виплачує отриманий кредит у певний термін і сплачує відсотки у ньому, то ломбард продає предмет застави аукціоні і покриває свої витрати.[2]

Таблиця 2.2

Порівняльна характеристика особливостей банківського та ломбардного кредитування

Критерії	**Ломбард**	**Комерційний банк**
Позичальники	Тільки фізичні особи, резиденти і нерезиденти, в тому числі клієнти з низьким рівнем дохідних ходів, що не мають кредитної історії, не отримують доступу до основних кредитних ринків	Фізичні та юридичні особи – резиденти
Застави	Рухоме майно, що належить позичальникам і призначене для особистого споживання	Рухоме та нерухоме майно, яке може відчужуватися відповідно до законів
Процентна ставка	Процентні ставки високі – до 200 - 250 %	Від 11 до 25% залежно від терміну кредитування
Платоспроможність клієнтів, їхня прибуткова база	Платоспроможність клієнтів, їх доходи, місце роботи, соціальний статус не з’ясовуються	Доходи позичальників та їх платоспроможність мають першорядне значення
Процедура оформлення кредитної угоди	Терміни оформлення угоди мінімальні і складають від 15 хв до 2 год.	Оформлення кредиту займає від 2 до 5 днів.

Сума кредитів, що видаються	Мінімальна сума не обмежена, максимальна сума обмежена	Мінімальна та максимальна суми обмежені
Терміни кредитування	Лише короткострокові кредити – від одного до трьох місяців (за законом – до одного року). За необхідності термін кредитування продовжується	Коротко-, середньо-, довгострокові кредити – від кількох місяців до 25…30 років
Цілі кредитування	Кредит не носить цільового характеру і забезпечується незалежно від мети	Вказівка мети отримання кредиту є обов'язковою
Можливість дострокового погашення	Дострокове повернення кредитної суми тільки вітається та не оподатковується додатковою комісією	Дострокове погашення кредитів не заохочується, обмовляється окремим пунктом договору
Ліцензування учасників фінансового ринку, констіль троль за діяльністю	Отримувати ліцензію не потрібно, діяльність ломбардів регулюється лише законами України	Фінансова діяльність банків ліцензується, здійснюється суворий контроль у режимі постійного моніторингу
Оподаткування доходів	Багато ломбардів подають свідомо помилкову інформацію про одержувані доходи, занижують суми податкових виплат або не платять їх взагалі	Доходи комерційних банків оподатковуються підвищеними податками
Зв'язок з тіньовою економікою та корупцією	Ломбардний бізнес привабливий для кримінальних структур, тісний зв'язок	Мінімальний і жорстко припиняється

Джерело: розроблено автором

У кризовий час частими клієнтами ломбардів стали громадяни, які тимчасово втратили роботу, особи, які мають складнощі з виконанням кредитних зобов'язань перед банками у зв'язку зі зниженням заробітної плати, а також бізнесмени, яким потрібні оборотні кошти для збереження бізнесу. Все частіше послугами ломбардів стали користуватися багаті громадяни та представники середнього класу.

Багатьом клієнтам подобається особиста участь в оцінці заставного майна, можливість індивідуального підходу щодо термінів і суми позики. Термін кредитування в ломбарді можна збільшити будь-якої миті, що неможливо при

оформленні кредиту в банку. Дострокове повернення кредитної суми лише вітається та не обкладається додатковим штрафом.

2.2. Особливості розвитку ринку ломбардних послуг у сучасних умовах України

Слід розпочати з нормативно-правової бази сфери ломбардних послуг України: Закон України «Про фінансові послуги та державне регулювання ринків фінансових послуг», Закон України «Про заставу», Закон України «Про забезпечення вимог кредиторів та реєстрацію обтяжень». [6][7]

Багато хто вважає, що в Україні реалізується «англійська» модель ломбардного ринку: головний дохід приносить лише кредитна діяльність, а не реалізація самої застави (близько 70 % позичальників віддавши борг, без проблем отримують заставу). Загалом, ломбарди розташовуються у спільній ніші з невеликими фінансовими компаніями та кредитними кооперативами, але через свою вузьку спеціалізацію суворо конкурують з ними лише у певній групі товарів. Ломбардний ринок має позитивні перспективи розвитку, здебільшого за допомогою збільшення списку прийнятого заставу майна.

На сьогоднішній день, найуніверсальнішими й важливими перспективами зростання є лише ломбарди. Вони включають єдиний клієнтський сегмент і активно конкурують за цілим рядом товарів. Цільовими клієнтами, як правило, є ті особи, які не мають можливості або бажання кредитуватися у великих банках (не входять до банківських стандартів, мають суттєві проблеми зі своєю кредитною історією, їм потрібне миттєве фінансування та інше). На сьогоднішній момент, щорічний приріст всього портфеля цих компаній не перевищує позначки 40 відсотків. За цієї умови зберігається ціла низка кількісних проблем: характерні спекулятивні вкладення, зумовлені малою фінансовою грамотністю громадян країни. Занадто високі ставки реалізації, які спричинені відсутністю недорогих пасивів на тривалий термін; ризики надлишкової позики, пов'язані з відсутністю взаємин з бюро історій кредитів та інше. [2]

На українському ринку характерною ознакою є те, що ломбарди видають кредити тільки за наявності застави. Заставою може бути практично будь-яке майно фізичної особи. Завдяки цьому у ломбарда мінімальні ризики неповернення кредиту та має можливість реалізовувати предмети застави. Вартість застави завжди перевищує розмір виданого кредиту, який наданий клієнту.[10]

Основна ознака прибутковості ломбарду – це висока процента ставка за кредитом. Станом на 2020 рік середньозважена річна ставка становить 217, 3%. Також спостерігається зростання обсяг виданих ломбардних кредитів за період з 2019 по 2021 рік. На рис. 2.1 зображено показники діяльності ломбардів.

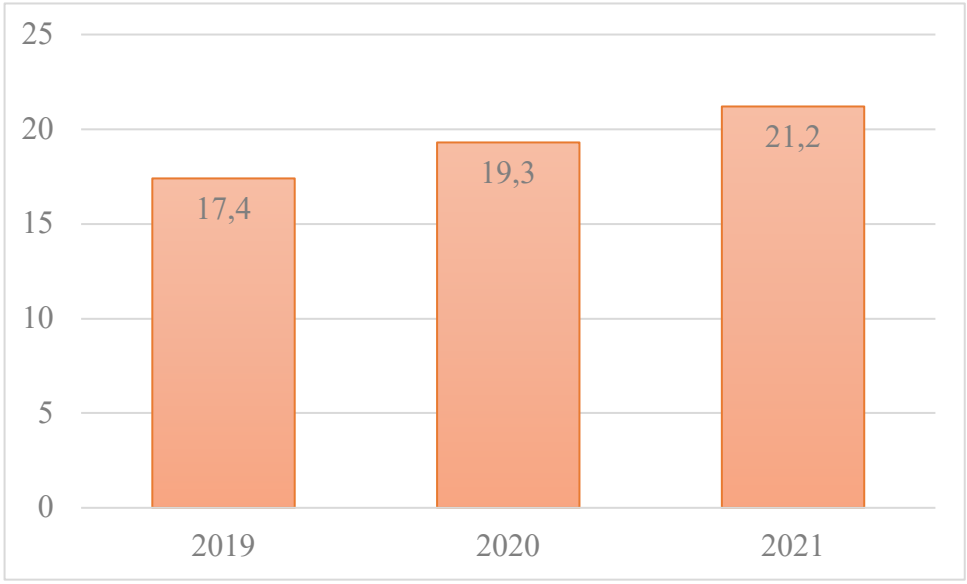

Рис. 2.1. Обсяг виданих кредитів

Джерело: розроблено автором на основі [1]

Зростання відбувається помірно, зумовлене зростанням попиту на такий вид кредитування тому, що неможливість отримання кредитів в інших банківських установах з низьким рівнем доходів.

Найбільшу частку майна, що надається під заставу займають ювелірні вироби, вони забезпечують 78% кредитів, далі йде побутова техніка – 21%, все інше займає лише 1% кредитів.

