Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра _____Програмної інженерії_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

Дослідження використання прогностичного аналізу та Big Data в електронній комерції
(тема)

Виконав: студент 2 курсу, групи ПЗСм-19-1___
спеціальності 121- Інженерія програмного забезпечення
(код і повна назва спеціальності)
Освітньо-професійної програми
Програмне забезпечення систем_____
(повна назва освітньої програми)

_____Слєпєнкова Є. С._____
(прізвище, ініціали)
Керівник _____проф. Дудар З. В._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2020 р.

Харківський національний університет радіоелектроніки

Факультет_Комп'ютерних наук_____

Кафедра_Програмної інженерії_____

Рівень вищої освіти_____другий (магістерський)_____

Спеціальність _121-Інженерія програмного забезпечення_____
(код і повна назва)

освітньо-професійна програма _Програмне забезпечення систем_____
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові _Слєпєнковій Єлизаветі Сегіївні_____
(прізвище, ім'я, по батькові)

1. Тема роботи _ Дослідження використання прогностичного аналізу та Big Data в електронній комерції _____
затверджена наказом по університету від "30" _жовтня _20 20 р № _1490 СТ_
2. Термін подання студентом роботи до екзаменаційної комісії
18 грудня 2020 р.
3. Вихідні дані до роботи *предиктивний аналіз, предиктивні моделі, алгоритми передбачування, великі дані, пояснювальна записка. Використовувати python, jupyter*_____
4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз проблемної галузі і постановка задачі, огляд стану проблеми електронній комерції на даний час, аналіз метрик аналіз методів обробки даних, аналіз існуючих рішень, алгоритм впровадження предиктивного аналізу, аналіз отриманих результатів*
_____

5. Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф. Дудар З.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка *
1.	Аналіз предметної галузі	16 вересня	виконано
2.	Формулювання проблеми	24 вересня	виконано
3.	Постановка задачі	29 вересня	виконано
4.	Дослідження використання прогностичного аналізу та Big Data в електронній комерції	10 жовтня	виконано
5.	Опис алгоритму	25 жовтня	виконано
6.	Підготовка пояснювальної записки	20 листопада	виконано
7.	Спецчастина	30 листопада	виконано
8.	Підготовка презентації та доповіді	3 грудня	виконано
9.	Попередній захист	7 грудня	виконано
10.	Нормоконтроль, рецензування	9 грудня	виконано
11.	Занесення диплома в електронний архів	11 грудня	виконано
12.	Допуск до захисту у зав. кафедри	14 грудня	виконано

Дата видачі завдання 1 вересня 2020 р.

Студент _____
(підпис)

Керівник роботи _____ проф. Дудар З. В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Атестаційна робота магістра містить: 89 с., 42 рис., 7 табл., 2 додатки, 46 джерел.

ПРОГНОСТИЧНА АНАЛІТИКА, ПРЕДИКТИВНА АНАЛІТИКА, ІНТЕРНЕТ-МАГАЗИН, ЕЛЕКТРОННА КОМЕРЦІЯ, BIG DATA, DATASET.

Об’єктом дослідження є моделі та алгоритми прогностичної аналітики при роботі з великими наборами даних в електронній комерції.

Метою роботи є дослідження використання моделей, які можуть бути використані для аналізу даних метрик та подальшого використання в онлайн-магазинах.

Методи розробки базуються на моделях та алгоритмах для обробки зібраних даних про купівельний досвід користувача інтернет-магазину.

У результаті роботи були здійснені дослідження та проаналізовані алгоритми, які у майбутньому будуть використані для реалізації складної програмної системи.

FORECAST ANALYSIS, PREDICTIVE ANALYTICS, ONLINE STORE, E-COMMERCE, BIG DATA, DATASET.

The object of study of the model and algorithms of predictive analytics when working with large data sets in e-commerce.

The aim of the work is to study the use of models that can be used to analyze metric data and further use in online stores.

Development methods are based on models and algorithms for processing the collected data on the shopping experience of the online store user.

As a result, research was carried out and algorithms were analyzed, which in the future will be used to implement a complex software system.

ЗМІСТ

ВСТУП

Прогностична аналітика – це використання даних, статистичних алгоритмів та методів машинного навчання для виявлення ймовірності майбутніх результатів на основі історичних даних. Мета полягає в тому, щоб вийти за межі знання того, що сталося, і дати найкращу оцінку того, що станеться в майбутньому.

Прогностична аналітика залучила підтримку широкого кола організацій. Прогнозується, що глобальний ринок досягне близько 10,95 млрд. доларів до 2022 року, зростаючи зі складеним річним рівнем зростання (CAGR) близько 21 відсотка між 2016 і 2022 роками, згідно зі звітом 2017 року

Організації звертаються до прогностичної аналітики, щоб допомогти вирішити складні проблеми та відкрити нові можливості.

Прогнозування поведінки покупців – це цікаве і складне завдання. У контексті електронної комерції вирішення цієї проблеми є приоритетним порівняно з багатьма новими проблемами, виникаючими серед проблем традиційного бізнесу.

Метою роботи є дослідження виявлення способів використання моделей, які можуть бути використані для аналізу метрик онлайн-магазинів та подальшого використання в онлайн-магазинах.

У цьому дослідженні будуть вивчатися фактори, які впливають на процес прийняття рішень покупцями в Інтернеті: потреби покупців, популярність товарів та переваги споживачів. Крім того, використовуючи дані про закупівлі та рейтинги товарів на веб-сайті електронної комерції, будуть досліджуватися методи кількісної оцінки сили факторів: використання асоціацій між товарами для прогнозування потреб споживачів; поєднання спільної фільтрації та ієрархічної байєсівської моделі дискретного вибору для вивчення переваг споживачів; побудова моделі регресії на основі опори, для розрахунку популярності продукції.

Таким чином, метою роботи буде дослідження використання прогностичної аналітики в системах електронної комерції для великих наборів даних. Прогностична аналітика дозволяє знаходити в великому масиві даних залежні один від одного області. Це дозволяє систематизувати масив даних і аналізувати їх за допомогою спеціальних методів. Такі методи предиктивної аналітики можна застосовувати для визначення моделі поведінки відвідувача веб-ресурсу, наприклад, з метою персоналізації сайту в залежності від його попередньої історії. Крім того, предиктивна аналітика вміє знаходить стандартні типи закономірностей: асоціація, послідовність, класифікація, кластеризація (сегментування) і прогнозування.

Для цього будуть проведені дослідження різних моделей та алгоритмів застосування предиктивного аналізу в електронній комерції. Результати цього дослідження будуть мати цінність для невеликих та середнього розміру компаній, які не використовують ще цей унікальний інструмент розробки маркетингових стратегій, який дозволить робити цінні висновки, знижувати ризики, приймати виважені рішення і надавати клієнтам персональне обслуговування. Подолання обмежень у розвитку дозволить компаніям рости та розвиватися на ринку електронної комерції, де дуже велика конкуренція.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Електронна комерція

Електронна комерція – це фінансові або торгові транзакції, що здійснюються за допомогою мереж. Ланцюжки глобальних бізнес-процесів, пов'язаних з проведенням транзакцій, постійно удосконалюються. Статистика, свідчить про значні темпи зростання цифрової економіки за останні роки. По суті, електронна комерція – це ядро, яке стрімко народжується в світі цифрової економіки і є результатом вибухових темпів розвитку телекомунікаційних потужностей за останнє десятиліття.

Розвиток глобальних маркетплейсів і їх зростаюча доступність дозволяє будь-якій людині практично в будь-якій точці світу не тільки купувати, але і продавати – без бар'єрів, як на внутрішньому, так і на зовнішніх ринках, знижуючи витрати на виробництво і торгівлю, а крім того – економити час. Німецький портал Statista оцінює світовий ринок електронної комерції 2017 року $ 1,5 трлн, в 2019 – майже в $ 2 трлн, а до 2022 прогнозує його послідовне зростання до $ 2,5 трлн. Сьогодні до електронної комерції відносять електронну покупку або продаж товарів через онлайн-сервіси або через Інтернет, мобільну комерцію, електронні перекази коштів, управління ланцюгами поставок, інтернет-маркетинг, онлайн-обробку транзакцій, електронний обмін даними (EDI), системи управління запасами і автоматизовані системи збору даних. Електронна комерція заснована на технологічних досягненнях напівпровідникової промисловості і є найбільшим сектором електронної промисловості [1].

У той час, як електронний бізнес займається всіма аспектами ведення онлайн-бізнесу, електронна комерція відноситься конкретно до угод між товарами і послугами.

Зростання електронної комерції змусило IT-персонал вийти за рамки проектування і обслуговування інфраструктури та розглянути численні

аспекти, орієнтовані на клієнтів: такі, як конфіденційність і безпека даних споживачів. При розробці IT-систем і додатків для електронної комерції необхідно враховувати нормативні вимоги, що стосуються управління даними, правил конфіденційності інформації, що дозволяє встановити особу і протоколи захисту інформації [2].

Є причина, по якій електронна комерція продемонструвала такий вибухове зростання за останні пару років. Дійсно, оскільки Інтернет стає важливою вимогою повсякденному житті, компанії вчаться користуватися численними перевагами електронної комерції. Такими як:

– глобальний ринок. Фізичний магазин завжди буде обмежений географічною зоною, яку він може обслуговувати. Інтернет-магазин або будь-який інший тип бізнесу електронної комерції, має весь світ в якості свого ринку;

– цілодобова доступність. Ще одна велика перевага ведення онлайн-бізнесу - це те, що він завжди відкритий;

– зниження затрат. Підприємства електронної комерції отримують вигоду від значно більш низьких експлуатаційних витрат. Оскільки немає необхідності наймати торговий персонал або підтримувати фізичну вітрину, основні витрати електронної комерції йдуть на складування і зберігання продуктів;

– управління запасами. Підприємства електронної комерції можуть автоматизувати управління запасами, використовуючи електронні інструменти для прискорення процедур замовлення, доставки і оплати. Це економить підприємствам мільярди на експлуатаційних витратах і витратах на запаси;

– цільової маркетинг. Маючи доступ до такої великої кількості даних про клієнтів і можливість стежити за купівельними звичками клієнтів, а також за новими галузевими тенденціями, підприємства електронної комерції можуть залишатися гнучкими і формувати свої маркетингові зусилля, щоб забезпечити більш індивідуальний досвід і знайти більше нових клієнтів.

– обслуговування нішевих ринків. Управляти нішевим бізнесом непросто. Масштабування нішевого продукту, щоб стати популярним, вимагає зусиль. З іншого боку, виходячи на світовий ринок, роздрібні продавці електронної комерції можуть побудувати високоприбутковий нішевий бізнес без будь-яких додаткових інвестицій. Використовуючи можливості онлайн-пошуку, клієнти з будь-якого куточка світу можуть знайти і придбати вашу продукцію;

– робота з будь-якого місця. Найчастіше ведення бізнесу електронної комерції означає, що вам не потрібно сидіти в офісі з 9 до 5 або мучитися день у день в дорозі. Ноутбук і хороше підключення до Інтернету - це все, що потрібно для управління своїм бізнесом з будь-якої точки світу.

Таким чином, електронна комерція – це просто процес покупки і продажу продукції за допомогою електронних засобів, таких як мобільні додатки і Інтернет. Електронна торгівля належить як до онлайн-роздрібної торгівлі, так і до онлайн-покупок, а також до електронних транзакцій.

Електронна комерція дозволяє купувати і продавати товари у глобальному масштабі двадцять чотири години на добу, не несучи при цьому тих же накладних витрат. Для кращої маркетингової комбінації і кращого коефіцієнта конверсії підприємство електронної комерції також повинно мати фізичну присутність [3].

1.2 Предиктивна аналітика

Предиктивна аналітика – це категорія аналітики даних, спрямована на прогнозування майбутніх результатів на основі історичних даних і таких аналітичних методів, як статистичне моделювання і машинне навчання. Наука предиктивної аналітики може генерувати майбутні ідеї зі значним ступенем точності. За допомогою складних інструментів і моделей предиктивної

аналітики будь-яка організація тепер може використовувати минулі і поточні дані для надійного прогнозування тенденцій і поведінки на мілісекунди, дні або роки в майбутньому.

Предиктивна аналітика отримала підтримку з боку широкого кола організацій, при цьому прогнозується, що до 2022 року світовий ринок досягне приблизно 10,95 млрд доларів США, а сукупний річний темп зростання (CAGR) буде складати близько 21 відсотка в період з 2016 по 2022 рік, згідно з опублікованим звітом 2017 року. компанії Zion Market Research.

В основі предиктивної аналітики лежить широкий спектр методів і технологій, включаючи Big Data, інтелектуальний аналіз даних, статистичне моделювання, машинне навчання і різні математичні процеси. Організації використовують прогнозну аналітику для аналізу поточних і історичних даних для виявлення тенденцій і прогнозування подій і умов, які повинні відбутися в певний час, на основі наданих параметрів.

За допомогою предиктивної аналітики організації можуть знаходити і використовувати шаблони, що містяться в даних, для виявлення ризиків і можливостей. Наприклад, моделі можуть бути розроблені для виявлення взаємозв'язків між різними факторами поведінки. Такі моделі дозволяють оцінювати або обіцянку, або ризик, пов'язаний з певним набором умов, направляючи інформоване ухвалення рішень по різним категоріям ланцюжків поставок і закупівель.

Предиктивна аналітика робить погляд в майбутнє більш точним і надійним, ніж попередні інструменти. Роздрібні продавці часто використовують прогнозні моделі для прогнозування потреб в товарних запасах, управління графіками відвантаження і настройки макетів магазинів для максимізації продажів. Авіакомпанії часто використовують предиктивну аналітику для визначення цін на квитки з урахуванням минулих тенденцій в сфері подорожей [4].

Шляхом оптимізації маркетингових кампаній за допомогою прогнозного аналізу, організації можуть також створювати нові відгуки

клієнтів або покупки, а також розширювати можливості перехресних продажів. Прогнозні моделі можуть допомогти підприємствам залучати, утримувати та збільшувати кількість своїх найцінніших клієнтів.

Предиктивна аналітика також може використовуватися для виявлення і припинення різних типів злочинної поведінки до того, як буде завдано будь-якої серйозної шкоди. Використовуючи предиктивну аналітику для вивчення поведінки і дій користувачів, організація може виявляти незвичайні дії, починаючи від шахрайства з кредитними картами і закінчуючи корпоративним шпигунством і кібератаками.

Інструменти предиктивної аналітики дозволяють користувачам в режимі реального часу отримувати глибоке уявлення про практично нескінченний спектр бізнес-операцій. Інструменти можуть використовуватися для прогнозування різних типів поведінки і шаблонів. Наприклад, як розподіляти ресурси в певний час, коли поповнювати запаси або коли найкраще запускати рекламну кампанію, ґрунтуючись на прогнозах на основі аналізу даних, зібраних за певний період часу. .