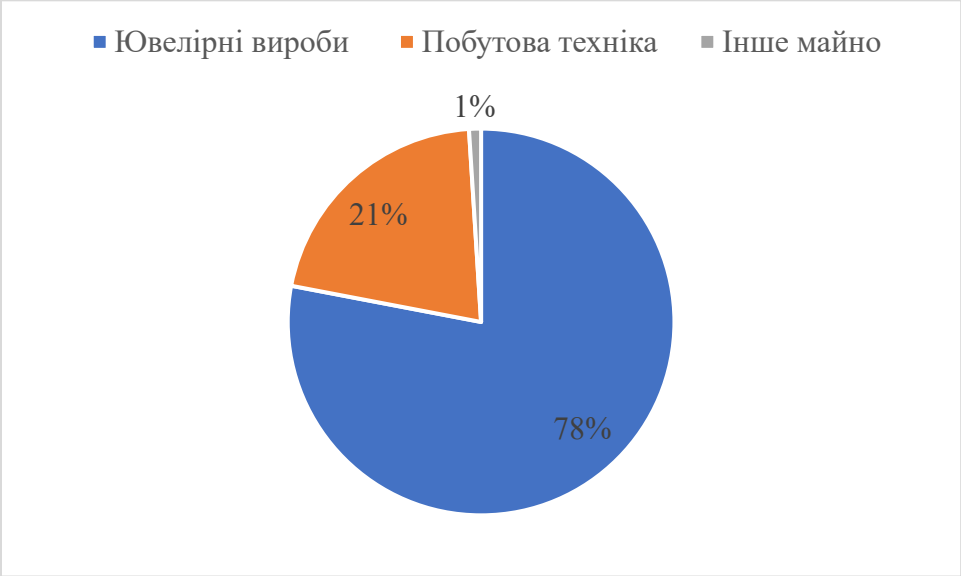

Рис 2.2. Структура видів застави

Джерело: розроблено автором на основі [1]

Вже не перший рік лідери українського ринку ломбардних послуг займають ломбард «Скарбниця» та ломбард «Благо». Вони на двох мають більше ніж 600 відділень та розташовані у 100 містах України.

Ломбард «Скарбниця» на даний момент найбільш технологічний на ринку України. У них є можливість керувати своїм кредитом через мобільний додаток. Також мають сервіс SkarbId в мобільному додатку ломбарду, через який додаються електроні документи та є можливість отримувати послуги у будь-якому відділенні без документів через смартфон. Достатньо велика кількість отримання послуг з отримання кредиту під заставу. [8]

Ломбард «Скарбниця» має найбільшу мережу своїх магазинів техніки серед інших ломбардів. Станом на 2022 рік загальна кількість магазинів налічує 184. Найбільше покриття відділень у місті Києві – 48, а загалом відділення представлені у 82 містах України. [9]

Отже, можна зробити висновки, що в умовах фінансової кризи ломбардний бізнес України отримав додатковий імпульс для подальшого розвитку. Попит на послуги ломбардів з кожним роком зростає.

Імідж ломбардів, як серйозних кредитних інститутів парабанківської системи країни, дедалі більше набирає сили. Сучасні ломбарди становлять конкуренцію комерційним банкам над ринком споживчого кредитування, а по

швидкості оформлення, видачі кредитів та з інших параметрів навіть перевершують їх.

Ломбардна індустрія значно розширила сферу впливу: сьогодні послугами ломбардів можуть користуватися всі верстви населення, хоча спочатку за допомогою до неї зверталися переважно клієнти з низьким рівнем доходів, що не мають доступу до банківських кредитних ресурсів.

Ломбардні мікрокредити дозволяють надавати фінансову допомогу населенню, знижувати соціальну напруженість, допомагають підтримувати малий бізнес, сприяють подоланню бідності та безробіття у суспільстві.

РОЗДІЛ 3

Методологія побудови системи Data Science для ломбарда «Скарбниця»

У цьому розділі проведемо аналіз бізнес-показників магазинів з продажу техніки ломбарда «Скарбниця» та побудову системи Data Science для прогнозування продажів магазину.

Для будь-якого підприємства є важливим аналіз бізнес-показників з метою подальшого розвитку та прибутку. Впровадження DS – проєктів має декілька етапів, які описані в першому розділі роботи.

Було обрано для аналізу три київських відділення магазинів з продажу техніки. Спочатку було розроблено процедуру вивантаження даних із системи обліку, так як загальна кількість магазинів перевищує 200 відділень та сформовано структуровані файли за певний період продажів. Вихідні дані сформовані за період з 01.05.2021 року по 30.04.2022 рік, показники, які містяться у базі даних, представлені у табл. 3.1.

Таблиця 3.1

Показники бази даних

Показник	Значення
sell_date	Дата та час продажу
weekday_num	День продажу: 1 - Понеділок, 2 - Вівторок і т.д.
isCredit	Товар продано у кредит
isNonCash	Товар продано за безготівковим розрахунком (платіжна картка/кредит)
isInternet	Замовлення товару здійснено в інтернет магазині
techgroup	Група товару
manufacturer_name	Виробник товару (Бренд)
itemcnt	Кількість проданого товару

itemprice	Ціна товару за одиницю
trading_period_days	Кількість днів з моменту доступності товару до продажу
shop_id	Ідентифікатор магазину
operator_id	Ідентифікатор продавця

Джерело: розроблено автором

Загальна кількість проданого товару 4172. Відділення №512 здійснило 1442 продажі товару, відділення №523 – 568 продажів, а найбільшу кількість продажів має відділення №864 – 2706.

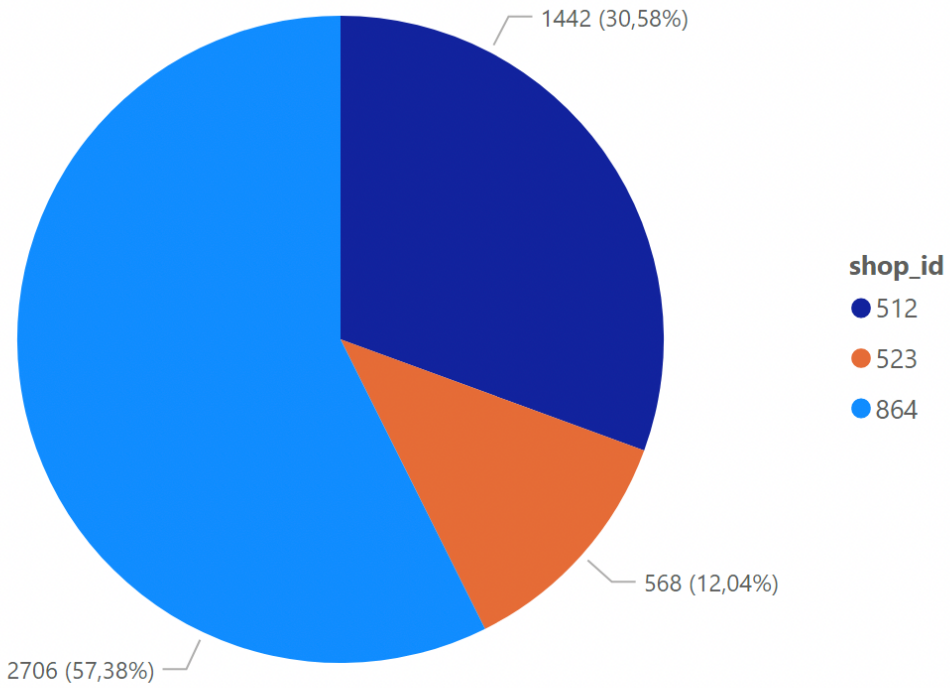

Рис. 3.1. Загальна кількість проданого товару

Джерело: розроблено автором

Також проаналізовано загальну суму продажів – 8428339 грн. По відділенням маємо такий розподіл: відділення №864 – 4745482 грн.; відділення №512 – 2610524 грн.; відділення №523 – 1072333 грн.