Практично всі прихильники предиктивної аналітики використовують інструменти, надані одним або декількома зовнішніми розробниками. Багато таких інструментів адаптовано до потреб конкретних підприємств і відділів. Основні постачальники програмного забезпечення і послуг для прогнозної аналітики:

- Acxiom;
- IBM;
- Information Builders;
- Microsoft;
- SAP;
- SAS Institute;
- Tableau Software;
- Teradata;
- TIBCO Software.

Моделі – це основа предиктивної аналітики, шаблони, які дозволяють користувачам перетворювати минулі і поточні дані в корисні ідеї, створюючи позитивні довгострокові результати.

Користувачі моделей мають доступ до майже нескінченного діапазону методів прогнозного моделювання. Багато методів унікальні для конкретних продуктів і послуг, але ядро загальних методів, таких як дерева рішень, регресія і навіть нейронні мережі, тепер широко підтримуються в широкому спектрі платформ предиктивної аналітики.

Дерева рішень, один з найпопулярніших методів, засновані на схематичній діаграмі в формі дерева, яка використовується для визначення курсу дій або для відображення статистичної ймовірності. Метод розгалуження також може показати всі можливі результати конкретного рішення і те, як один вибір може привести до іншого.

Методи регресії часто використовуються в банківських, інвестиційних та інших фінансових моделях. Регресія допомагає користувачам прогнозувати вартість активів і розуміти стосунки між змінними, такими, як ціни на товари і акції.

На передньому краї методів предиктивної аналітики знаходяться нейронні мережі – алгоритми, розроблені для визначення основних взаємозв'язків в наборі даних, імітуючі спосіб функціонування людського розуму.

Прихильники предиктивної аналітики мають легкий доступ до широкого спектру статистичних алгоритмів, алгоритмів інтелектуального аналізу даних і машинного навчання, призначених для використання в моделях прогнозного аналізу. Алгоритми зазвичай розробляються для вирішення конкретного бізнес-завдання або серії проблем, поліпшення існуючого алгоритму або надання деяких унікальних можливостей.

Наприклад, алгоритми кластеризації добре підходять для сегментації клієнтів, виявлення спільнот та інших завдань, пов'язаних з соціальними мережами. Щоб поліпшити утримання клієнтів або розробити систему

рекомендацій, зазвичай використовуються алгоритми класифікації. Алгоритм регресії зазвичай вибирається для створення системи кредитного рейтингу або для прогнозування результату багатьох подій, що залежать від часу.

Хоча предиктивної аналітика існує вже кілька десятиліть, час цієї технології прийшов. Все більше і більше організацій звертаються до предиктивної аналітики, щоб збільшити свій прибуток і конкурентну перевагу [5]. Причини цьому:

– зростаючі обсяги і типи даних, а також зростання інтересу до використання даних для отримання цінної інформації;

– більш швидкі і дешеві комп'ютери;

– програмне забезпечення, більш просте у використанні;

– більш жорсткі економічні умови і необхідність конкурентної диференціації.

Оскільки інтерактивне і просте у використанні програмне забезпечення стає все більш поширеним, прогнозна аналітика більше не є прерогативою математиків і статистиків. Бізнес-аналітики та бізнес-експерти також використовують ці технології.

Маркетингова атрибуція полягає в тому, щоб віддавати належне там, де це необхідно і незалежно від того, наскільки складна для користувача модель зважування, яка використовується у вашій платформі веб-аналітики. Привласнення балів маркетинговим контактам виключно на основі присутності, частоти і порядку точок взаємодії невірно. Частої присутності точки взаємодії в циклах взаємодії з клієнтом недостатньо, щоб гарантувати, що ця точка взаємодії призведе до покупки клієнта. Це заважає погодженню кореляції з причинно-наслідковим зв'язком.

Ґрунтуючись на цьому визначенні, будь-який підхід, який не враховує ймовірність конверсії певного типу клієнтів незалежно від каналу збуту, не зможе точно оцінити зростання продажів цих каналів збуту.

Предиктивна аналітика сприяє розвитку Big Data : компанії збирають величезну кількість даних про клієнтів у реальному часі, а предиктивна

аналітика використовує ці історичні дані в поєднанні з оглядом клієнтів для прогнозування майбутніх подій. Предиктивна аналітика дозволяє організаціям використовувати Big Data (як збережені, так і в режимі реального часу) для переміщення з історичного погляду на перспективу замовника [6].

Наприклад, магазини, які використовують дані програм лояльності, можуть аналізувати минулу поведінку покупки, щоб передбачити купони або акції, в яких клієнт найбільше бере участь або купує в майбутньому. саналітика також може бути застосована до поведінки веб-переглядачів клієнтів, щоб забезпечити клієнту персоналізований досвід роботи

Незважаючи на те, що предиктивна аналітика швидко набирає популярність у галузі Big Data, компанії не поспішають впроваджувати технологію насамперед через причини складності та недоступності. Однак у міру просування ринку аналітики даних розробляються більш доступні та прості у використанні рішення. Ці більш універсальні рішення можуть бути інтегровані компаніями з електронної комерції на різних платформах.

Предиктивна аналітика надає підприємствам електронної комерції більш глибоке розуміння звичок та переваг клієнтів

Ринок роздрібної торгівлі в Інтернеті розвивається стрімкими темпами, і клієнти активно шукають більш привабливий досвід роздрібної торгівлі. Щоб досягти успіху на високодинамічному ринку, підприємства електронної комерції повинні мати можливість бути на крок попереду своїх клієнтів. Вони повинні бути в змозі передбачити, що клієнти шукають у своєму магазині електронної комерції [7].

Можливості прогнозованого пошуку, які можуть бути вбудовані у аналітичне рішення, дозволять підприємствам електронної комерції в режимі реального часу аналізувати їхні поведінки, що відбулися при натисканні, історію покупок та налаштування продукту.

Предиктивна аналітика дозволить здійснювати постійний аналіз даних клієнтів, тоді як можливості машинного навчання надаватимуть найбільш релевантні результати та рекомендації користувачам.

Не всі клієнти взаємодіють із магазином електронної комерції однаково. Кожен клієнт унікальний, а їхня поведінка в Інтернеті буде відрізнятися залежно від індивідуальних смаків та уподобань. Предиктивна аналітика допомагає оцінити різні змінні елементи в поведінці клієнтів. Це призведе до бажаного залучення та відповідей від клієнта, зробивши їх електронну комерцію дуже персоналізованою.

Предиктивна аналітика також може допомогти бізнесу електронної комерції визначити оптимальні ціни на свою продукцію шляхом ефективного аналізу настроїв клієнтів щодо ціноутворення

Підприємства електронної комерції можуть використовувати потенціал прогностичної предиктивної аналітики, щоб запропонувати розширені рекомендації щодо продуктів та акцій

Подібно до того, як торговий представник може давати персоналізовані рекомендації потенційним клієнтам у фізичному роздрібному магазині, клієнти в магазинах електронної комерції мають такі ж рекомендації. У цю все більш цифрову епоху, коли більшість клієнтів віддають перевагу покупкам в Інтернеті, не виходячи з дому чи офісу, відповідні рекомендації щодо продуктів швидко стали головним фактором успіху бізнесу в галузі електронної комерції.

Щоб забезпечити здоровий коефіцієнт конверсії у своїх магазинах електронної комерції, інтернет-магазини докладають концентрованих зусиль для покращення можливостей рекомендацій щодо продуктів на своїх платформах електронної комерції. Використовуючи потенціал аналітики, інтернет-роздрібні торговці можуть отримати відповідну інформацію про окремих клієнтів. Це допоможе їм запропонувати цільові рекомендації щодо товарів на основі аналізу минулої історії покупок, моделей перегляду магазинів та найпопулярніших продуктів або предметів у конкретному ціновому діапазоні [8].

Таким чином, предиктивна аналітика дозволяє бізнесу електронної комерції приймати більш швидкі, більш відповідні критичні бізнес-рішення, які позитивно впливатимуть на результативність бізнесу.

Впровадження предиктивної аналітики як частини аналітичного рішення, що використовується електронною комерцією, може призвести до значної конкурентної переваги для роздрібної торгівлі електронною комерцією. Однак перед остаточним розгортанням моделі предиктивної аналітики повинні бути ретельно перевірені, щоб переконатися, що вони функціонують так, як очікувалося [9].

Підприємства електронної комерції також повинні періодично контролювати моделі предиктивної аналітики, щоб мінімізувати можливість помилок в аналізі даних.

1.3 Big data в електронні комерції

Аналітика Big Data – це можливість вивчення великої кількості даних з метою виявлення прихованих закономірностей, кореляцій, ринкових тенденцій, переважних споживачів та інших ідей, які можуть допомогти підприємствам відповідати певнім змінам. Данні різняться по об'єму, різноманітності та швидкості. Це можуть бути структуровані дані, такі як сховища даних SQL; неструктуровані дані, такі як файли документів; або потокова передача даних із датчиками в реальному часі, що використовується в IoT. Хоча в цій концепції немає нічого нового, через те, що компанії, які раніше аналізували свої дані вручну, з Big Data отримують безліч переваг. Це такі, як економія часу та витрат, ефективність, більш розумні бізнес-шаги, більш ефективні операції, більший високий прибуток і більші щасливі клієнти [10].

Big Data складаються з двох типів цінної інформації: структурованої та неструктурованої. Структуровані дані - це звичайна інформація, яка містить ім'я та адресу. Неструктуровані дані збираються з таких місць, як соціальні мережі, і включає лайки, твіти, кліки та відео. Ця інформація важлива для підприємства електронної комерції, оскільки вона може оптимізувати обслуговування клієнтів.

Інтелектуальний аналіз даних за допомогою Big Data допомагає краще розуміти покупця. Покупець може не визначити, що кожне його клацання постійно відслідковується. Все, що досліджує покупець, записується від входу до виходу. Виходячи з цього, продавець виявляє їх інтерес до перегляду та історії покупок і отримує представлення про їх покупців моделі: кількість їх спроб і коли вони купують більше всього [11].

З Big Data продавці електронної комерції також виграли в управлінні своїми запасами. Тепер вони можуть організовувати клієнтів на основі їхніх покупецьких введень, демографічних даних та періоду часу і, відповідно, звільнятися від своїх надлишкових запасів. Як у вечірній час, так і у святкові сезони, здійснюються найвищі високі покупки, тому деякі власники інтернет-магазинів змінюють ціни на свої продукти / послуги протягом цього періоду часу і навіть запускають маркетингову рекламу в соціальних мережах для залучення клієнтів. І останнє, але не менш важливе: Big Data дозволяють легко перенацілювати клієнтів. Рекомендуючи товари, що зацікавили покупця, електронна комерція забезпечує більш високий рівень задоволеності клієнтів на основі оптимального клієнтського досвіду.

Ще одне з переваг с для підприємства електронної торгівлі полягає в тому, що магазини електронної торгівлі збільшують продажі. За допомогою аналізу повідомлень клієнтів інтернет-магазинів тепер можна знайти свої пропозиції та рекомендації, забезпечуючи більш інтерактивний та різноманітний клієнтський досвід. Проводячи кампанії, надаючи купони та скидки на основі минулих звітів про витрати, вони тепер приваблюють величезний трафік та хороші доходи.

Кінцевий успіх полягає в тому, щоб відчувати клієнта. Погане обслуговування може назавжди відлякати клієнта. У подальшому, щоб продовжити, треба надати їм найкраще, і Big Data були б дуже корисними. Використовуючи дані Big Data , інтернет-магазини постійно слідують за досвідом покупця, щоб краще реагувати на їх потреби. Від відповіді на їх запрошення до інформування про нові пропозиції та відстеження їх товарів - Big Data допомагає власникам інтернет-магазинів підтримувати більш прості та довгострокові відносини з клієнтами. Врешті, чим більш задовільні клієнти, тим більше вони будуть користуватися пропонованими послугами.

Одно з найкращих переваг Big Data полягають у тому, що тепер Інтернет-магазини можуть відслідковувати свої запаси та своєчасно забезпечуватися запасами, щоб задовольнити потреби клієнтів. Постачальнику електронної комерції стало надзвичайно вигідним знати спроби продукту на їх веб-сайті та використовувати цю можливість, щоб закупати продукти, які шукає покупець і яких зараз немає в магазині [12].

Big Data залучені до неймовірних змін у промисловості електронної комерції, але є безкоштовні можливості, які ще потребують вивчення. Багато компаній вже почали використовувати аналіз в режимі реального часу для вдосконалення продажу та збільшення прибутку, надаючи обслуговування клієнтам на місці. Наступною найбільшою подією стане об'єднання Інтернету речовин та Big Data. Результатом цієї вдалої комбінації буде виведення промислової електронної комерції на досконало новий рівень.

1.4Постановка задачі

Метою атестаційної роботи є дослідження моделей та алгоритмів предиктивного аналізу для використання та застування у системах електронної комерції. Проблема, яка вирішується в даній роботі – це те, що в

електронній комерції дуже малий відсоток застосування аналізу даних, зібраних про користувачів та незнання того, що робити з цими даними для покращення продажів. Вирішенням проблеми є створення універсального алгоритму для обробки великих масивів даних зібраних про користувачів для створення прогностичних звітів щодо шляхів покращення користувальницького досвіду та підвищення прибутку компанії.

Задачею атестаційної роботи є:

– проведення аналізу предметної галузі, де буде досліджений стан використання предиктивного аналізу на поточний час;

– дослідження основних метрик електронної комерції, які є актуальними для використання предиктивного аналізу;

– дослідження різних моделей роботи предиктивного аналізу та шляхи їх застосування в електронній комерції;

– дослідження особливостей збору великих наборів даних про покупців та алгоритмів обробки таких даних;

– створення алгоритму для прогнозування стану системи електронної комерції в найближчому майбутньому;

– донесення мети про можливість інтегрування предиктивного аналізу у системи електронної комерції різного рівня та розміру.

Результатом проведеного дослідження буде актуалізація проблеми використання предиктивного аналізу в електронній комерції та рекомендації щодо шляхів ефективної обробки великої кількості даних, зібраних про користувачів.

2 АНАЛІЗ МЕТОДІВ ДОСЛІДЖЕННЯ

2.1 Методи дослідження

Для даного дослідження був обраний теоретичний метод дослідження, тому що він пов’язаний з більш глибоким аналізом фактів, з проникненням у сутність досліджуваних явищ, з пізнанням та формулюванням законів, тобто з поясненням реальної дійсності. Теоретичний метод дослідження спрямований на саме аналіз фактів, статистики та результатів.

Методи теоретичного рівня: абстрагування, ідеалізація, формалізація, аналіз і синтез, індукція і дедукція, аксіоматика, узагальнення та ін. На теоретичному рівні проводяться логічні дослідження зібраних фактів, розробка понять, суджень та виконання умовиводів. У процесі цієї роботи співвідносяться попередні наукові уявлення з новими, що виникають. На теоретичному рівні наукове мислення звільняється від емпіричного опису, створюється теоретичне узагальнення. Таким чином, новий теоретичний зміст знань надбудовується над емпіричними знаннями. На теоретичному рівні пізнання науковці використовують логічні методи подібності або відмінності, розробляють нові системи знань або вирішують завдання подальшого узгодження теоретично розроблених систем з накопиченими новими експериментальними результатами.