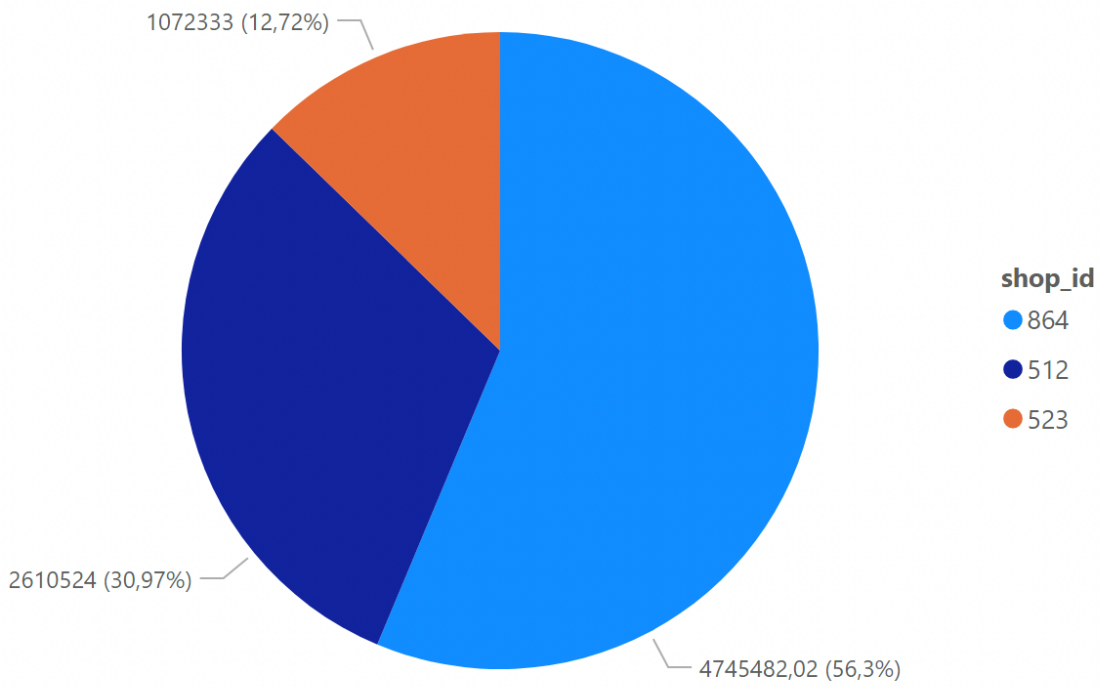

Рис. 3.2. Загальна сума продажів

Джерело: розроблено автором

Наступний етап включає в себе аналіз техніки по групам товарів та їх виробників. Загальна кількість груп товарів 21 та 588 вирбників (бренд товару). Найбільшу кількість продажів по групам товарів займають мобільні телефони – 2179 (46,2%), будівельна техніка – 730 (15,48%) та побутова техніка 475 (10,07%), а найменший відсоток має група – садові інструменти.

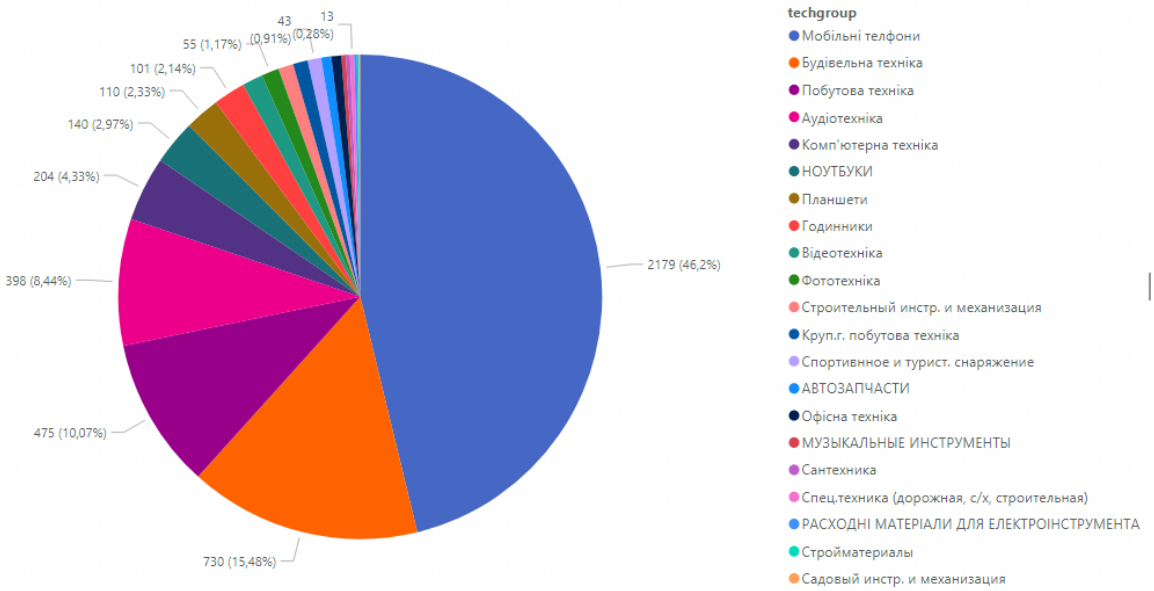

Рис. 3.3. Загальна кількість продажів по групам товарів

Джерело: розроблено автором

Далі на рис. 3.4 продемонстровано, яких виробників найбільше продано за трьома групами товарів, а саме мобільні телефони, будівельна техніка, побутова техніка, бо їх найбільша кількість продажів. Топ 3 бренди за кількістю продажів мобільних телефонів: «SAMSUNG» – 457 (20,9%), «XIAOMI» – 373 (17,12%), «APPLE» – 214 (9,82%).

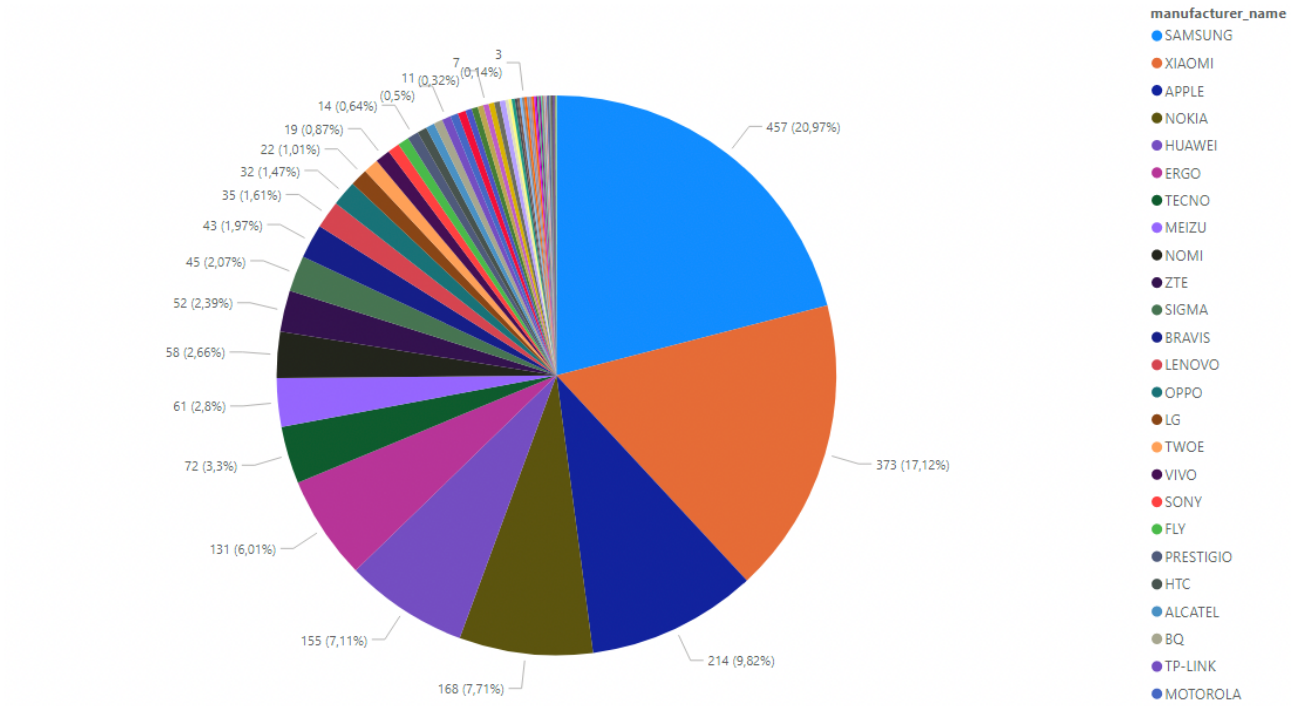

Рис 3.4. Загальна кількість продажів мобільних телефонів

Джерело: розроблено автором

Топ бренди за кількістю продажів будівельної техніки: «Дніпро-М» – 114 (15,62%), «BOSCH» – 100 (13,7%), «MAKITA» – 49 (6,71%), результати представлені на рис. 3.5.