2.2 Аналіз метрик

Метрика – це послідовно визначаємий вимір продуктивності веб-сайту, який піддається кількісній оцінці. Приклади відповідних показників електронної торгівлі варіюються від коефіцієнта конверсії до середньої вартості замовлення, від коефіцієнта відмови від кошика до джерел трафіку.

Google Analytics, соціальні мережі, інтернет-магазин, сторінки продуктів, домашні сторінки, каси і візки для покупок – все це багаті джерела даних, які збирають, піддаються кількісній оцінці. Дані є готові для інтерпретації та вимірювання тенденцій з плином часу [13].

КПЕ – це ключовий показник ефективності. У той час як метрика - це будь-який кількісний вимір, КПЕ – важлива метрика. Це цифри, які відстежуються на предмет зростаннях [14].

Хоча відвідування сайтів може мати важливе значення, замовлення можуть бути ключовим показником ефективності. Зазвичай оцінка відбувається по невеликій кількості критично важливих чисел. Це і є КПЕ.

Різниця між метрикою і КПЕ, в тому, що метрики вимірюють процес, а КПЕ вимірюють продуктивність цих процесів. Іншими словами, ключові показники ефективності – це суб'єктивні, конкретні цілі, які повинен досягти магазин. Далі будуть перераховані основні показники ефективності електронної торгівлі.

2.2.1 Середня вартість придбання

Вартість залучення клієнтів є критично важливим показником для підприємств електронної комерції. Він вимірює середні граничні витрати на придбання одного додаткового клієнта.

Він використовується для того, щоб визначити, чи ефективно використовується маркетинговий бюджет, коли про нього йдеться, у поєднанні із середньою вартістю замовлення та ціною проведеного часу клієнтом на сайті при аналізі прибутковості відділу.

Врешті-решт, це показник успіху чи невдачі спільних маркетингових зусиль, і коли останні є очевидним, це сигналізує про необхідність повного оновлення вашої маркетингової стратегії.

$$\frac{\text{Загальна вартість маркетингу}}{\text{Загальна кількість нових покупців}} = \text{Вартість придбання клієнта}$$

Середня вартість придбання – це гра чисел. Простіше кажучи, вартість придбання клієнта обчислюється шляхом складання всіх маркетингових витрат та усереднення цієї кількості на кожного нового клієнта:

2.2.2 Довічна цінність клієнта

У маркетингу ціна за життя клієнта, ціна клієнта протягом усього життя або вартість за весь час – це прогнозування чистого прибутку, що приписується всім майбутнім відносинам із клієнтом. Модель прогнозування може мати різний рівень витонченості та точності, починаючи від грубої евристики до використання складних методів прогнозної аналітики.

Довічна вартість клієнта також може бути визначена як грошова оцінка відносин із клієнтом, виходячи із теперішньої вартості прогнозованих майбутніх грошових потоків від відносин із клієнтами.

Цінність клієнта протягом усього життя є важливою метрикою, оскільки вона являє собою верхню межу витрат на залучення нових клієнтів. З цієї причини це важливий елемент при розрахунку окупності реклами, витраченої на моделювання маркетингової суміші.

2.2.3. Середня вартість замовлення

Середня вартість замовлення (СВЗ) – це показник електронної комерції, який вимірює середню загальну суму кожного замовлення, розміщеного у продавця протягом певного періоду часу. СВЗ – це одна з найважливіших

метрик, яку повинні знати інтернет-магазини, визначаючи такі ключові бізнес-рішення, як витрати на рекламу, макет магазину та ціни на товари.

Формула для розрахунку СВЗ – це дохід, поділений на кількість замовлень.

$$\frac{\text{Дохід}}{\text{Кількість замовлень}} = \text{Середня вартість замовлення}$$

СВЗ визначається з використанням продажів за замовлення, а не продажів за клієнта. Хоча один клієнт може повертатися кілька разів, щоб зробити покупку, кожне замовлення буде враховано в СВЗ окремо.

Середня вартість замовлення не описує валовий прибуток або норму прибутку, але дає уявлення про те, як ці цифри виникають.

2.2.4 Рівень утримання та частка постійних клієнтів

Відсоток клієнтів, які повертаються – це співвідношення людей, які здійснили первинну покупку, до тих, хто повернувся та зробив другу (або третю чи четверту) покупку. Це відсоток повторних клієнтів.

$$\frac{\text{Клієнти, які повертаються}}{\text{Загальна кількість споживачів}} * 100 = (\%)$$

Щоб розрахувати відсоток клієнтів, що повертаються, просто розділіть кількість клієнтів, що повертаються, на загальну кількість клієнтів і помножте на 100, щоб перетворити на відсоток. Це можна розрахувати на основі різноманітних часових рамок, таких як щодня, щотижня або щомісяця.

2.2.5 Коефіцієнт конверсії

Ефективність маркетингу конверсій вимірюється коефіцієнтом конверсії: кількість клієнтів, які здійснили транзакцію, поділена на загальну кількість відвідувачів веб-сайту. Коефіцієнти конверсії електронних вітрин зазвичай низькі. Маркетинг на основі конверсій може збільшити це число, а також доходи в Інтернеті та відвідуваність веб-сайтів [15].

Маркетинг конверсій намагається вирішити низькі онлайн-конверсії за допомогою оптимізованого обслуговування клієнтів, що вимагає складної комбінації персоналізованого управління досвідом клієнтів, веб-аналітики та використання зворотного зв'язку із клієнтами, щоб сприяти вдосконаленню потоків процесів та дизайну сайту.

Зосередження на покращанні потоку веб-сайтів, онлайн-каналах обслуговування клієнтів та онлайн-маркетингу конверсій досвіду, зазвичай розглядається як довгострокова інвестиція, а не як швидке вирішення. Збільшення відвідуваності сайту за останні 10 років мало що зробило для збільшення загальної конверсії ставки, тому маркетинг конверсій зосереджується не на залученні додаткового трафіку, а на перетворенні наявного трафіку.

Це вимагає активної взаємодії зі споживачами за допомогою аналітики в режимі реального часу, щоб визначити, чи відвідувачі розгублені та чи мають ознаки відмови від сайту.

Потім необхідно розробляти інструменти та повідомлення, щоб інформувати споживачів про наявні товари, і в кінцевому підсумку переконати їх здійснити онлайн-конверсію.

В ідеалі клієнт підтримуватиме стосунки після продажу через кампанії підтримки або повторного залучення. Маркетинг конверсій впливає на всі фази життєвого циклу клієнта, і для полегшення переходу від однієї фази до іншої використовується кілька рішень маркетингових конверсій.

2.2.6. Середня норма прибутку

Валовий прибуток – це показник, який вимірює прибуток як частку доходу. Більш висока валова рентабельність або норма валової рентабельності означає, що бізнес є високорентабельний. Нижча валова рентабельність або норма валової рентабельності означає, що отримується менше прибутку з кожною проданою одиницею.

Валова маржа розраховується як:

$$\text{Дохід} - \text{ вартість реалізованих товарів} = \text{Валова націнка}$$

Вартість реалізованих товарів включає купівельну вартість товару, а також будь-які інші витрати, які безпосередньо пов'язані з товаром, такі як доставка, імпортні мита та складування.

Як правило, валова маржа вказується як відсоток від доходу. У цьому випадку це називається "ставка валової рентабельності":

$$\left(\frac{\text{Валова маржа}}{\text{Дохід}}\right) * 100$$

Однак маркетингові витрати, витрати на персонал та інші операційні витрати до вартості реалізованих товарів не включаються.

2.2.7. Рівень відмови від кошика

Коефіцієнт відмови від кошика – це відсоток покупців в Інтернеті, які додають товари у віртуальний кошик, а потім відмовляються від нього перед

завершенням покупки. Він показує рівень зацікавлених потенційних клієнтів, які виїжджають, нічого не купуючи, порівняно із загальною кількістю створених кошиків.

$$\left(1 - \left(\frac{\text{здійснені покупки}}{\text{створені кошикидля покупок}}\right)\right) * 100 = \text{коефіцієнт відмови від кошика}$$

Швидкість відмови від кошика обчислюється шляхом ділення загальної кількості здійснених покупок на кількість створених кошиків. Віднімітьрезультат від одиниці, а потім помножте на 100 для показника відмови.

2.2.8 Важливість метрик для предиктивного аналізу

Робота з предиктивною аналітикою тісно пов’язана з метриками та ключовими показниками ефективності, які необхідні для подальшого створення прогнозів. Дані цих показників фіксуються різними системами обліку, включаючи CRM та програмне забезпечення для бізнес-аналітики. CRM зберігають велику кількість даних про кожного клієнта сайту, його контактні дані та історію активності на сайті. Саме за допомогою виділення ключових метрик, вирішується проблема розподілу інформації для аналізу на корисну та ту, яка не грає ролі.

При створенні прогнозів важливо виділити в великих наборах даних лише ті показники, несуть в собі інформацію, прогноз якої буде актуальним на даний момент для цієї компанії. Це треба зробити для того, щоб скоротити час обробки інформації алгоритмом. Наприклад, якщо ми маємо 30 колонок інформації зібраних про користувачів, але лише 10 з них стосуються КПЕ є сенс не обробляти цю інформацію при створенні прогнозу, а обробляти лише ті показники які на даний час є провідними для бізнесу.

Основною метою компанії при роботі з предиктивною аналітикою повинно бути створення актуальних та чітких звітів для подальшого планування розвитку бізнесу.

2.3 Аналіз методів обробки даних

Інструменти предиктивної аналітики засновані на декількох різних моделях і алгоритмах, які можна застосовувати до широкого спектру сценаріїв використання.

2.3.1 Модель класифікації

Модель класифікації в деякому сенсі є найпростішою з декількох типів моделей предиктивної аналітики. Вона поміщає дані в категорії на основі того, що дізнається з історичних даних.

Класифікаційні моделі найкраще підходять для відповіді на питання «так» чи «ні», забезпечуючи широкий аналіз, який допомагає направляти рішучі дії. Широта можливостей моделі класифікації і легкість, з якою її можна перенавчити з новими даними – означає, що її можна застосовувати в багатьох різних галузях.

Моделі класифікації кількісно визначають взаємозв'язки в даних способом, який часто використовується для класифікації клієнтів або потенційних клієнтів по групах. Класифікаційні моделі не ранжують клієнтів по ймовірності вчинення певної дії. Модулі можна використовувати для категоризації клієнтів по їх перевагам щодо продуктів і стадіями життя. Інструменти класифікаційного моделювання можуть використовуватися для

розробки додаткових моделей, які можуть моделювати велику кількість індивідуальних агентів і робити прогнози.

2.3.2 Модель кластеризації

Модель кластеризації сортує дані в окремі вкладені смарт-групи на основі схожих атрибутів. Якщо взуттєва компанія електронної комерції хоче реалізувати цільові маркетингові кампанії для своїх клієнтів, вони можуть переглянути сотні тисяч записів, щоб створити індивідуальну стратегію для кожної людини

Використовуючи модель кластеризації, можна швидко розділяти клієнтів на схожі групи на основі загальних характеристик і розробляти стратегії для кожної групи в більшому масштабі.

Інші варіанти використання цього методу прогнозного моделювання можуть включати в себе угруповання здобувачів кредиту в «розумні кошика» на основі атрибутів позики, визначення районів в місті з високим рівнем злочинності та порівняльний аналіз даних клієнтів SaaS по групах для визначення глобальних моделей використання.

2.3.3 Модель прогнозу

Одна з найбільш широко використовуваних моделей предиктивної аналітики – прогнозна модель має справу з прогнозуванням значень показників, оцінюючи числове значення для нових даних на основі вивчення історичних даних.

Ця модель може застосовуватися всюди, де доступні історичні числові дані.

Прогнозні моделі використовуються для аналізу взаємозв'язку між конкретними характеристиками одиниці у вибірці і одним або декількома відомими атрибутами або характеристиками одиниці. Мета моделі – оцінити ймовірність того, що аналогічний пристрій в іншому зразку буде демонструвати певні характеристики. Прогностичні моделі часто виконують обчислення під час реальних транзакцій, наприклад, для оцінки ризику або можливості даного клієнта або транзакції, щоб прийняти рішення.

Доступні одиниці вибірки з відомими атрибутами і відомими характеристиками називаються «навчальною вибіркою». Одиниці в інших вибірках з відомими атрибутами, але невідомими характеристиками, називаються одиницями «поза вибірки». Одиниці поза вибірки не обов'язково мають хронологічну зв'язок з одиницями навчальної вибірки.

2.3.4 Модель викидів

Модель викидів орієнтована на аномальні записи даних в наборі даних. Він може ідентифікувати аномальні фігури або самі по собі, або в поєднанні з іншими числами і категоріями.

– запис сплеску звернень в службу підтримки, який може вказувати на збій продукту, який може привести до відкликання;

– виявлення аномальних даних в транзакціях або в страхових випадках для виявлення шахрайства;

– виявлення незвичайної інформації в журналах NetOps і виявлення ознак наближення незапланованого простою.

Модель викидів особливо корисна для предиктивної аналітики в роздрібній торгівлі та фінансах. Наприклад, при виявленні шахрайських

транзакцій модель може оцінювати не тільки суму, але також місце, час, історію покупок і характер покупки.

2.3.5 Модель часових рядів

Модель часових рядів складається з послідовності точок даних, отриманих з використанням часу в якості вхідного параметра. Він використовує дані за останній рік для розробки числової метрики і прогнозує наступні три-шість тижнів даних з використанням цієї метрики. Часовий ряд, як правило, моделюється через стохастичний процес Y (t), тобто послідовність випадкових величин. В умовах прогнозування ми опиняємося в момент часу t, і ми зацікавлені в оцінці Y (t + h), використовуючи лише інформацію, доступну в момент часу t.

Сценарії використання цієї моделі включають кількість щоденних дзвінків, отриманих за останні три місяці, продажу за останні 20 кварталів або кількість пацієнтів, які звернулися в конкретну лікарню за останні шість тижнів. Це потужний засіб розуміння того, як одинична метрика розвивається з плином часу, з точністю, що перевищує прості середні. Він також враховує пори року або події, які можуть вплинути на показник.

Якщо власник салону хоче передбачити, скільки людей, ймовірно, відвідають його бізнес, він може вдатися до грубого методу усереднення загальної кількості відвідувачів за останні 90 днів. Однак зростання не завжди є статичним або линійним, і модель часових рядів може краще моделювати експоненціальне зростання і краще узгоджувати модель з тенденцією компанії. Власник також може прогнозувати кілька проектів або декількох регіонів одночасно, а не тільки по одному за раз.

2.3.6 Порівняння методів обробки даних

Модель прогнозу є однією з найпоширеніших моделей прогнозної аналітики. Вона обробляє прогнозування метричних значень шляхом оцінки значень нових даних на основі результатів історичних даних. Вона часто використовується для генерації числових значень в історичних даних, коли таких немає. Однією з найбільших сильних сторін даної моделі є її здатність вводити кілька параметрів.