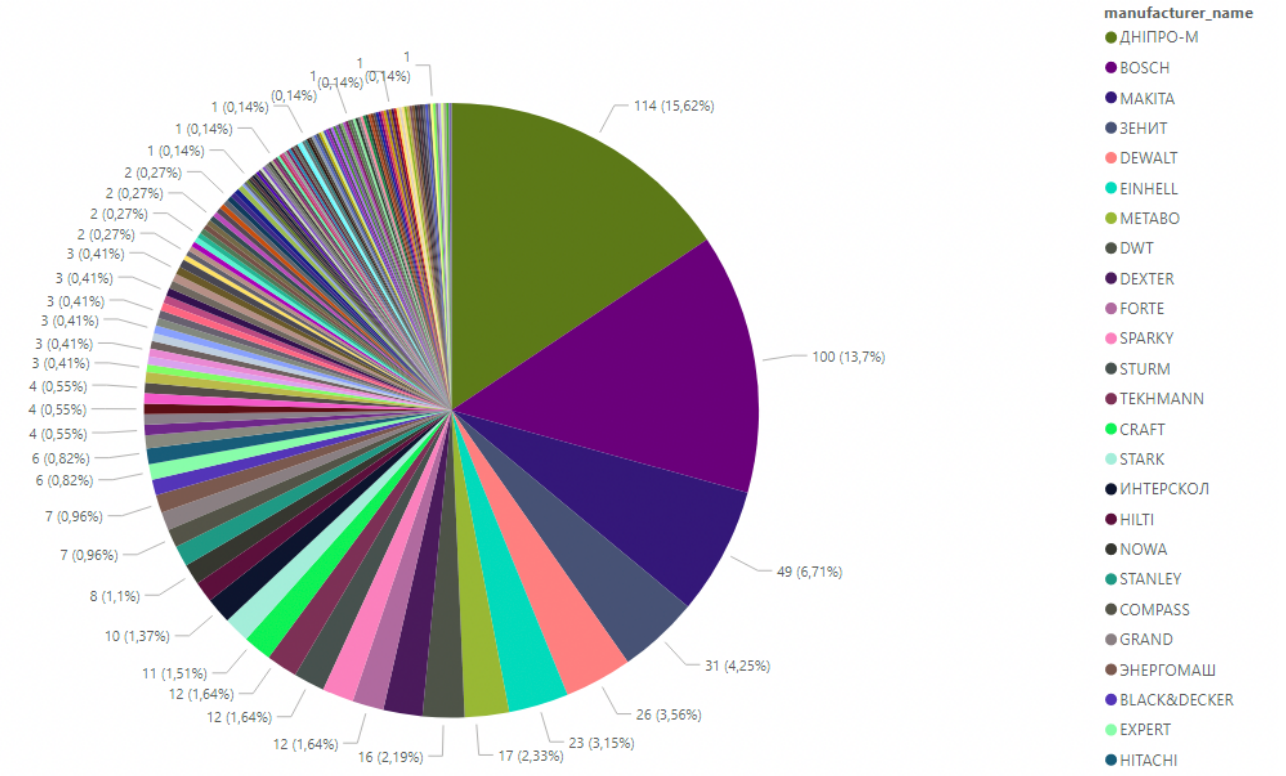

Рис. 3.5. Загальна кількість продажів будівельної техніки

Джерело: розроблено автором

Топ бренди за кількістю продажів побутової техніки: «GLO» – 130 (27,37%), «PHILIPS» – 34 (7,16%), «TEFAL» – 17 (3,58%), результати представлені на рис. 3.6.

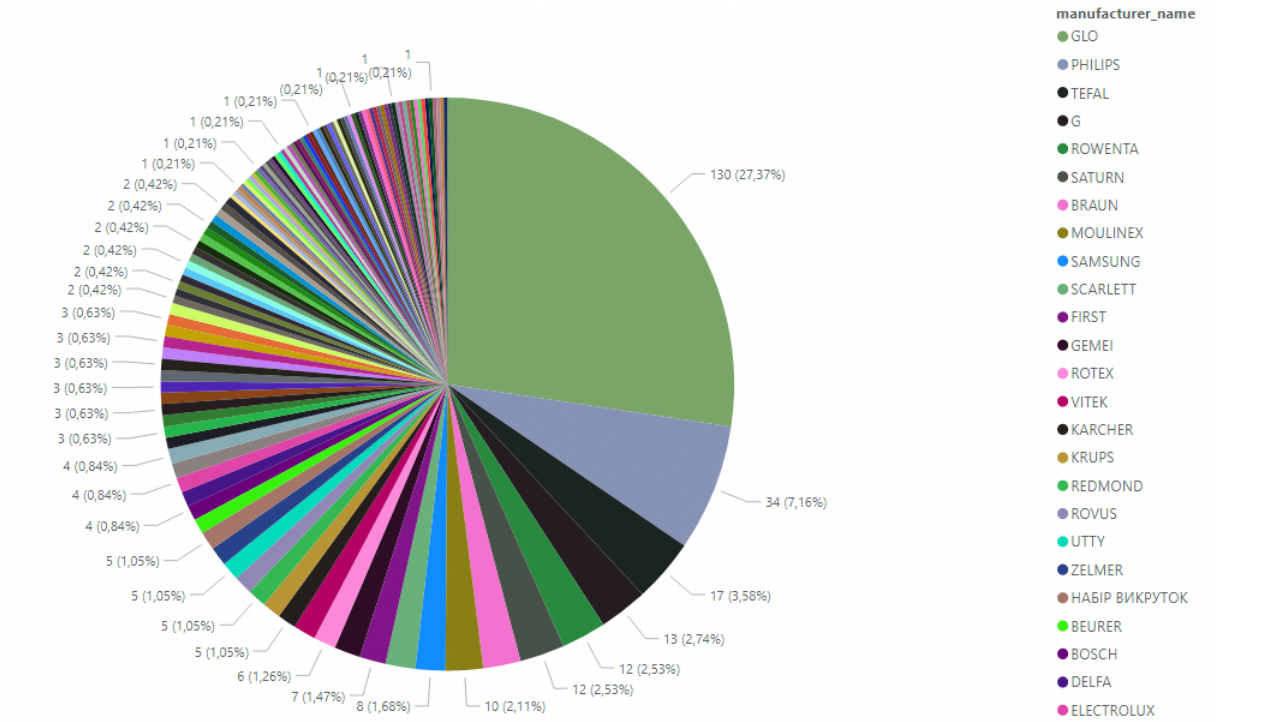

Рис. 3.6. Загальна кількість продажів побутової техніки

Джерело: розроблено автором

Також проаналізовано у якій спосіб купляли товари, результати представлені на рис. 3.7.

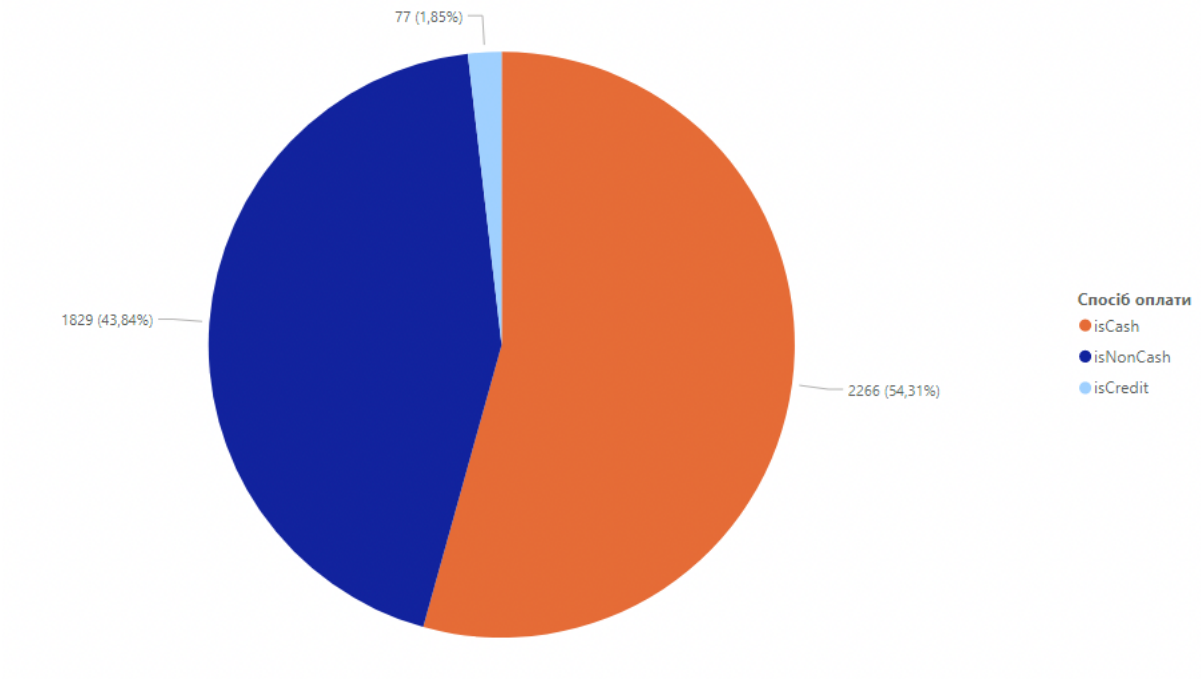

Рис. 3.7. Відображення розподілу між способами оплати

Джерело: розроблено автором

Переважна більшість обирали спосіб оплати готівкою – 2266 (54,31%), безготівковий розрахунок – 1829 (43,84%).