Класифікаційна модель є одною з найпоширеніших моделей прогнозної аналітики. Ці моделі працюють шляхом класифікації інформації на основі історичних даних. Моделі класифікації можуть застосовуватися в різних галузях, таких як фінанси та роздрібна торгівля, що пояснює, чому вони настільки поширені в порівнянні з іншими моделями.

Хоча класифікаційні та прогнозні моделі працюють з історичними даними, модель викидів працює з аномальними записами даних у наборі даних. Як випливає з назви, аномальні дані стосуються даних, що відхиляються від норми. Попередні моделі корисні в галузях, де виявлення аномалій може заощадити організаціям мільйони доларів, а саме у роздрібній торгівлі та фінансах. Однією з причин того, чому дана модель настільки ефективна у виявленні шахрайства, є те, що вона, може бути використана для виявлення аномалій. Оскільки частота шахрайства є відхиленням від норми, ця модель швидше за все прогнозує шахрайство до його виникнення

У той час як класифікаційні та прогнозні моделі зосереджуються на історичних даних, модель часових рядів зосереджується на даних про аномалії. Модель часових рядів фокусується на даних, де час є вхідним параметром. Модель часових рядів працює, використовуючи різні точки даних (взяті з даних попереднього року) для розробки числової метрики, яка буде прогнозувати тенденції протягом певного періоду.

Модель кластеризації бере дані та сортує їх у різні групи на основі загальних атрибутів. Можливість розділити дані на різні набори даних на основі конкретних атрибутів особливо корисна в певних програмах, таких як маркетинг.

Моделі предиктивної аналітики не є монолітом. Існують різні моделі, розроблені для конкретних функцій та дизайну. Іншими словами можна сказати, що вибір конкретної моделі залежить від архітектури системи та потреб бізнесу. Оскільки при аналізі використовуються метрики та КПЕ, саме від обраного показнику залежить, яку модель варто застосувати при створенні прогнозу.

2.4 Алгоритми прогнозування

В цілому алгоритми предиктивної аналітики можна розділити на дві групи: машинне навчання і глибоке навчання.

Машинне навчання включає в себе структурні дані, які ми бачимо в таблиці. Алгоритми для цього включають як лінійні, так і нелінійні різновиди. Лінійні алгоритми навчаються швидше, тоді як нелінійні краще оптимізовані для вирішення проблем, з якими вони можуть.

Глибоке навчання – це різновид машинного навчання, який більш популярний для роботи з аудіо, відео, текстом і зображеннями [16].

При прогнозному моделюванні машинного навчання можна застосовувати кілька різних алгоритмів. Універсального алгоритму предиктивної аналітики не існує, тому різні моделі мають свої сильні та слабкі сторони. Нижче наведені деякі з найбільш поширених алгоритмів, які використовуються в моделях прогнозної аналітики, описаних вище.

2.4.1 Random Forest

Random Forest, мабуть, найпопулярніший алгоритм класифікації, здатний як до класифікації, так і до регресії. Він може точно класифікувати великі обсяги даних.

Назва «Random Forest» походить від того факту, що алгоритм є комбінацією дерев рішень. Кожне дерево залежить від значень випадкового вектора, що відбирається незалежно з однаковим розподілом для всіх дерев в «лісі». Кожен з них вирощений в максимально можливій мірі.

Алгоритми предиктивної аналітики намагаються досягти мінімально можливої помилки або за допомогою «підвищення» (метод, який регулює вага спостереження на основі останньої класифікації), або «підсумовування» (який створює підмножини даних з навчальних вибірок, обраних випадковим чином із заміною). Random Forest використовує мішки.

Якщо є багато вибіркових даних, замість того, щоб тренуватися з усіма з них, можна взяти підмножину і потренуватися на ній, а також взяти іншу підмножину і навчитися на ній (допускається перекриття). Все це можна робити паралельно. Для отримання середнього значення з даних беруться кілька вибірок (рис. 2.1).

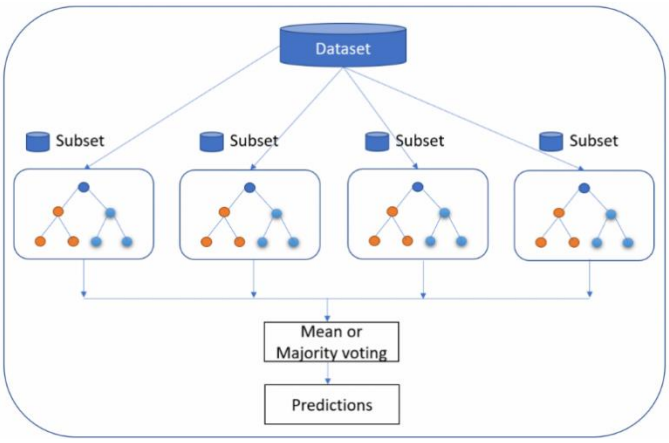

Рисунок 2.1 – Діаграма роботи алгоритму «Random Forest»

Хоча окремі дерева можуть бути «слабкими учнями», принцип випадкового лісу полягає в тому, що разом вони можуть становити одного «сильного учня».

Популярність моделі Random Forest пояснюється її різними перевагами:

– точність і ефективність при роботі з великими базами даних;

– кілька дерев зменшують дисперсію і зміщення меншого набору або одного дерева;

– стійкий до перенавчання;

– може обробляти тисячі вхідних змінних без видалення змінних;

– може оцінити, які змінні важливі при класифікації;

– надає ефективні методи оцінки відсутніх даних;

– зберігає точність, коли велика частина даних відсутня.

2.4.2 Узагальнена лінійна модель для двох значень

Узагальнена лінійна модель (Generalized Linear Model) – більш складний варіант загальної лінійної моделі. Остання модель вимагає порівняння впливу декількох змінних на безперервні змінні, перш ніж витягувати з масиву різних розподілів, щоб знайти модель «найкращої відповідності».

Припустимо, необхідно дізнатися, як покупці купують зимові пальта. Регулярна лінійна регресія може показати, що на кожний негативний градус різниці в температурі купується додатково 300 зимових пальто. Хоча здається логічним, що ще 2100 пальто можуть бути продані, якщо температура впаде з 9 градусів до 3, менш логічно, що якщо вона впаде до -20, ми побачимо, що число збільшиться точно в тій же мірі.

Узагальнена лінійна модель звузила б список змінних, ймовірно, припускаючи, що є збільшення продажів за межами певної температури і зменшення або вирівнювання продажів при досягненні іншої температури.

Перевага цього алгоритму в тому, що він дуже швидко навчається. Мінлива відповіді може мати будь-яку форму експоненціального типу розподілу. Узагальнена лінійна модель також може мати справу з категоріальним предикторами, хоча її відносно просто інтерпретувати. Додатково до всього, вона забезпечує чітке розуміння того, як кожен із предикторів впливає на результат, і досить стійкий до перенавчання. Однак для цього потрібні відносно великі набори даних, і лінійна модель схильна до викидів.

2.4.3 Модель з градієнтним посиленням

Модель з градієнтним посиленням створює модель прогнозування, що складається з ансамблю дерев рішень (кожне з яких є «слабким учнем», як у випадку з Random Forest), перш ніж узагальнювати (рис. 2.2). Як випливає з назви, вона використовує техніку «прискореного» машинного навчання, на відміну від упаковки, що використовується в Random Forest. Вона використовується для моделі класифікації.

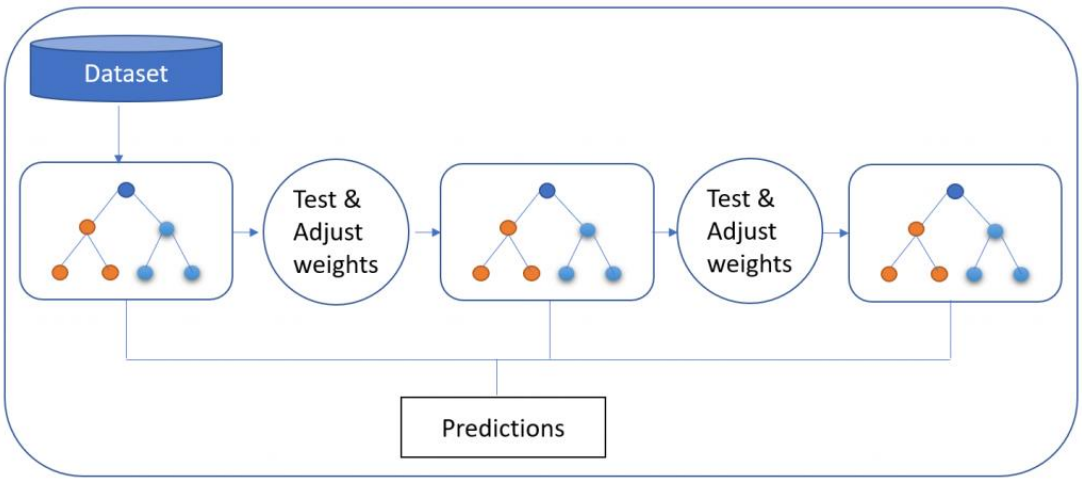

Рисунок 2.2 – Діаграма роботи алгоритму «Модель з градієнтним посиленням»

Відмінною рисою GBM є те, що вона будує свої дерева по одному за раз. Кожне нове дерево допомагає виправити помилки, зроблені раніше навченим дерева – на відміну від моделі випадкового лісу, в якій дерева не мають відношення. Вона дуже часто використовується в рейтингу з машинним навчанням, наприклад, в пошукових системах Yahoo і Яндекс.

Завдяки підходу МГП дані стають більш виразними, а результати порівняльного аналізу показують, що метод МГП краще з точки зору загальної повноти даних. Однак, оскільки кожне дерево будується послідовно, це також може тривати довше. Проте, вважається, що його більш низька продуктивність призводить до кращого узагальнення.

2.4.4 Метод k-середніх

Дуже популярний і високошвидкісний алгоритм K-середніх включає в себе розміщення немаркованих точок даних в окремі групи на основі подібності.

Цей алгоритм використовується для моделі кластеризації. Наприклад, користувач 1 і користувач 2 входять в першу групу, а користувач 3 і користувач 4 – в другу. Користувач 1 і користувач 2 мають дуже схожі характеристики, але користувач 2 і користувач 3 мають дуже різні характеристики.

Метод k-середніх намагається з'ясувати, якими є загальні характеристики окремих людей, і групує їх разом (рис 2.3).

Це особливо корисно, коли у вас великий набір даних і ви хочете реалізувати індивідуальний план – це дуже складно зробити з одним мільйоном чоловік.

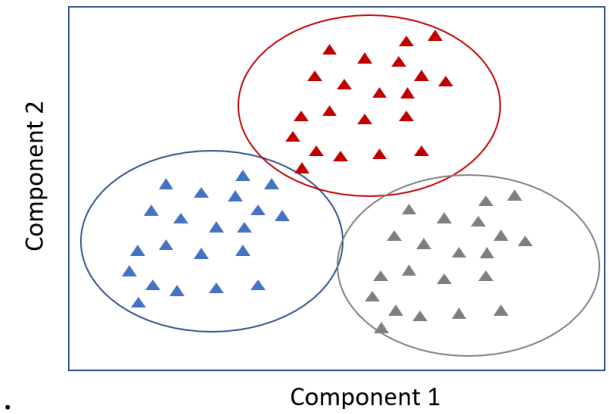

Рисунок 2.3 – Діаграма роботи алгоритму «Метод k-середніх»

Метод k-середніх повідомляє, що прогноз для елемента повинен бути середнім значенням n-найближчих елементів до цього елементу на основі функцій наборів. KNN працює для додатків класифікацій та регресій, його можна швидко навчати та легко реалізовувати, але при тестуванні більших наборів даних він може бути повільним.

2.4.5 Prophet

Алгоритм Prophet використовується в моделях часових рядів і прогнозів. Це алгоритм з відкритим вихідним кодом, розроблений Facebook, який використовується всередині компанії для прогнозування (див. рис. 2.4).

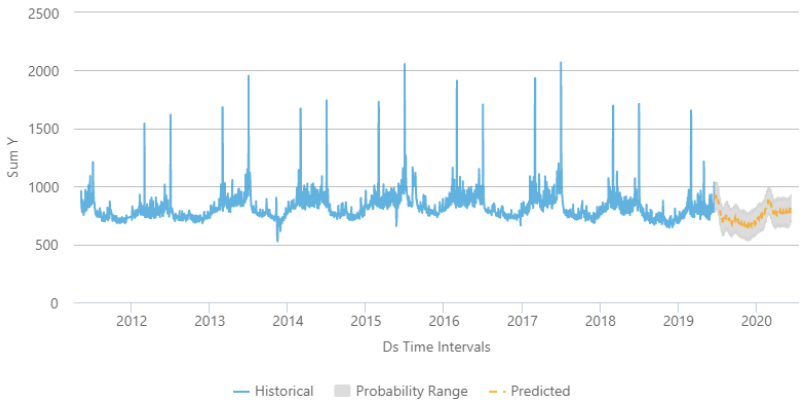

Рисунок 2.4 – Діаграма роботи алгоритму «Prophet» за роками

Алгоритм Prophet дуже корисний при плануванні потужності, наприклад при розподілі ресурсів та встановлення цілей продажів. Через непослідовність рівня продуктивності повністю автоматизованих алгоритмів прогнозування та їх негнучкості успішно автоматизувати цей процес було складно. З іншого боку, прогнозування вручну вимагає багатогодинної праці висококваліфікованих аналітиків. Prophet не просто автоматичний; він також досить гнучкий, щоб включати евристику і корисні припущення. Швидкість, надійність і стійкість алгоритму при роботі з безладними даними зробили його популярним альтернативним вибором для часових рядів і моделей прогнозування. Це цінне як для експертів-аналітиків, так і для менш досвідчених у прогнозуванні.

2.5 Висновки та майбутні перспективи

Існує довга історія використання прогнозних моделей у завданнях прогнозування. Раніше статистичні моделі використовувались як прогностичні моделі, які базувались на вибіркових даних великого набору даних. З удосконаленням у галузі інформатики та розвитком комп’ютерних технік були розроблені новіші техніки, і впродовж певного періоду впроваджувалися все кращі та кращі алгоритми. Розробки в галузі штучного інтелекту та машинного навчання змінили світ обчислень, де впроваджуються інтелектуальні обчислювальні методи та алгоритми. Моделі машинного навчання мають дуже хороший досвід використання їх як прогнозних моделей. Штучні нейронні мережі принесли революцію в галузі прогнозної аналітики. На основі вхідних параметрів можна передбачити вихід або майбутнє будь-якого значення. Зараз, завдяки досягненням у галузі машинного навчання та розвитку методів глибокого навчання, спостерігається тенденція використання моделей глибокого навчання в прогностичній аналітиці.

3 ПЕРЕДБАЧАЮЧА МОДЕЛЬ ПОВЕДІНКИ КЛІЄНТІВ

У цьому розділі досліджується модель прогнозування купівельної поведінки COREL (CustOmer Purchase pREdiction ModeL) [17]. Нехай c_k буде клієнтом; d_i и d_j будуть продуктами. Коли c_k купив d_i за час t, COREL може повернути n найбільш ймовірних куплених продуктів c_k після t часу.