На рис. 3.8. також представлено кількість та сума продажів товарів за 11 місяців, а саме з травня 2021 року по квітень 2022 року.

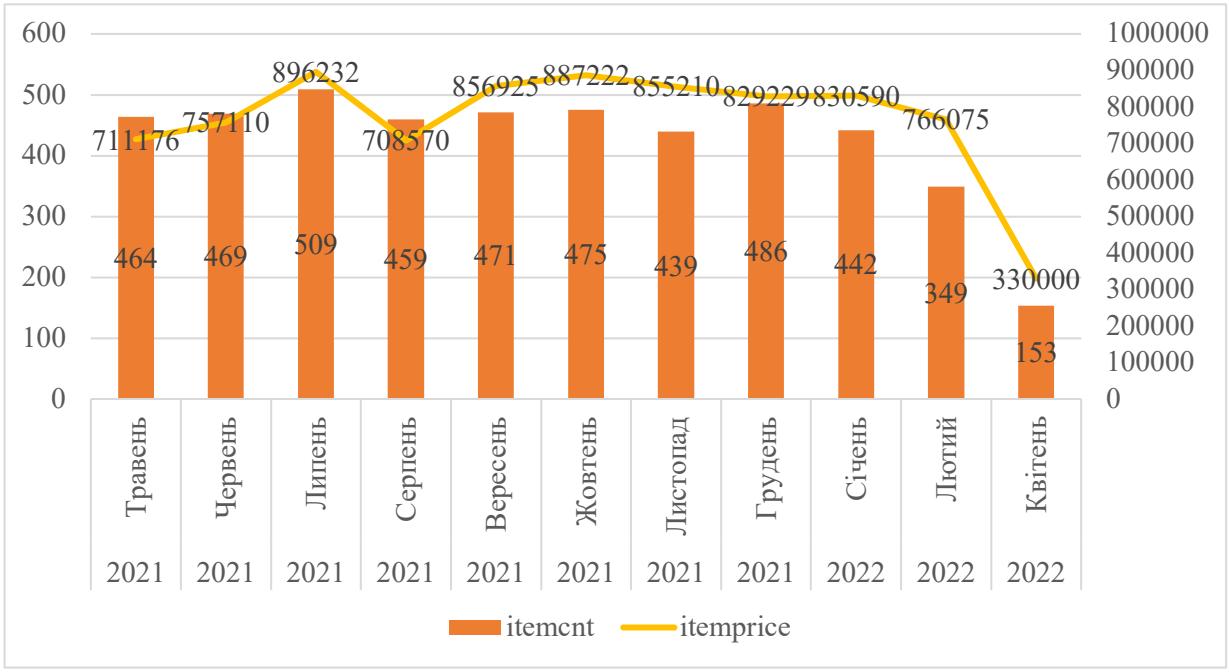

Рис 3.8. Кількість та сума продажів товарів по місяцям

Джерело: розроблено автором

Найбільша кількість та сума продажів у липні 2021 року, на другому місці за кількістю продажів грудень 2021 року, але за сумою продажів друге місце займає жовтень 2021 року. В даних відсутній березень 2022 року через те, що в Україні 24 лютого 2022 року розпочалася повномасштабна війна з Росією, тому в цей час у Києві відділення магазинів були зачинені. У квітні магазини відновили свою роботу, але не вийшли на той рівень продажів, які були в минулих місяцях.

Наступний етап це побудувати модель для прогнозу. Ціль проекту спрогнозувати суму продажів на рік в розрізі магазину, групи та виробників товарів, використовуючи планову кількість товарів.

У роботі використовувалася мова програмування Python. Для обробки та аналізу даних було використано бібліотеку Pandas, для побудови моделей було використано бібліотеку Sklearn. Бібліотека Sklearn найчастіше використовується у сфері Data Science та Machine Learning.[17]

Спочатку виконувалася попередня обробка даних. У табл. 3.2 представлено початкові дані.

Таблиця 3.2

Початкові дані магазинів техніки

sell_date	weekday_num	isCredit	isNonCash	isInternet	techgroup	manufacturer_name	itemcnt	itemprice	trading_period_days	shop_id	operator_id
15.11.2021 22:55	1	0	0	0	Мобільні телефони	NOKIA	1.0	210.0	4	512	8B91005056AD35EB11E18DDB13B28ACA
15.11.2021 21:59	1	0	1	0	Ноутбуки	LENOVO	1.0	3100.0	14	512	8B91005056AD35EB11E18DDB13B28ACA
15.11.2021 20:03	1	0	1	0	Комп'ютерна техніка	CROWN	1.0	110.0	325	864	9647005056AD35EB11EB5596CF69748C
15.11.2021 19:46	1	0	1	0	Будівельна техніка	HILTI	1.0	1550.0	106	864	9647005056AD35EB11EB5596CF69748C
11.12.2021 14:27	6	0	0	0	Будівельна техніка	BOSCH	1.0	3879.0	267	864	86B6001EC9DAD1F611DE7CD26EC23027
15.11.2021 17:22	1	0	0	0	Мобільні телефони	SAMSUNG	1.0	680.0	321	864	9647005056AD35EB11EB5596CF69748C
15.11.2021 17:53	1	0	1	0	Мобільні телефони	SAMSUNG	1.0	3700.0	66	864	9647005056AD35EB11EB5596CF69748C
...	…	…	…	…	…	…	…	…	…	…	…

Джерело: розроблено автором

Цільова змінна «itempricettl» – загальна сума продажів. Категоріальні змінні: «techgroup» – група товарів, «manufacturer_name» – виробник товару (Бренд), «shop_id» – ідентифікатор магазину. Далі було зроблено групування товарів за рік та отримано кількість продажів по кожній групі товару та дані було переведено до вигляду унітарного коду (Hot Encoding) за допомогою бібліотеки Pandas. У табл. 3.3 представлено дані після обробки.[22]

Таблиця 3.3

Дані після обробки

itempricettl	itemcnt	techgroup_Автозапчастини	techgroup_Аудіотехніка	techgroup_Побутова техніка	techgroup_Відеотехніка	techgroup_Комп'ютерна техніка	...	manufacturer_name_ЭЛЕКТРОМАШ	manufacturer_name_ЭЛПРОМ	manufacturer_name_ЭНЕРГОМАШ	shop_id_512	shop_id_523	shop_id_864
0	550.0	1.0	1	0	0	0	...	0	0	0	0	0	1
1	1400.0	1.0	1	0	0	0	...	0	0	0	1	0	0
2	315.0	5.0	1	0	0	0	...	0	0	0	1	0	0
3	1000.0	1.0	1	0	0	0	...	0	0	0	0	0	1
4	315.0	5.0	1	0	0	0	...	0	0	0	1	0	0
5	252.0	4.0	1	0	0	0	...	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Джерело: розроблено автором

Також було сформовано окремі DataFrame для залежної і незалежних змінних. Першою була побудована модель лінійної регресії за допомогою функції LinearRegression() з бібліотеки Sklearn. Результати моделі представлені у табл. 3.4.[42]

Таблиця 3.4

Результати моделі лінійної регресії

Показник	Результат
R_square	0.8361580575915755
Max error	224292.75
Mean squared error	312979664.45402366
alpha	711168591439885.5

Джерело: розроблено автором

Коефіцієнт детермінації – 83 %, вказує на гарний зв'язок між факторами регресії та залежної змінної. Але інші показники оцінки моделі мають дуже великі значення. Це вказує на те, що в моделі існує залежність між незалежними змінними.

Тому наступним кроком була побудова гребневої регресійної моделі з урахуванням мультиколінеарності незалежних змінних. Модель була побудована за допомогою функції Ridge() з бібліотеки Sklearn. Результати моделі представлені у табл. 3.5.[23]

Таблиця 3.5

Результати гребневої моделі регресії

Показник	Результат
R_square	0.8310656646383467
Max error	222967.80463739685
Mean squared error	322707426.5541365
alpha	-1026.4768921666537

Джерело: розроблено автором

Цей крок до суттєвих змін результатів не призвів, тому ці прогнози моделей поганої якості.

Через низьку якість лінійних регресійних моделей – висока максимальна та середньоквадратична помилка, було прийнято рішення побудувати регресійну модель CART. Для недопущення перенавчання дерева рішень, уся вибірка ділиться на тренувальну та тестову в відношенні 80% до 20% відповідно, максимальна глибина дерев (max_depth) встановлена на значені 17. Всі незалежні зміні – X та залежна зміна – Y відповідно. У табл. 3.6 представлено якість моделі на тестовій вибірці.[24][35]

Таблиця 3.6

Результати моделі на тестовій вибірці

Показник	Результат
R_square	0.5133281270309371
Max error	28143.0
Mean squared error	26918055.838662464

Джерело: розроблено автором

Судячи з результатів, зменшення максимальної помилки в 10 разів, а середньоквадратична помилка зменшилась в 50 разів. Тому якість моделі прийнятна для навчання моделі на всіх даних. Інформація моделі на повній вибірці представлена у табл. 3.7.

Таблиця 3.7

Результати моделі на повній вибірці

Показник	Результат
Відсоток помилок <100 грн.	72 %
Відсоток помилок <500 грн.	80%
R_square	0.9989590554070125
Max error	12365.550347222223
Mean squared error	1988468.182440912

Джерело: розроблено автором

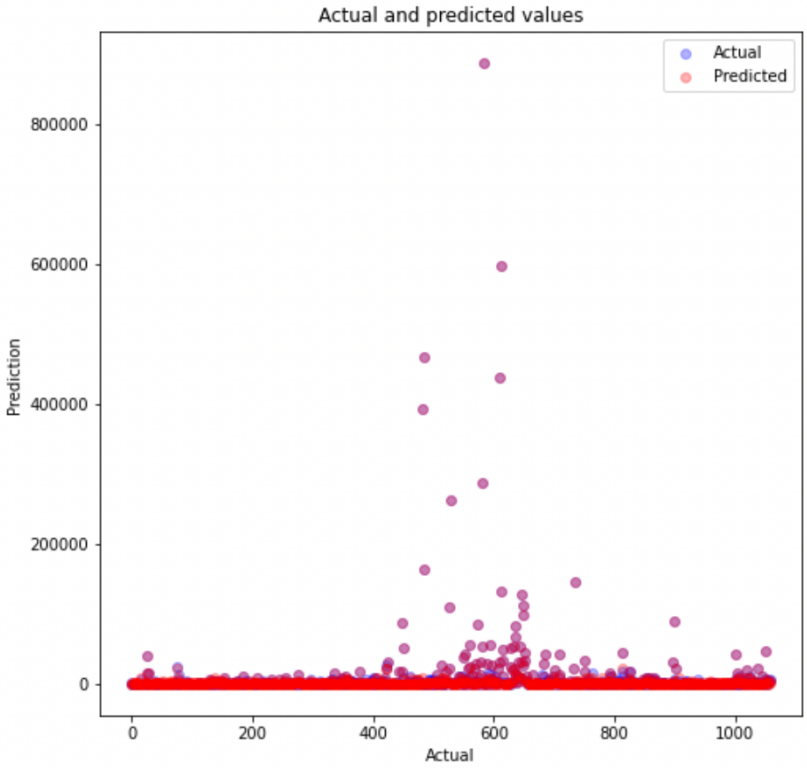

Рис. 3.9. Фактичні та прогнозні значення

Джерело: розроблено автором

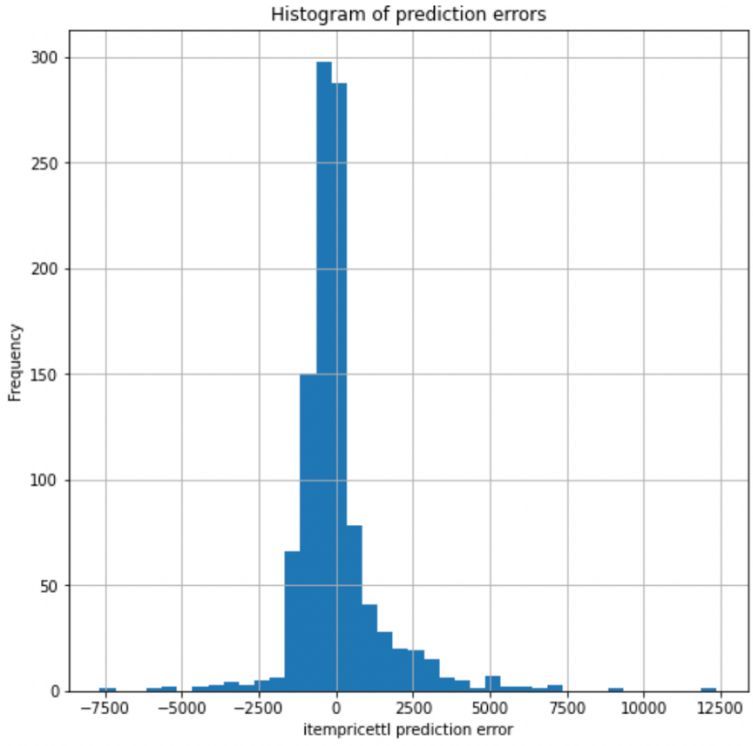

Рис 3.10. Помилки прогнозування

Джерело: розроблено автором

Модель показала гарний результати, особливо високий коефіцієнт детермінації, що вказує на тісний зв’язок між змінними. Було враховано кількість

помилок в заданих значеннях <100 грн. та <500 грн. відповідно. Тому модель дозволяє прогнозувати річні продажі товарів, 80% передбачень з точністю до 500 грн. в розрізі груп та виробників товарів.

Далі у табл. 3.8 наведено дані, які використовувалися на практиці для моделі прогнозу. У даних представлена інформація по групам товарів, виробнику, кількість продажів та ідентифікатор магазина. Результати прогнозу представлені у табл. 3.9.

Таблиця 3.8

Дані для прогнозу моделі

techgroup	manufacturer_name	itemcnt	shop_id
Мобільні телефони	NOKIA	10	864
Ноутбуки	LENOVO	40	864
Комп'ютерна техніка	CROWN	37	864
Будівельна техніка	BOSCH	24	864
Мобільні телефони	SAMSUNG	290	864
Мобільні телефони	ERGO	15	864
Мобільні телефони	APPLE	120	864
Планшети	SAMSUNG	61	864
Будівельна техніка	ДНІПРО-М	7	864
Мобільні телефони	OPPO	18	864

Джерело: розроблено автором

Скрипт програми для прогнозу наведено на рис. 3.11.

```
# Завантаження даних для прогнозу
data_pred = pd.read_csv('to_predict.csv', sep=';')
data_pf = pd.get_dummies(data_pred, columns=cat_features)
# Виділення списку
for x in features:
    if x not in data_pf.columns:
        data_pf[x] = 0
y_pred = model_tr.predict(data_pf)
data_pred[target] = y_pred
print(data_pred)
```

Рис. 3.11. Скрипт програми

Джерело: розроблено автором

Таблиця 3.9

Результати прогнозу моделі

	techgroup	manufacturer_name	itemcnt	shop_id	itempricettl
0	**Мобільні телефони**	NOKIA	10	864	20436.00
1	**Ноутбуки**	LENOVO	40	864	14643.00
2	**Комп'ютерна техніка**	CROWN	37	864	56200.00
3	**Будівельна техніка**	BOSCH	24	864	51231.00
4	**Мобільні телефони**	SAMSUNG	290	864	888072.00
5	**Мобільні телефони**	ERGO	15	864	54620.00
6	**Мобільні телефони**	APPLE	120	864	466791.00
7	**Планшети**	SAMSUNG	61	864	132930.00
8	**Будівельна техніка**	ДНІПРО-М	7	864	4733.86
9	**Мобільні телефони**	OPPO	18	864	9008.66

Джерело: розроблено автором

Отже, роблячи висновок, за результатами проведених моделювань можна стверджувати, що модель регресії дерево рішень дала кращі результати ніж лінійна чи гребенева регресія відповідно, які не показали задовільного результату.

Також у ході дослідження було проаналізовано та представлено у вигляді візуалізації бізнес-показників магазинів техніки ломбарду «Скарбниця». Модель, яка була побудована у роботі може у подальшому використовуватися для прогнозів річних продажів, що підвищує якість прийняття рішень у майбутньому для підприємства. У додатку А представлено повний скрипт Python з усіма етапами побудови моделей.

ВИСНОВКИ

Отже, розкрито теоретичні засади побудови системи Data Science на підприємстві, проведено порівняння DS з іншими галузями обробки даних, проаналізовані етапи життєвого циклу DS – проекту, проведено аналіз ринку ломбардних послуг. Також було проаналізовано бізнес-показники магазинів ломбарду «Скарбниця» та побудовано модель прогнозу продажів підприємства.