Оскільки c_k купив d_i за t час, ймовірність того, що c_k також купить d_j після t часу, може бути

$$p(d_j|c_k, d_i) = \frac{p(d_i|c_k, d_j)p(d_j, c_k)}{p(d_i, c_k)}$$

Припустимо, що c_k і d_i не залежать одне від одного, тобто c_k може купити будь-який продукт за час t, ймовірність може бути

$$p(d_j|c_k, d_i) = \frac{p(d_i|d_i)p(d_j, c_k)}{p(d_i)}$$

де p ($d_j|c_k$) - ймовірність того, що c_k купить d_jp($d_j|d_i$) - це ймовірність того, що покупець, який купив d_i, також купить d_j.p(d_j) – апріорна ймовірність d_j.

Нехай ω={d_1,…,d_i -1, d_i+1,…, d_m} – набір продуктів-кандидатів. Ми обчислюємо p ($d_j|c_k$, d_i) для кожного продукту d_j ∈ω, а потім ранжуємо їх. Коли передбачається, що завжди апріорна ймовірність продукту p (d_j) однакова для всіх продуктів, p (d_j) можна ігнорувати. Тому COREL може бути

$$p(d_j|c_k, d_i) \propto p(d_j|c_k)p(d_j|d_i)$$

COREL можна розуміти як двоетапний підхід: використання $p(d_j|d_i)$ для побудови колекції продуктів ω, в якій продукти пов'язані з di; використовуючи $p(d_j|d_i)$, щоб вибрати найбільш ймовірних кандидатів з ω.

Отже, найбільш важливим завданням для побудови моделі COREL є оцінка обох параметрів $p(d_j|d_i)$ і $p(d_j|c_k)$

3.1 Оцінка ймовірності покупки наступного товару клієнтом

Параметр $p(d_j|d_i)$ представляє ймовірність того, що d_j також буде куплений тим же покупцем після покупки d_i. Параметр можна оцінити, досліджуючи зв'язок між d_i і d_j, яка може бути розрахована за допомогою аналізу ринкового кошика. Коли обидва продукти входять в одну і ту ж ринкову корзину, зазвичай вважається, що між ними існує зв'язок. Використання оцінки максимальної правдоподібності

$$p(d_j|d_i) = \frac{|d_i \cap d_j|}{|d_i|}$$

де $|d_i|$ позначає кількість придбаного продукту d_i; $|d_i \cap d_j|$ - частота одночасної появи продуктів d_i і d_j в одній ринкової кошику.

Однак експеримент показує, що набір кандидатів, побудований з використанням формули вище, настільки малий, що COREL не може досягти хорошої продуктивності.

Тому пропонується створити асоціацію категорії, а потім підібрати кандидатів із асоційованої категорії товару. Як правило, веб-сайти електронної комерції присвоюють свої продукти багаторівневим категоріям. Наприклад, Adidas (www.adidas.com) має три категорії рівнів своєї продукції [18]. Для одного елемента «Adidas Superstar» його категорії від першого рівня до

третього рівня – це «Жінки-> Взуття-> Кросівки». Ми генеруємо асоціації категорій у третьому рівні категорій. Thr (d_i) позначає третю категорію товару d_i. Тому,

$$p(d_j|d_i) = |Thr(d_i) \cap \frac{Thr(d_j)}{|Thr(d_i)|}$$

Експерименти, про які повідомляється у розділі 4, демонструють, що асоціації категорій можуть розширити колекцію кандидатів. COREL досягає кращих показників, відбираючи найкращі n асоційованих категорій, ніж використання першої формули.

3.2 Оцінка ймовірності придбання продукту

Параметр p($d_j|c_k$) вказує на ймовірність того, що клієнт c_k придбає продукт d_j. Однак точно оцінити параметр практично неможливо. Відповідно до нашого досвіду онлайн-покупок, ми можемо виявити, що параметр залежить від двох факторів: популярності d_j та переваги c_k. Припустимо, що обидва фактори не залежать один від одного, формально p($d_j|c_k$) приблизно обчислюється за формулою:

$$p(d_j|c_k) \approx Hot(d_j) * Preference(c_k, d_j)$$

Вважається, що товар, який купується набагато частіше і має вищі рейтинги, ніж інші, буде більш популярним у процесі прийняття рішень покупцями.

Тому ми розробляємо модель, яка називається Heat model Hot (d_j), для розрахунку популярності продукції [19].

Перевага клієнта також відіграє важливу роль під час прийняття рішення про придбання

3.2.1 Модель Heat

Окрім потреб замовника, кількість відгуків та оцінок товару також відіграє важливу роль при прийнятті рішення про закупівлю замовника. Якщо товар отримує багато низьких оцінок або мізерні відгуки, вважається, що клієнт буде вагатися у придбанні товару, навіть якщо у нього є сильна потреба в ньому. Ми використовуємо відгуки та інформацію про продаж товару, щоб розрахувати його популярність. Вони включають Qr, кількість оглядів; Qs, середнє значення оцінок; Qa, кількість днів з моменту поставки; Qu, кількість днів з часу останнього огляду. Ми використовуємо вектор для представлення добутку d_i, в якому є чотири елементи (Qr, Qs, Qa, Qu).

Попередні дослідження показали, що SVR (Support Vector Regression) – чудовий інструмент для прогнозування завдань [20]. Модель розробляється на основі SVR для розрахунку популярності продуктів, яка називається Heat model Hot (d_i). Отримавши продукт із чотирма атрибутами Qr, Qs, Qa, Qu, Heat, модель може обчислити оцінку його популярності. Комплект поїздів є необхідним компонентом для вивчення моделі Heat.

Наскільки відомо, жоден з веб-сайтів електронної комерції не містить маркованих даних про популярність продуктів. Однак ми можемо спостерігати, що відвідувач може вирішити, який із двох продуктів є більш популярним в Інтернет-магазинах. На основі спостереження ми використовуємо наступні чотири кроки для створення моделі Heat.

По-перше**,** спираючись на підхід краудсорсингу, розробляється платформа, на якій учасники повинні вибрати більш популярний продукт із пари продуктів, що відображаються на веб-сторінці

Система працює з наступними кроками. Вибирає будь-які два продукти A і B з бази даних продуктів; відображає коефіцієнти Qr, Qs, Qa та Qu обох продуктів на веб-сторінці; учасник обирає з них більш популярного; Якщо Hot (A)> Hot (B), генерується два екземпляри набору ланцюгів.

У отриманому наборі ланцюгів таблиці 1 один примірник включає п’ять полів: Err_Qr, Err_Qs, Err_Qa, Err_Qu та мітку. Qr (A) позначає елемент Qr вектора A.

Таблиця 1 – Екземпляри набору ланцюгів

Err_Qr	Err_Qs	Err_Qa	Err_Qu	label
Qr(A)- Qr(B)	Qs(A) -Qs(B)	Qa(A)- Qa(B)	Qu(A)- Qu(B)	1
Qr(B)- Qr(A)	Qs(B) -Qs(A)	Qa(B)- Qa(A)	Qu(B)- Qu(A)	-1

По-друге**,** ми створюємо модель логістичної регресії $f(\varphi)$, яка може порівняти популярність двох продуктів. φ у моделі є вектором, де елементи, позначені як Err_Qr, Err_Qs, Err_Qa, Err_Qu, представляють різницю між елементами обох порівняних векторів продукту.

$$f(\varphi) = \frac{\exp\big(\pi(\varphi)\big)}{1 + \exp\left(\pi(\varphi)\right)}$$

де $\pi(\varphi) = \beta_0 + \beta_1 \times Err\_Qr + \beta_2 \times Err\_Qs + \beta_3 \times Err\_Qa + \beta_4 \times Err\_Qu$

Ми використовуємо набір поїздів, отриманий на кроці 1, для підготовки моделі логістичної регресії. Далі створюємо алгоритм, що використовує модель логістичної регресії $f(\varphi)$ для розрахунку популярності продуктів, описується наступним чином.

Таблиця 2 – Алгоритм 1

Алгоритм 1. Розрахунок популярності продуктів
Вхідні дані: колекція продуктів ω, модель логістичної регресії f (φ) Результат: популярність продуктів у ω кроках: 1. P ←[]

Продовження таблиці 2

Алгоритм 1. Розрахунок популярності продуктів	
2.	Для кожної пари<a,b>, a,b∈ω, a≠b
3.	φ =V(a)-V(b)
4.	оцінка=f(φ)
6.	P[a]= P[a]+оцінка-0.5; P[b]= P[b]+0.5-оцінка
9.	Кінець
10.	нормалізувати P до діапазону [0,1]
11.	Повернення P

В алгоритмі 1 масив P зберігає обчислену популярність усіх продуктів у ω в діапазоні [0, 1].

Використовуючи алгоритм 1, ми розраховуємо популярність для кожного продукту в наборі ω, і далі генеруємо набір поїздів для моделі SVR. Два приклади набору поїздів наведені в наступній таблиці, де оцінка стосується популярності товару, а Ln (Qr) – природний журнал атрибута Qr.

Таблиця 3 - Приклади набору поїздів

Ln(Qr)	Qs	Ln(Qa)	Ln(Qu)	score
1.0986	4	5.9054	5.8833	0.23539
0.69315	5	6.0497	5.9636	0.32821

У цьому дослідженні досліджується ε-SVR і μ-SVR, поєднуючись із поліноміальним ядром та радіальною базовою функцією, які використовуються як функція ядра SVR відповідно. Оскільки загальних вказівок для визначення параметрів SVR мало, це дослідження варіює параметри для вибору оптимальних значень для найкращого виконання прогнозу. Експериментальні результати показують, що □-SVR з радіальною базовою функцією може досягти найкращих показників у нашому дослідженні.

Враховуючи набір точок даних, $\{(X_1,z_1),\ldots,(X_m,z_m)\}$, таких, що $X_i \in R^n$ є вхідним, а $z_i \in R^1$ є цільовим, стандартною формою ε-SVR є

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{1} \xi_i + C \sum_{i=1}^{1} \xi_i^*$$

На тему:

$$W^{T\phi(x_i)} + b - z_i \leq \varepsilon + \xi_i$$
$$z_i - W^{T\phi(X_i)} - b \leq \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0, i = 1, \ldots, l$$

Коли модель Heat досягає найкращих показників, параметри ε-SVR становлять C=1 та ε=0.3.

3.3 Вивчення переваг клієнтів

Економічні моделі вибору зазвичай припускають, що прихована корисність людини є функцією переваги бренду та атрибутів [21]. Спільна фільтрація може бути використана для оцінки рейтингу клієнта для одного товару в електронній комерції шляхом використання рейтингів товарів, зроблених клієнтами зі схожим смаком. Однак СФ не враховує перевагу споживача щодо ціни та торгової марки продуктів, які відіграють важливу роль у прийнятті рішення про покупку. Прогнозується рейтинг c_k для d_j, використовуючи спільну фільтрацію, CF(c_k,d_j), а потім пропонуємо ієрархічну байєсівську модель дискретного вибору для вивчення переваг клієнтів перед ціною та брендом, DC(c_k,d_j). CF(c_k,d_j)*DC(c_k,d_j) відноситься до переваг замовника c_k до продукту d_j.

CF(c_k,d_j) обчислюється у формулі :

$$CF(c_k, d_j) = \sum\nolimits_{s \in S} \frac{Sim(c_k.s) \times rating(s, d_j)}{|S|}$$

де S позначає набір клієнтів, який складається з 10 найкращих найбільш подібних клієнтів з c_k; рейтинг (s, d_j) відноситься до рейтингу, який замовник виставляє для продукту d_j.

Можливі значення рейтингу визначаються за числовою шкалою від 0 (сильно не подобається) до 5 (сильно подобається). Sim (c_k, s) вказує на подібність між споживачами c_k і s, яку можна обчислити, використовуючи косинусну міру. Вектор характеристик клієнта визначається як сукупність рейтингів товарів. Наприклад, вектор ознак c_k, V(c_k)=(0, 4, 1, 0, 5) представляє, що c_k не придбав продукт d_1 (або він/вона дав 0 рейтингове значення) і дав d_2 рейтингове значення 4 тощо.

$$Sim(c_k, c_i) = \frac{V(c_k)V(c_i)}{|V(c_k)||V(c_l)}$$

Експерименти, про які повідомляється в далі, аналізують вплив CF на ефективність COREL. Експериментальні результати показують, що модель, що поєднує p(d_i|d_j) з CF, може перевершити основні моделі, використовуючи лише p(d_j|d_i) або CF для прогнозування купівельної поведінки клієнта.

Пропонується ієрархічна Байєсівська модель дискретного вибору, щоб дізнатись про переваги замовника c_k's щодо ціни та бренду [22]. Застосовуючи модель, можна підрахувати, наскільки клієнт c_k віддає перевагу продукту d_j, DC(c_k, d_j).

Поділимо ціну та бренд кожного товару відповідно на три рівні: висока, середня та низька; великий, помірний і малий бренд. Таким чином, вектор ознак x товару d має шість двійкових ознак x=(p_hi, p_me, p_lo, b_la, b_mo, b_sm), що відповідає трьом рівням цін і трьом маркам рівнів, відповідно. Лише один із трьох рівнів цін у векторі ознак має значення 1, тоді як інші дорівнюють 0. Наприклад, (p_hi=1, p_me = 0, p_lo = 0) вказує, що ціна товару знаходиться на високому рівні. Особливості торгової марки також

підпорядковуються правилу. Наприклад, (b_la=0, b_mo =0, b_sm=1) означає, що товар належить малому бренду

$$DC(c_k, d_j) = P(y_j = 1) = \frac{1}{1 + \exp(-V(d_j, c_k)}$$

Функція корисності:

$$u(d_j, c_k) = V(d_j, c_k) + e_{jk}$$

$$V(d_j, c_k) = \beta_1 \times p_{hi} + \beta_2 \times p_{me} + \beta_3 \times p_{lo} + \beta_4 \times b_{la} + \beta_5 \times b_{mo} + \beta_6 \times b_{sm}$$

$P(y_j=1)$ позначає ймовірність вибору товару d_j. Шість коефіцієнтів $\beta_1 \sim \beta_6$ у функції корисності визначаються особливостями замовника. Це означає, що кожен клієнт може зіткнутися з функцією корисності з різними коефіцієнтами. Використовуємо наступні функції для побудови вектора характеристик клієнта.

– R (нещодавно): кількість місяців, що минули з моменту останньої покупки клієнта

– F (Частота): кількість покупок за останні 12 місяців.

– M (монетарна): сума вартості від замовника за останні 12 місяців.

– Sd, що є середньоквадратичним відхиленням цін загальної закупівлі продукції замовника. Цінність виявляє звичку покупця в Інтернеті. Менший Sd означає, що споживачеві подобається купувати фіксовану різноманітність товару, тоді як більший Sd вказує на те, що клієнт не заперечує щодо ціни на товари в Інтернеті.

Вік, який позначає часовий інтервал у році від поточної дати до дати, коли клієнт вперше придбав товар на веб-сайті.

Кожен клієнт може мати перевагу щодо ціни та бренду товарів. Наприклад, хтось віддає перевагу продуктам великого бренду, тоді як інші не дбають про бренд товарів за умови, що вони дешеві [23]. В ієрархічній

байєсівській моделі коефіцієнти корисної функції визначаються особливостями клієнтів.

Використання *B* позначаєβ_1~β_6.

$$B = Z\Delta + U; \quad u_i \sim N(0, V\beta)$$

Уже споживач може мати перевагу щодо ціни та бренду товарів. Матриця *Z* містить особливості клієнтів. Матриця коефіцієнтів Δ має нормальний розподіл із середнім значенням vec(Δ) та матрицями коваріації, заданими добутком Кронекера з A-1 та *V*β.

$$\beta_n \sim N\left(\Delta` Z_n, V_\beta\right)$$
$$vec(\Delta) \sim N\left(vec(\bar{\Delta}), A^{-1} \otimes V_\beta\right)$$
$$V_\beta \sim IW(v, V)$$

Оператор vec створює вектор стовпця з матриці шляхом складання векторів стовпців Δ [24]. Гіперпараметр *V*β має інвертований пріоритет Вішарта. Ми встановлюємо неінформативні попередні *v*, *V*, Δ і *A* до v=m+3, *V*=*v***I*, Δ =0 , де *m* - кількість коефіцієнтів у функції корисності.

Параметри ієрархічної Байєсівської моделі можна описати за допомогою DAG (спрямованого ациклічного графіку) на малюнку 2.

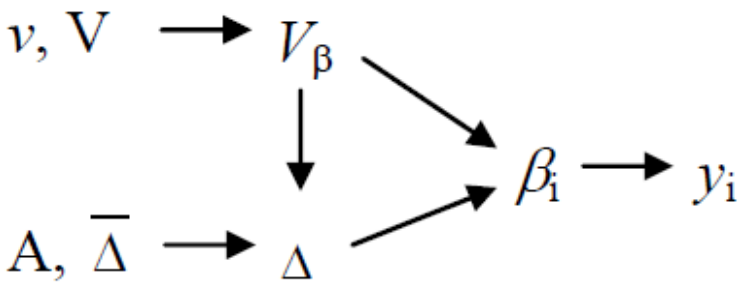

Рисунок 3.1 - Спрямований ациклічний графік

Використовується алгоритм MCMC-мегаполіс для оцінки параметрів в ієрархічній байєсівській моделі, в якій прийнято нормальний розподіл щодо пропозицій для алгоритму MCMC [24]. Функцією вірогідності журналу є:

Таблиця 4 – Алгоритм 2

Алгоритм 2. Використовуйте MCMC-алгоритм ненависті мегаполісу для оцінки параметрів
Кроки: 1. ініціюючий β_{old} 2. взяти з $V_\beta \mid v, V \sim IW(v+n, V+S)$ $vec(\Delta) \mid \Delta, A, V\beta \sim N(vec(\Delta), A^{-1} \otimes V\beta)$ 3. черпати з 4. намалювати $\beta_{new} \sim N(\beta_{old}, V_\beta)$ 5. Обчислити $\alpha(\beta_{old}, \beta_{new}) \sim \min(1, p(\beta_{new})q(\beta_{new}, \beta_{old}) / p(\beta_{old})q(\beta_{old}, \beta_{new}))$ Де $p(\beta_{new}) / p(\beta_{old}) = \exp(L(X, Y, \beta_{new}) - L(X, Y, \beta_{old}))$, $q(\beta_{new}, \beta_{old}) / q(\beta_{old}, \beta_{new}) = \exp\{(\beta'_{new} - Z\Delta) * V_\beta * (\beta_{new} - (Z\Delta)') - (\beta'_{old} - Z\Delta) * V_\beta * (\beta_{old} - (Z\Delta)'))\}$ 6. Якщо $\alpha<1$, то 7. $\beta_{old} = \beta_{new}$ з імовірністю α 8. ще 9. $\beta_{old} = \beta_{new}$ 10. Кінець 11. Перейдіть до кроку (2) до кінця циклу

Використовуючи збережені розіграші, ми можемо побудувати задній розподіл коефіцієнтів. Рисунок 3.6 ілюструє задній розподіл трьох коефіцієнтів p_hi, p_me та p_lo для одного клієнта. Можна помітити, що середній розподіл становить приблизно -2,7, 0,8 та 0,5, відповідно.

З бальної оцінки трьох коефіцієнтів можна зробити висновок, що замовник, як правило, відкидає продукцію з високою ціною і, як правило, віддає перевагу продуктам із середньою ціною та низьким цінам. Навчена модель розкриває більше інформації про клієнтів.

Рисунок 3.5 – Задній розподіл коефіцієнтів замовника

Щоб вивчити ієрархічну байєсівську модель дискретного вибору, необхідно знати вибір клієнтів у кінцевому альтернативному наборі. Однак у контексті електронної комерції ми можемо знати лише те, що придбали клієнти, а не знати, від чого клієнт відмовився у своєму виборі.

При вивчанні ієрархічної Байєсової моделі дискретного вибору, як позитивні, так і негативні вибірки є необхідними компонентами. Розглядаючи придбані товари як позитивні дані, ми розробляємо методику генерування одного негативного примірника з позитивного [25].

Одним із випадків у наборі даних поїздів є вектор ознак придбаного товару в поєднанні з етикеткою. Шість особливостей p_hi, p_me, p_lo, b_la, b_mo, b_sm в екземплярі представляють рівень ціни та рівень бренду товару відповідно.

Таблиця 5 – Позитивний екземпляр

p_hi	p_me	p_lo	b_la	b_mo	b_sm	label
1	0	0	1	0	0	1

Коли кожна ознака в позитивному екземплярі інвертована, ми можемо отримати негативний екземпляр.

Таблиця 6 – Негативний екземпляр

p_hi	p_me	p_lo	b_la	b_mo	b_sm	label
0	1	1	0	1	1	0

4 АЛГОРИТМ ВПРОВАДЖЕННЯ ПРЕДИКТИВНОГО АНАЛІЗУ

Модель предиктивного аналізу допомагає підвищити ефективність організації та досягти успішних результатів на підприємстві за допомогою даних, статистики та методів машинного навчання.

Для проведення інтелектуального аналізу слід виконати деякі основні дії, які продемонстровані на рисунку 4.1.

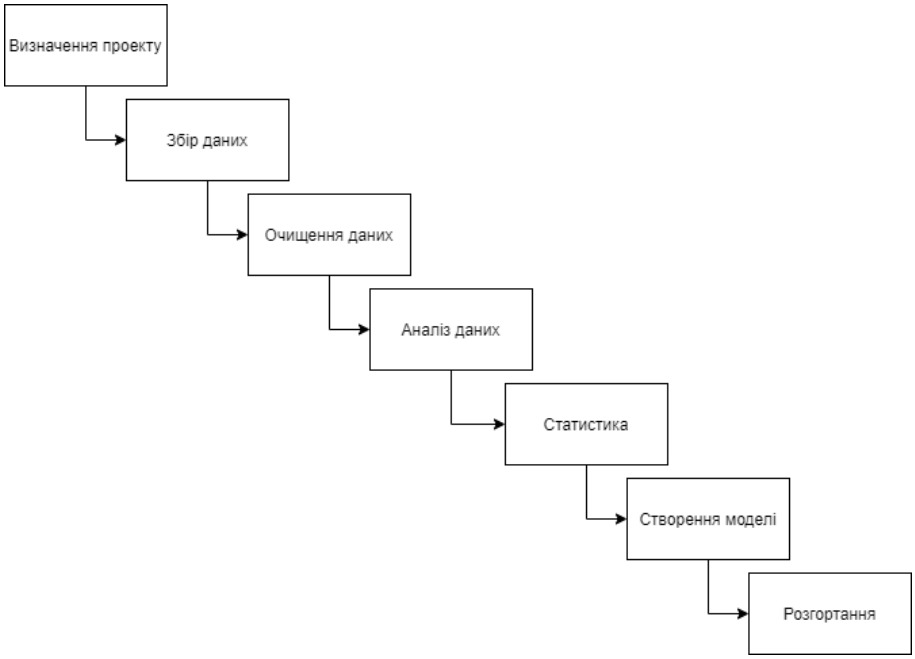

Рисунок 4.1 – Алгоритм створення прогностичної моделі

Визначення постановки проблеми становить в собі визначення результатів проекту, обсяг зусиль, цілі, визначення наборів даних, які будуть використовуватись [26].

Збір даних передбачає збір необхідних деталей, необхідних для аналізу. Він включає історичні або минулі дані з уповноваженого джерела, щодо яких слід провести предиктивний аналіз.

Очищення даних – це процес, в якому вдосконалюються набори даних. У процесі очищення даних видаляються непотрібні та помилкові дані. Це передбачає видалення зайвих даних та дублікатів даних із наборів даних. В

алгоритмі машинного навчання дані не можна використовувати в нормальному вигляді, оскільки вони є способом отримання, тому дані потрібно розробити, перш ніж використовувати їх у моделях машинного навчання. Цей прийом застосовується для вирішення проблем, які ще не відомі екстрактору знань. Це називається роботою попередньої обробки. Правильне форматування та очищення даних є важливим для попередньої обробки. Попередня обробка даних складається з наступних дій.

Набір даних імпортується, зберігаючи файл набору даних у форматі CSV. В рамках очищення даних потрібно видалити деякі стовпці, які не сприяють досягненню кінцевих результатів алгоритму. Тут випадають ідентифікатор елемента та ідентифікатор виходу.

Відсутні дані – це те, що потрібно маніпулювати, щоб не залишалося розбіжностей між даними, що надходять у модель.

Оскільки концепції машинного навчання використовують математичні моделі для вирішення задач, потрібно переконатися, що є достатньо числових даних, що підтверджують гіпотезу, щоб отримати найкращі результати,

Масштабування функцій – це метод, за допомогою якого ми масштабуємо дані до точного та масштабованого розміру з метою підвищення точності та зменшення помилок. Це в основному запобігає великим розбіжностям точок даних, які будуть використовуватися в алгоритмі, і дозволяє досягти кращих результатів

Вилучення незалежних та залежних змінних: залежними змінними є цільові або вихідні змінні, які потрібно остаточно оцінити, а потім порівняти між собою. Незалежні змінні – це ознаки або вхідні змінні, які неможливо змінити будь-якими способами, і відповідно цілі прогнозуються.

Щоб уникнути переобладнання, два окремі набори даних не імпортуються для тренування та випробування. Отже, розділення здійснюється в одному наборі даних. Набір навчальних даних – це дані, які потрібні для навчання моделі. Тестові набори даних – це ті, які можна використовувати для прогнозування результатів тесту.

Аналіз даних передбачає дослідження даних. Дані досліджуються та ретельно аналізуються, щоб визначити деякі закономірності або нові результати з набору даних. На цьому етапі знаходиться корисна інформація та робиться висновок, виявляючи деякі закономірності чи тенденції.

На етапі побудови прогнозної моделі ми використовуються різні алгоритми для побудови прогнозних моделей на основі спостережуваних закономірностей. Це вимагає знання python, R, Statistics та MATLAB тощо. Також перевіряється гіпотезу, використовуючи стандартні статистичні моделі.

Під час розгортання модель починає працювати в реальному середовищі, і це допомагає у щоденних обговореннях і робить її доступною для використання.

Прогнозування продажів є загальним і важливим використанням машинного навчання (ML). Прогнози продажів можна використовувати для визначення базових показників та визначення поступового впливу нових ініціатив, планування ресурсів у відповідь на очікуваний попит та проектування майбутніх бюджетів.

Першим кроком є завантаження даних і перетворення їх у структуру, яку потім буде використано для кожної моделей. Далі на рисунку 4.2 приведено кодування.

```
def load_data():
    return pd.read_csv("data/train.csv"))


def monthly_sales(data):
    data = data.copy()
    data.date = data.date.apply(lambda x: str(x)[:-3])
    data = data.groupby('date')['sales'].sum().reset_index()
    data.date = pd.to_datetime(data.date)
    data.to_csv('../data/monthly_data.csv')
    return data
data = load_data()
monthly_data = monthly_sales(data)
```

Рисунок 4.2 – Перетворення даних

У необробленому вигляді кожен рядок даних являє собою один день продажів в одному з десяти магазинів. Основна мета – передбачити щомісячні продажі, тому спочатку треба об’єднати всі магазини та дні в загальний щомісячний продаж.

Якщо скласти графік загального щомісячного обсягу продажів з часом, можливо побачити, що середньомісячний обсяг продажів з часом збільшується, а це означає, що дані не є стаціонарними. Щоб зробити його нерухомим, необхідно обчислити різницю між продажами за кожен місяць і додати це до кадру даних як новий стовпець. Кодування приведено на рисунку 4.3.

```
def get_diff (data):
    data ['sales_diff'] = data.sales.diff ()
    data = data.dropna ()
    return datastationary_df = get_diff (monthly_data)
```

Рисунок 4.3 – Обчислення різниці в місяцях продажів за місяць

Тепер, коли дані представляють щомісячні продажі, і їх перетворено на стаціонарні, можна налаштувати дані для різних типів моделей. Для цього треба визначити дві різні структури: одна буде використана для моделювання ARIMA, а інша – для решти моделей.

Для моделі ARIMA [27] знадобляться лише індекс дати і стовпці залежної змінної (різниця у продажах) (див.рис.4.4).

```
def generate_arima_data(data):
    dt_data = data.set_index('date').drop('sales', axis=1)
    dt_data.dropna(axis=0)
    dt_data.to_csv('../data/arima_df.csv')
    return dt_data
arima_data = generate_arima_data(stationary_df)
```

Рисунок 4.4 – Генерація даних для моделі

Для інших наших моделей створено новий фрейм даних, де кожна функція відображає продажі за попередній місяць. Щоб визначити, скільки

місяців слід включити в набір функцій, необхідно спостерігати за графіками автокореляції та часткової автокореляції та використовувати правила вибору лагів у моделюванні ARIMA. Таким чином, можна підтримувати послідовний огляд для ARIMA та регресивних моделей.

Виходячи з вищевикладеного, обрано свій період огляду до 12 місяців. Отже, створено фрейм даних, який має 13 стовпців, 1 стовпець для кожного з 12 місяців і стовпець для залежної змінної, різниці в продажах.

Зараз є дві окремі структури даних, структура ARIMA, яка включає індекс дати та часу, і контрольована структура, яка включає лаги як функції. Кодування створення контрольованої структури даних приведено на рисунку 4.5.

```
def generate_supervised(data):
    supervised_df = data.copy()

    #create column for each lag
    for i in range(1,13):
        col = 'lag_' + str(i)
        supervised_df[col] = supervised_df['sales_diff'].shift(i)

    supervised_df = supervised_df.dropna().reset_index(drop=True)
    supervised_df.to_csv('../data/model_df.csv', index=False)

    return supervised_df
model_df = generate_supervised(stationary_df)
```

Рисунок 4.5 – Створення контрольованої структури даних

Для створення та оцінки всіх моделей використано ряд допоміжних функцій, які виконують наступні функції.