За результатами дослідження можна зробити такі висновки:

1. Data Science – це процес видобутку великих наборів необроблених даних, як структурованих, так і неструктурованих, для виявлення закономірностей та отримання корисної інформації. Це міждисциплінарна галузь, і основи DS включають статистику, висновки, інформатику, прогнозну аналітику, розробку алгоритмів машинного навчання та нові технології для отримання інформації з великих даних.

2. DS напряму взаємопов’язано з другими галузями обробки та моделювання даних, таких як Big Data, Data Analytics, Data Science, Data Minnig Machine Learning. Для кращого розуміння процесів роботи у сфері DS представлено сім кроків, які складають життєвий цикл DS: розуміння бізнесу, інтелектуальний аналіз даних, очищення даних, дослідження даних, конструювання ознак, прогнозне моделювання та візуалізація даних.

3. В умовах фінансової кризи ломбардний бізнес України отримав додатковий імпульс для подальшого розвитку. Попит на послуги ломбардів з кожним роком зростає. Імідж ломбардів, як серйозних кредитних інститутів парабанківської системи країни, дедалі більше набирає сили. Сучасні ломбарди становлять конкуренцію комерційним банкам над ринком споживчого кредитування, а по швидкості оформлення, видачі кредитів та з інших параметрів навіть перевершують їх. Ломбардна індустрія значно розширила сферу впливу: сьогодні послугами ломбардів можуть користуватися всі верстви населення, хоча спочатку за допомогою до неї зверталися переважно клієнти з низьким рівнем доходів, що не мають доступу до банківських кредитних ресурсів.

4. Ломбард «Скарбниця» на даний момент найбільш технологічний на ринку України. У них є можливість керувати своїм кредитом через мобільний додаток. Також мають сервіс SkarbId в мобільному додатку ломбарду, через який додаються електроні документи та є можливість отримувати послуги у будь-якому відділенні без документів через смартфон. Достатньо велика кількість отримання послуг з отримання кредиту під заставу. Ломбард «Скарбниця» має найбільшу мережу своїх магазинів техніки серед інших ломбардів. Станом на 2022 рік загальна кількість магазинів налічує 184. Найбільше покриття відділень у місті Києві – 48, а загалом відділення представлені у 82 містах України.

5. Було побудовано та проаналізовано декілька моделей, а саме лінійну та гребеневу регресію відповідно, які не дали результатів. Тому було побудовано регресійну модель CART. Модель показала гарний результати, особливо високий коефіцієнт детермінації, що вказує на тісний зв’язок між змінними. Було враховано кількість помилок в заданих значеннях <100 грн. та <500 грн. відповідно. Тому модель дозволяє прогнозувати річні продажі товарів, 80% передбачень з точністю до 500 грн. в розрізі груп та виробників товарів.

6. У ході дослідження було проаналізовано та представлено у вигляді візуалізації бізнес-показників магазинів техніки ломбарду «Скарбниця». Модель, яка була побудована у роботі може у подальшому використовуватися для прогнозів річних продажів, що підвищує якість прийняття рішень у майбутньому для підприємства.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Біла книга «Майбутнє регулювання діяльності ломбардів» / НБУ. Київ, 2020. 24 с. URL: https:// bank.gov.ua/admin_uploads/article/White_paper_lombard_2020.pdf?v=4
2. Бойко С. В., Кеба А. А. Сучасний стан та перспективи розвитку ломбардів в Україні. Інтернаука. Серія «Економічні науки». 2017. № 16. URL: https://www. inter-nauka.com/uploads/public/15115321262724.pdf
3. Булій Н. О. Особливості ломбардної діяльності на ринку фінансових послуг України / Н. О. Булій // Науковий журнал: тенденції фінансового ринку / ЛНУ імені Івана Франка. – Львів, 2012. – №23. – С. 34-39
4. Гнот Т.В. Алгоритми Data Science у моделюванні бізнес- процесів. Економіка і суспільство, 2017 (Випуск 12)
5. Дребот Н. П., Танчак Я. А., Миколишин М. М. Тенденції розвитку небанківських фінансових установ на ринку фінансових послуг України. Науковий вісник НЛТУ України. 2020. Т. 30. № 1. С. 109–114.
6. Закон України «Про заставу» від 02.10.1992 р. № 2654-XII [Електронний ресурс]. – Режим доступу: http://zakon2.rada.gov.ua/laws/show/2654-12
7. Закон України «Про фінансові послуги та державне регулювання ринків фінансових послуг» від 12.07.2001р. № 2664-III [Електронний ресурс]. – Режим доступу: http://zakon2.rada.gov.ua/laws/show/2664-14
8. Ломбард «Скарбниця», URL: https://www.skarb.com.ua/.
9. Магазин техніки «Техноскарб», URL: https://ua.tehnoskarb.ua/.
10. Положення про порядок надання фінансових послуг ломбардами // Розпорядження Держфінпослуг України від від 26.04.2005 № 3981 [Електронний ресурс]. – Режим доступу: http://zakon2.rada.gov.ua/laws/show/z0565-05
11. Самойленко Л. Б. Можливості та проблеми застосування технологій big data вітчизняними компаніями. Ефективна економіка. 2018. No 1.
12. Фоусет Т. Data Science для бізнесу. Як збирати, аналізувати і використовувати дані, 2019.

13. Цивільний кодекс України від 16.01.2003 р. № 435-IV [Електронний ресурс]. – Режим доступу: http://zakon2.rada.gov.ua/laws/show/435-15
14. Bakir, H., Chniti, G., Zaher, H. 2018. E-Commerce Price Forecasting Using LSTM Neural Networks. International Journal of Machine Learning and Computing, 8, 169-174. https://doi.org/10.18178/ijmlc.2018.8.2.682.
15. Chakure A. Random Forest Regression, 29.06.2019. URL: https://towardsdatascience.com/random-forest-and-its-implementation-1824ced454f
16. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C et al. CRISP-DM 1.0: Step-by-step data mining guide; 2000.
17. Deily N. Python Insider / N. Deily. — The Python Core Developers, 2019. — 240 p.
18. Donges N. A complete guide to the Random Forest algorithm, 16.06.2019. URL: https://builtin.com/data-science/random-forest-algorithm
19. Freedman D.A. Statistical Models: Theory and Practice. Cambridge University Press. 2009. P. 221−228
20. Galton F. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland. 1886. P. 246–263
21. García-Santillán, A., Ramos-Hernández, J. J., Hernández-González, S., Rivera-Santiago P. Pawnshops: Are They Really a Solution for the Vulnerable Population? Ecoforum Journal. 2017. Vol. 6. Is. 1. URL: http://www. ecoforumjournal.ro/index.php/eco/article/viewFile/554/341
22. Garreta R. Learning scikit-learn: Machine Learning in Python Paperback / R. Garreta, G. Moncecchi // Packt Publishing, 2013. — P. 183-190.
23. Ghadban K., Ghazi S. Choosing Ridge Parameter for Regression Problems. Communications in Statistics – Theory and Methods. 2005. P. 1177–1182
24. Godbole S. Discriminative methods formulti—labeled classification / S. Godbole, S. Sarawagi // Adv.Knowl. Discov. Data Mining, 2004. — P. 22- 30.
25. Gruber M. Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators. Boca Raton: CRC Press. 1998. P. 7–15

26. Hoffmann M. Von Industrial Big Data zu Smart Data: Wie aus Produktionsdaten Erkenntnisse werden. Informatik Aktuell 2019, 2019.
27. Kamthania, D., Pawa, A., Madhavan, S.S. 2018. Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business. Journal of computing and information technology, 26, 57-68. https://doi.org/10.20532/cit.2018.1003863.
28. Meier H, Boßlau M. Design and Engineering of Dynamic Business Models for Industrial Product-Service Systems. In: Shimomura Y, Kimita K, editors. The Philosopher's Stone for Sustainability. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, p. 179–184.
29. Olha Fedirko , Tetiana Zatonatska, Tomasz Wołowiec, Stanisław Skowron Data Science and Marketing in E- Commerce Amid COVID-19 Pandemic//European Research Studies Journal Volume XXIV, Special Issue 2, 2021 – pp. 3-16
30. Parveen, N., Santhi, M.V.B.T., Burra, L.R., Pellakuri, V., Pellakuri, H. 2021. Women's ecommerce clothing sentiment analysis by probabilistic model LDA using RSPARK. Materials Today: Proceedings, (in press). https://doi.org/10.15388/Ekon.2020.2.6.
31. Pliskunova, O., Klochko, R. 2020. Classification of e-commerce customers based on Data Science techniques. CEUR Workshop Proceedings. Available at: http://ceurws.org/Vol-2649/paper2.pdf.
32. Powers D. The Problem of Area Under the Curve // International Conference on Information Science and Technology, 2012. — P. 12-17.
33. PythonAnywhere. 2021. Plans and Pricing. Available at: https://www.pythonanywhere.com/pricing/.
34. Ross Q.J. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers. 1993. P. 122−131
35. Ross Q.J. Induction of decision trees. Machine Learning. 1986. P.81-106
36. Ruczinski I., Kooperberg C. Logic Regression, 2001-2005 URL: http://kooperberg.fhcrc.org/logic/documents/documents.html