– поділ тестового трену: розділ даних таким чином, що останні 12 місяців є частиною тестового набору, а решта даних використовується для навчання моделі;

– масштабування даних: за допомогою шкали min-max масштабуються дані так, щоб всі змінні потрапляли в діапазон від -1 до 1;

– зворотне масштабування: після запуску моделей використано цю допоміжну функцію, щоб змінити масштабування кроку 2;

– створення фрейм даних прогнозування: генерування фрейм даних, яке включає фактичні продажі, зафіксовані в тестовому наборі, та прогнозування результатів моделі, щоб можна було кількісно оцінити успіх;

– оцінка моделі: ця допоміжна функція зберігає середньоквадратичну помилку (rmse) та середню абсолютну помилку (mae) прогнозів для порівняння продуктивності п’яти моделей.

Для регресивних моделей ми можна використовувати структуру прогнозу відповідності бібліотеки scikit-learn [28] . Тому можна створити базову структуру моделювання, яку буде названо для кожної моделі. Функція нижче(див. рис. 4.6) викликає багато допоміжних функцій, описаних вище, для розділення даних, запуску моделі та виведення оцінок RMSE та MAE.

```
    def regressive_model(train_data, test_data, model, model_name):

        X_train, y_train, X_test, y_test, scaler_object =
            scale_data(train_data, test_data)

        mod = model
        mod.fit(X_train, y_train)
        predictions = mod.predict(X_test)
        original_df = load_data('../data/monthly_data.csv')
        unscaled = undo_scaling(predictions, X_test, scaler_object)
        unscaled_df = predict_df(unscaled, original_df)\
        get_scores(unscaled_df, original_df, model_name)
        plot_results(unscaled_df, original_df, model_name)

    train, test = tts(model_df)
    regressive_model(train, test, LinearRegression(),
'LinearRegression')
    regressive_model(train, test,
    RandomForestRegressor(n_estimators=100,max_depth=20),'RandomFores
t')
    regressive_model(train, test, XGBRegressor(n_estimators=100,
    learning_rate=0.2), 'XGBoost')
```

Рисунок 4.6 – Виклик допоміжних функцій

LSTM – це тип рекуррентної нейронної мережі, який особливо корисний для прогнозування з послідовними даними [29]. Для цього можна використати дуже простий LSTM. Для додаткової точності можна додати сезонні особливості та додаткову складність моделі (див. рис. 4.7).

```
    def lstm_model(train_data, test_data):
        X_train, y_train, X_test, y_test, scaler_object =
            scale_data(train_data, test_data)
        X_train = X_train.reshape(X_train.shape[0], 1,
X_train.shape[1])
        X_test = X_test.reshape(X_test.shape[0], 1, X_test.shape[1])
        model = Sequential()
        model.add(LSTM(4, batch_input_shape=(1, X_train.shape[1],
X_train.shape[2]), stateful=True))
        model.add(Dense(1))
        model.add(Dense(1))
        model.compile(loss='mean_squared_error', optimizer='adam')
        model.fit(X_train, y_train, epochs=200, batch_size=1,
verbose=1, shuffle=False)
        predictions = model.predict(X_test, batch_size=1)
        original_df = load_data('../data/monthly_data.csv')
        unscaled = undo_scaling(predictions, X_test, scaler_object,
lstm=True)
        unscaled_df = predict_df(unscaled, original_df)
        get_scores(unscaled_df, original_df, 'LSTM')
        plot_results(unscaled_df, original_df, 'LSTM')APIMA:
```

Рисунок 4.7 – Створення LSTM моделі

Модель ARIMA виглядає дещо інакше, ніж моделі вище. Було використано пакет статистичних моделей SARIMAX для навчання моделі та генерування динамічних прогнозів. Модель SARIMA розпадається на кілька частин.

- AR: представлений як p, є авторегресивною моделлю;
- I: представлений як d, це різничий термін;
- MA: представлений як q – це модель ковзного середнього;
- S: дозволяє нам додавати сезонний компонент.

У наведеному нижче коді на рисунку 4.8 визначено модель, а потім зроблено динамічні прогнози на останні 12 місяців даних. Для стандартних нединамічних прогнозів прогноз на наступний місяць робиться з використанням фактичних продажів за попередні місяці. На відміну від цього, для динамічних прогнозів прогноз на наступний місяць робиться з використанням прогнозованих продажів за попередні місяці.

```
    def sarimax_model(data):
        sar = sm.tsa.statespace.SARIMAX(data.sales_diff, order=(12,
0, 0), seasonal_order=(0, 1, 0, 12), trend='c').fit()
        start, end, dynamic = 40, 100, 7
        data['pred_value'] = sar.predict(start=start, end=end,
dynamic=dynamic)
        original_df = load_data('../data/monthly_data.csv')
        unscaled_df = predict_df(data, original_df)
        get_scores(unscaled_df, original_df, 'ARIMA')
        plot_results(unscaled_df, original_df, 'ARIMA')
```

Рисунок 4.8 –Створення динамічних прогнозів

Для порівняння продуктивності моделі розглянуто середньоквадратичну похибку (RMSE) та середню абсолютну похибку (MAE).

Ці вимірювання зазвичай використовуються для порівняння продуктивності моделі, але вони мають дещо іншу інтуїцію та математичне значення.

- MAE: середня абсолютна похибка в середньому повідомляє, наскільки віддалені прогнози від справжнього значення. У цьому випадку всі помилки набувають однакову вагу;
- RMSE: обчислюється RMSE, беручи квадратний корінь із суми всіх квадратних помилок. Коли формується квадрат, більші помилки мають більший вплив на загальну помилку, тоді як менші помилки не мають такої ваги на загальну помилку.

З допоміжних функцій вище використано get scores для обчислення балів RMSE та MAE для кожної моделі.

Ці оцінки були збережені у словнику та замариновані. Для порівняння цей словник було перетворено у фрейм даних Pandas[30] та побудовано результати (див. рис 4.9).

```
     def create_results_df():=
         results_dict = pickle.load(open("model_scores.p", "rb"))
         results_df = pd.DataFrame.from_dict(results_dict,
orient='index', columns=['RMSE', 'MAE', 'R2'])
         results_df = results_df.sort_values(by='RMSE',
ascending=False).reset_index()
         return results_df
     results = create_results_df()
```

Рисунок 4.9 – Фрейм даних Pandas

Це дає результати представлені на рисунку 4.10.

	index	RMSE	MAE
0	Random Forest	18599.232966	15832.750000
1	LinearRegression	16221.040791	12433.000000
2	ARIMA	14959.893467	11265.335749
3	LSTM	14638.748350	11951.083333
4	XGBoost	13574.792632	11649.666667

Рисунок 4.10 – Результати порівняння моделей

Очевидно, що, хоча результати моделей виглядають подібними на наведених вище графіках, вони різняться за ступенем точності.

Можна побачити, що в цілому модель XGBoost [31] мала найкращу продуктивність, за якою слідкували моделі ARIMA та LSTM. Тут застереження полягає в тому, що всі наведені моделі були виведені в їх найосновнішій формі, щоб продемонструвати, як їх можна використовувати для прогнозування продажів. Моделі були лише злегка налаштовані, щоб мінімізувати складність. Наприклад, LSTM може мати багато додаткових вузлів і шарів для підвищення продуктивності.

Щоб визначити, яка модель підходить для кокретного випадку використання, слід врахувати наступне.

– ступінь складності моделі проти інтерпретації, яка комфортна;

– моделі можуть бути налаштовані, а функції можуть бути спроектовані так, щоб включати сезонну інформацію, свята, вихідні тощо;

– необхідно зрозуміти, як буде використано результати та як надходитимуть дані для оновлення моделі;

– необхідно налаштувати моделі, використовуючи перехресну перевірку або подібні методи, щоб уникнути переобладнання даних.

5 РЕЗУЛЬТАТИ ВПРОВАДЖЕННЯ АЛГОРИТМУ ПРОГНОЗУВАННЯ ПРОДАЖІВ

Прогнозування продажів є важливою частиною сучасної бізнес-аналітики [32–34]. Це може бути складною проблемою, особливо у випадку відсутності даних, відсутніх даних та наявності відхилень. Продажі можна розглядати як часовий ряд. Існують деякі обмеження підходів до часових рядів для прогнозування продажів. Ось деякі з них:

– потрібно мати історичні дані протягом тривалого періоду часу, щоб визначити сезонність. Однак часто немає історичних даних для цільової змінної, наприклад у випадку запуску нового продукту. У той же час існують часові ряди продажів для подібного товару, і можна розраховувати, що новий продукт матиме подібну модель продажів;

– дані про продажі можуть мати багато відхилень та відсутні дані. Необхідно очистити викиди та інтерполювати дані, перш ніж використовувати підхід до часових рядів;

– треба взяти до уваги багато екзогенних факторів, які впливають на збут.

Прогнозування продажів – це швидше проблема регресії, ніж проблема часових рядів. Практика показує, що використання підходів до регресії часто може дати кращі результати порівняно з методами часових рядів. Алгоритми машинного навчання дозволяють знаходити закономірності в часових рядах. Можливо знайти складні закономірності в динаміці продажів, використовуючи контрольовані методи машинного навчання. Деякі з найпопулярніших – це алгоритми машинного навчання на основі дерев [35], наприклад, Random Forest [36], Gradiend Boosting Machine [37, 38]. Одне з основних припущень методів регресії полягає в тому, що закономірності, отримані в минулих даних, будуть повторюватися в майбутньому. У даних про продажі можна спостерігати кілька типів закономірностей та ефектів. Це: тенденція,

сезонність, автокореляція, закономірності, спричинені впливом таких зовнішніх факторів, як промо, ціноутворення, поведінка конкурентів. Можливо спостерігати шум у продажах. Шум, викликають фактори, які не включені до даного розгляду. У даних про продажі можліво спостерігати екстремальні значення – відхилення. Якщо потрібно провести оцінку ризику, слід взяти до уваги рівень шуму та екстремальні значення. Випадки можуть бути спричинені деякими конкретними факторами, наприклад, промо-акції, зниження ціни, погодні умови тощо. У цій роботі вивчається використання моделей машинного навчання для прогнозування часових рядів продажів. Прийнята одна модель, ефект узагальнення машинного навчання та складання декількох моделей.

Для даного аналізу використовуються історичні дані про продажі магазинів “Rossmann Store Sales” [39]. Ці дані описують продажі в магазинах Rossmann. Розрахунки проведено в середовищі Python з використанням основних пакетів pandas, sklearn, numpy, keras, matplotlib, seaborn. Для проведення аналізу був використаний блокнот Jupiter. На рисунку 5.1 показані типові часові ряди продажів, величини продажів є нормованими довільними одиницями.

Рисунок 5.1 – Типовий часовий ряд продажів

По-перше, проведено описову аналітику, яка є дослідженням розподілу продажів, візуалізації даних за допомогою різних парних сюжетів. Це корисно у

пошуку співвідношень та драйверів збуту, на яких треба зосередитись. Рисунки 5.2–5.4 показують результати дослідницького аналізу.