37. The Guardian [Електронний ресурс] URL: https://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume
38. Tibshirani R. Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society. 1996. P. 267–88
39. Tolles J. Logistic Regression Relating Patient Characteristics to Outcomes / J. Tolles,W. Meurer // JAMA, 2016. — P. 533.
40. Tsoumakas G. Random k-Labelsets: an ensemble method for multilabel classification [Електронний ресурс] / G. Tsoumakas, I. Vlahavas — 2007, vol. 4701. — P. 406-417. — Режим доступу: https://link.springer.com/chapter/10.1007/978-3-540-74958-5_38
41. Wang, Q., Cai, R., Zhao, M. 2020. E-commerce brand marketing based on FPGA and machine learning. Microprocessors and Microsystems, 103446. https://doi.org/10.1016/j.micpro.2020.103446.
42. Wei Y., Pere A., Koenker, R., He, X. Quantile Regression Methods for Reference Growth Charts. Statistics in Medicine. 2006. P. 1369–1382
43. Wirth R, Hipp J. CRISP-DM: Towards a Standard Process Model for Data Mining. In: Mackin N, editor. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining: 11th - 13th April 2000, Crowne Plaza Midland Hotel, Manchester, UK. Blackpool, Lancashire: Practical Application Company; 2000, p. 29– 39.

ДОДАТКИ

Додаток А

```
import pandas as pd
# модель регресії
from sklearn.linear_model import LinearRegression, Ridge #0.24.2
from sklearn import tree #0.24.2
from sklearn.metrics import max_error, mean_squared_error #0.24.2
from sklearn.model_selection import train_test_split #0.24.2

from google.colab import files
upload = files.upload()

data_L0 = pd.read_csv('original_data.csv')
data_L0.head(5)

target = 'itempricettl'

# категоріальні зміні
cat_features = ['techgroup', 'manufacturer_name', 'shop_id']

# змінні для групування
columns2groupby = cat_features
data_L0[target] = data_L0['itemcnt']*data_L0['itemprice']
data_L0[data_L0['itemcnt'] > 1].head(10)

# Групування даних за рік та отримання кількості продажів по кожній групі
data_L1 = data_L0.groupby(columns2groupby).agg({target:'sum',
                                                'itemcnt':'sum'}).reset_index()

# FIXME: отладочные строки
```

```
#data_L1.head(20)
data_L1.describe()
# Hot Encoding для категорільних параметрів
# зберігаємо дані з перекодованими категоріальними параметрами
data_L2 = pd.get_dummies(data_L1, columns=cat_features)
data_L2.to_csv('model.csv')
data_L2.head(20)
# FIXME: отладочные строки
data_L2.shape
data_L2.head(10)
#data_L2.describe()
# отримуємо список усіх незалежних змінних (всіх стовпців не рівних цільової змінної)
features = [x for i,x in enumerate(data_L2.columns) if x != target]
print(features)
# Окремі DataFrame для залежної та незалежних змінних
data_L3f = data_L2[features]
data_L3t = data_L2[[target]]
# FIXME: отладочные строки
data_L3f.shape
data_L3f.head(10)
data_L3f.describe()
# навчання моделі лінійної регресії
lm = LinearRegression()
model = lm.fit(data_L3f, data_L3t)
# Інформація про модель
print(f'R_square = {model.score(data_L3f, data_L3t)}')
print(f'Max error = {max_error(data_L3t[target], model.predict(data_L3f))}')
print(f'Mean squared error = {mean_squared_error(data_L3t[target], model.predict(data_L3f))}')
```

```
print(f'alpha = {model.intercept_[0]}')
coef = dict()
for idx, x in enumerate(data_L3f.columns):
    coef[x] = model.coef_[0][idx]

results = pd.DataFrame(coef.items(), columns=['Feature', 'Value']).sort_values(['Value'],ascending=False)
print(results)
# навчання моделі гребеневої регресії (враховуючи мультиколінеарності незалежних features)
lr = Ridge()
model_R = lr.fit(data_L3f, data_L3t)
# Інформація про модель
print(f'R_square = {model_R.score(data_L3f, data_L3t)}')
print(f'Max error = {max_error(data_L3t[target], model_R.predict(data_L3f))}')
print(f'Mean squared error = {mean_squared_error(data_L3t[target], model_R.predict(data_L3f))}')
print(f'alpha = {model_R.intercept_[0]}')
coef = dict()
for idx, x in enumerate(data_L3f.columns):
    coef[x] = model_R.coef_[0][idx]

results = pd.DataFrame(coef.items(), columns=['Feature', 'Value']).sort_values(['Value'],ascending=False)
print(results)
# Для недопущення перенавчання дерева рішень, вся вибірка поділяється на навчальну (80%) та тестову (20%)
X_train, X_test, y_train, y_test = train_test_split(data_L3f, data_L3t, test_size=0.2, random_state=9273)
```

```
# навчання моделі на частині даних
tr = tree.DecisionTreeRegressor(max_depth=17)
model_tr = tr.fit(X_train, y_train)

y_pred = model_tr.predict(X_test)
# якість моделі на тестовій вибірці
print(f'TEST R_square = {model_tr.score(X_test, y_test)}')
print(f'TEST Max error = {max_error(y_test, y_pred)}')
print(f'TEST Mean squared error = {mean_squared_error(y_test, y_pred)}')
# Якість прийнятна, навчання моделі на всіх даних для застосування
model_tr = tr.fit(data_L3f, data_L3t)

# якість моделі на повній вибірці
y_pred = model_tr.predict(data_L3f)
# розрахунок помилок
# % помилок, що не перевищують задане значення
diff = (data_L3t[target] - y_pred)
ttl_cnt = len(data_L3t.index)
p01 = len(diff[diff < 100])/ttl_cnt #100UAH
p10 = len(diff[diff < 500])/ttl_cnt #500UAH
print(diff)
print(ttl_cnt)
print(p01)
print(p10)

# Інформація по моделі
print('<100UAH error share = ', p01, ', <500UAH error share = ', p10)
print(f'R_square = {model_tr.score(data_L3f, data_L3t)}')
```

```
print(f'Max error = {max_error(data_L3t[target], y_pred)}') # максимальная ошибка
print(f'Mean squared error = {mean_squared_error(data_L3t[target], y_pred)}') # среднеквадратическая

from google.colab import files
upload = files.upload()

data_pred = pd.read_csv('to_predict.csv', sep=';')
print(data_pred)

# Завантаження даних для прогнозу
data_pred = pd.read_csv('to_predict.csv', sep=';')
data_pf = pd.get_dummies(data_pred, columns=cat_features)
# Виділення списку
for x in features:
    if x not in data_pf.columns:
        data_pf[x] = 0
y_pred = model_tr.predict(data_pf)
data_pred[target] = y_pred
print(data_pred)

y_pred = model_tr.predict(data_L3f)

from matplotlib import pyplot as plt

# Побудова графіків
#actual/predicted
_, ax = plt.subplots(figsize=(8,8))
```

```
ax.scatter(x = range(0, data_L3t[target].size), y=data_L3t[target], c = 'blue', label
= 'Actual', alpha = 0.3)
ax.scatter(x = range(0, y_pred.size), y=y_pred, c = 'red', label = 'Predicted', alpha
= 0.3)
plt.title('Actual and predicted values')
plt.xlabel('Actual')
plt.ylabel('Prediction')
plt.legend()
plt.show()

#errors
diff = data_L3t[target] - y_pred
diff.hist(bins = 40, figsize=(8,8))
plt.title('Histogram of prediction errors')
plt.xlabel(target +' prediction error')
plt.ylabel('Frequency')

#tree
fig = plt.figure(figsize=(30,15), facecolor='white')
_ = tree.plot_tree(model_tr,
                   feature_names=features,
                   class_names=target,
                   filled=True)
```