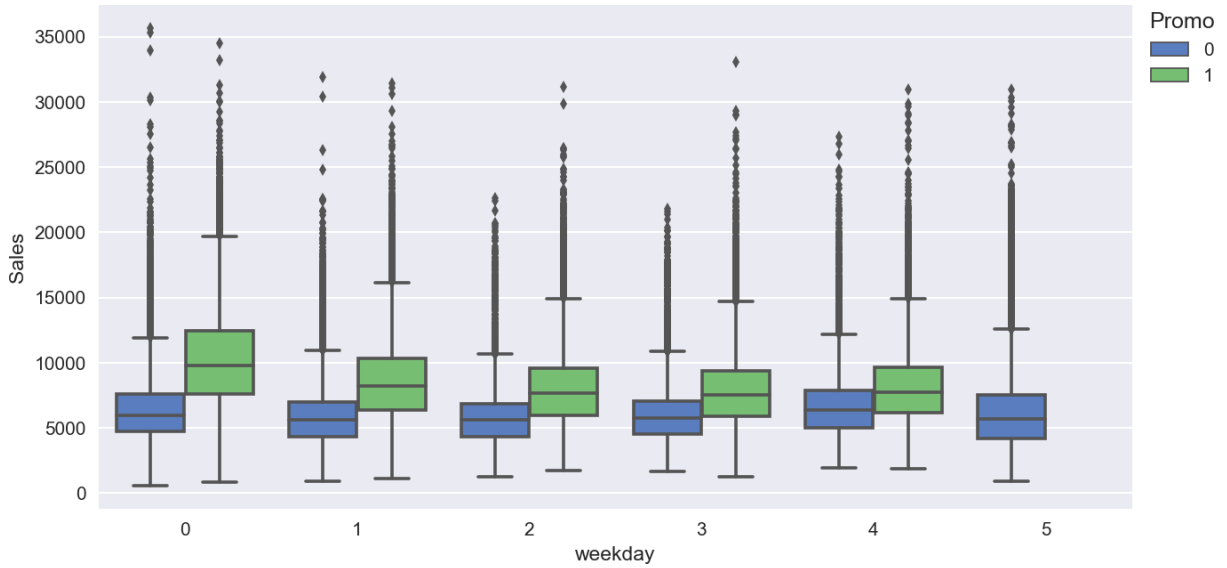

Рисунок 5.2 – Ділянки для розподілу продажів у порівнянні з днем тижня

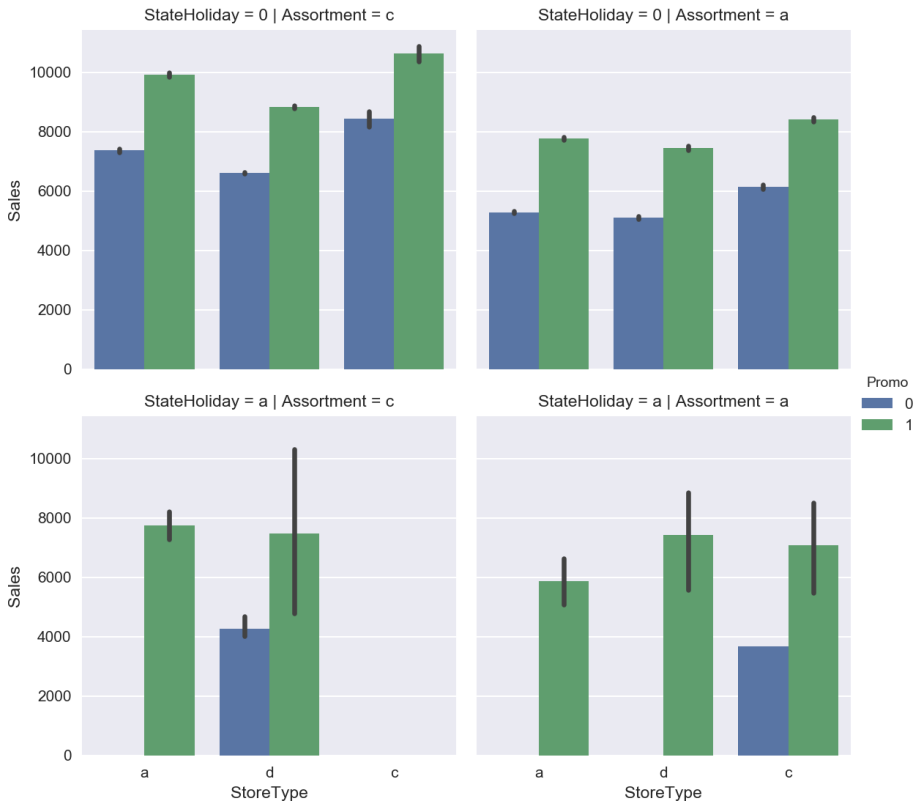

Рисунок 5.3 – Діаграми факторів для сукупних продажів

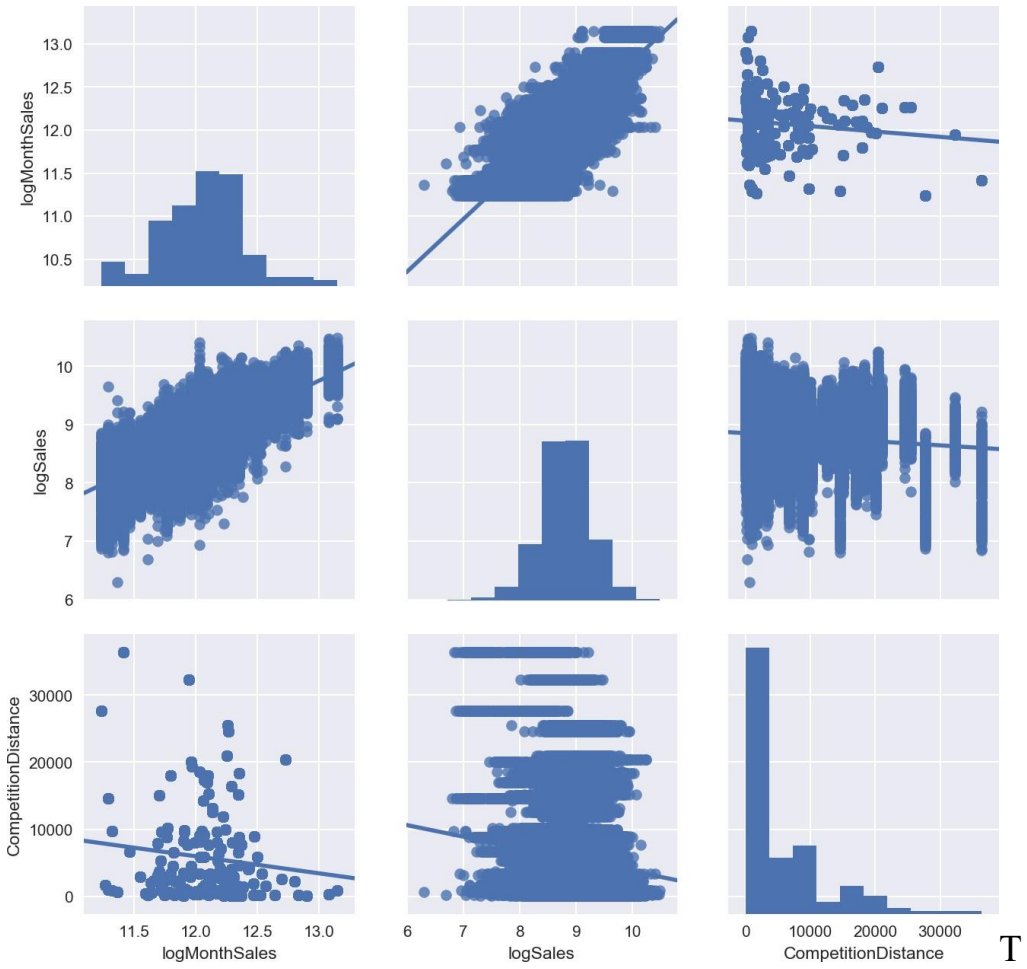

Рисунок 5.4 – Парні графіки log(MonthSales), log(Sales), CompetitionDistance

Особливістю більшості методів машинного навчання є те, що вони можуть працювати лише зі стаціонарними даними. У випадку невеликої тенденції, можна знайти зсув, використовуючи лінійну регресію на наборі перевірки.

Розглянемо підхід під наглядом машинного навчання з використанням історичних часових рядів продажів. Для тематичного дослідження використано алгоритм Random Forest [40]. Як коваріати використовано категоричні ознаки: промо, день тижня, день місяця, місяць.

Для категоріальних ознак застосовано одноразове кодування, коли одну категоріальну змінну було замінено на n двійкові змінні, де n – кількість унікальних значень категоріальних змінних.

На рисунку 5.5 показані прогнози часових рядів продажів.

Рисунок 5.5 – Прогнозування продажів (помилка набору поїздів: 3,9%, помилка набору перевірок: 11,6%)

На рисунку 5.6 показано важливість особливостей. Для оцінки помилок використано відносну середню абсолютну помилку (MAE), яка обчислюється як

$$error = \frac{MAE}{mean(Sales)} * 100\%$$

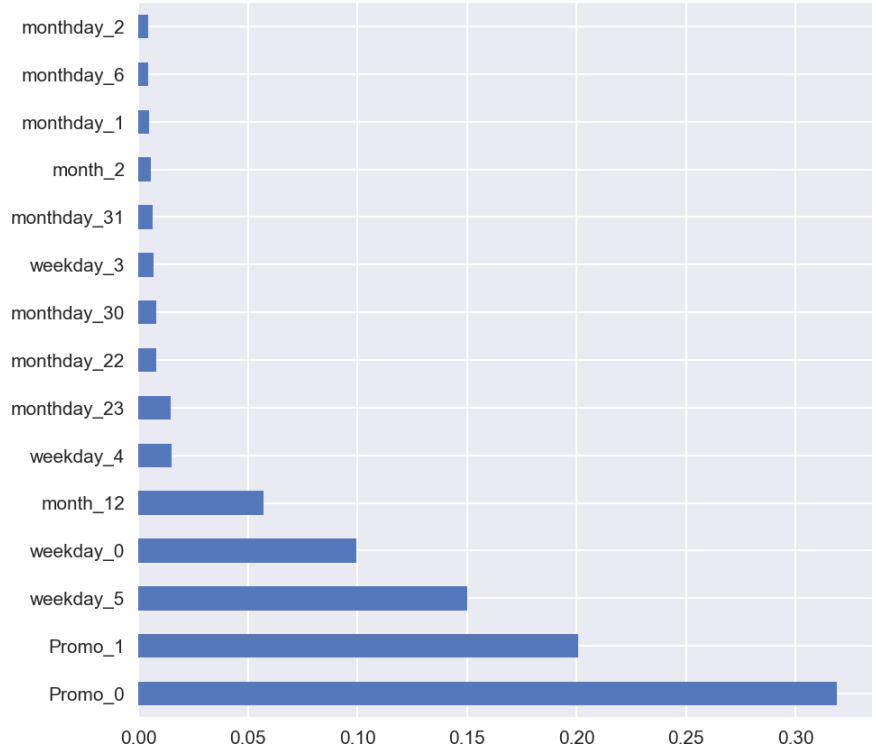

Рисунок 5.6 – Важливість особливості

На рисунку 5.7 показані прогнозні залишки для часових рядів продажів.

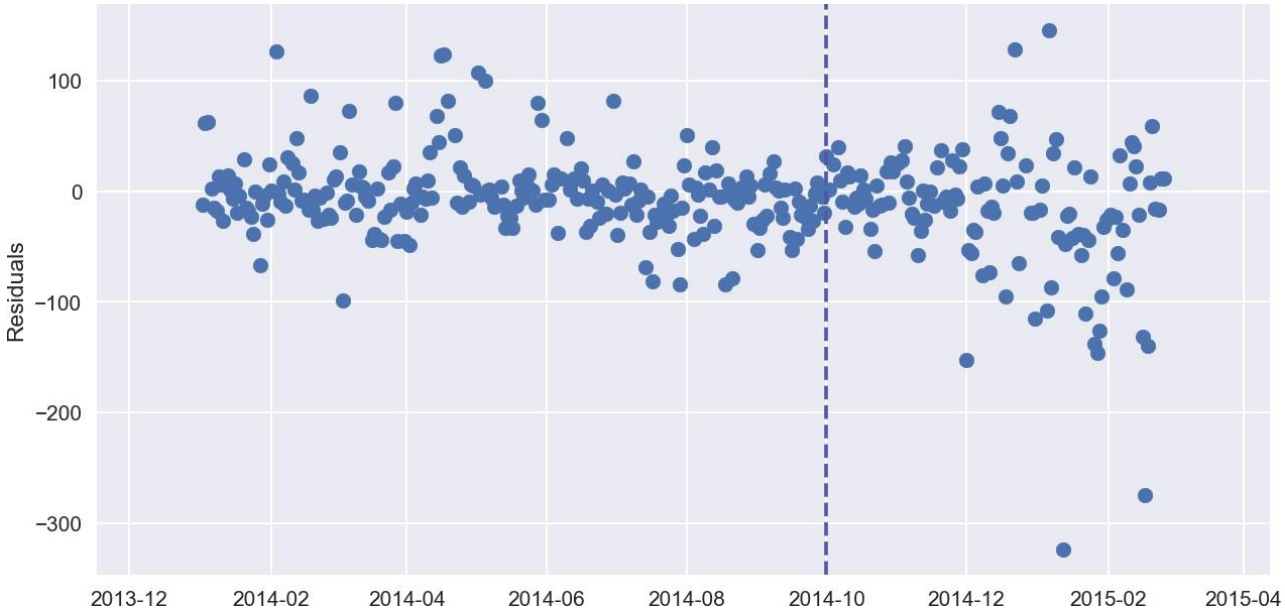

Рисунок 5.7 – Прогнозні залишки для часових рядів продажів

На рисунку 5.8 показано середнє значення залишків, що котиться.

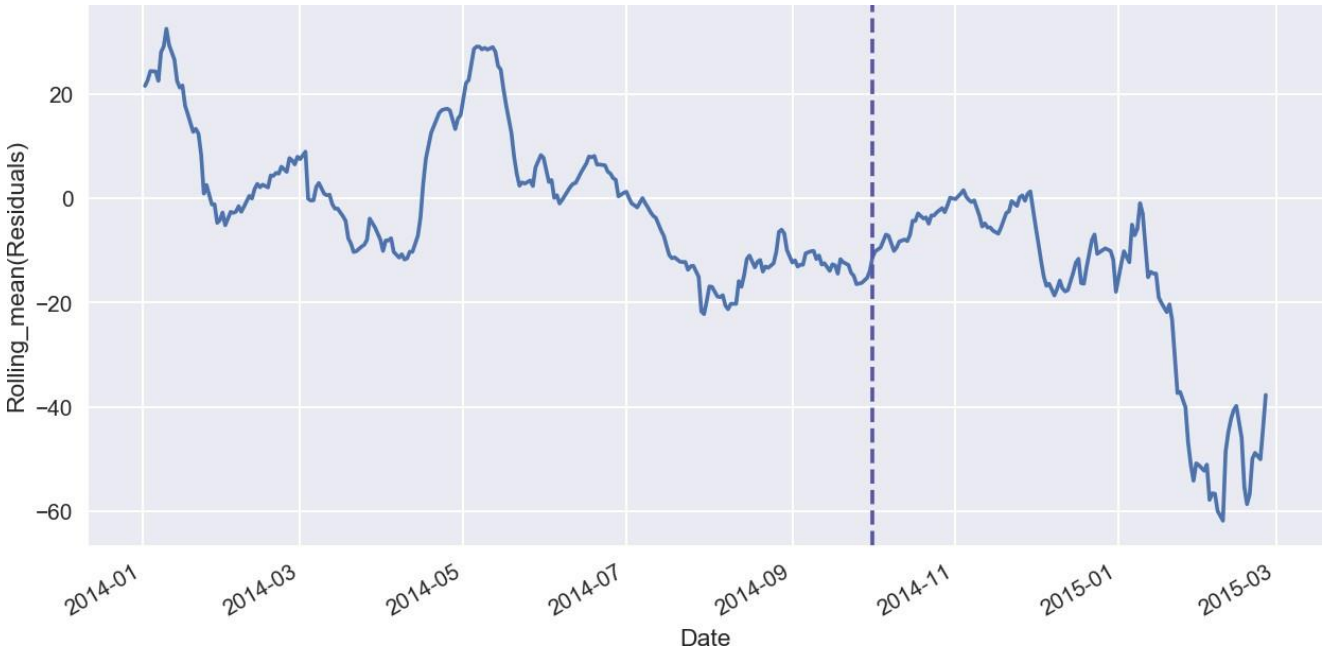

Рисунок 5.8 – Середнє значення кочення залишків

На рисунку 5.9 показано стандартне відхилення залишків прогнозу.

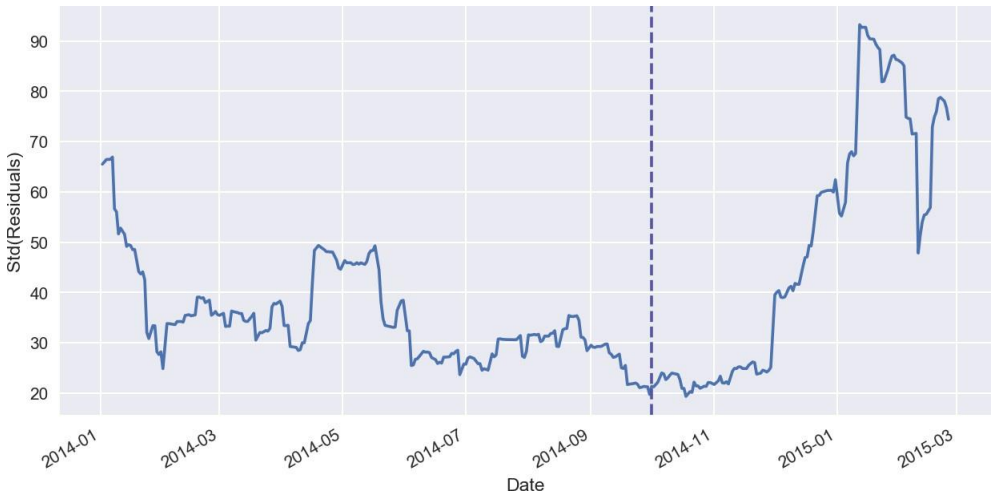

Рисунок 5.9 – Стандартне відхилення залишків прогнозу

У прогнозі можна спостерігати упередження щодо набору валідації, який є постійною (стабільною) заниженою або завищеною оцінкою продажів, коли прогноз буде вищим або нижчим щодо реальних значень. Це часто з’являється, коли застосовуються методи машинного навчання для нестаціонарних продажів. Можливо провести корекцію зміщення, використовуючи лінійну регресію на наборі перевірки. Необхідно відрізняти точність на наборі перевірки від точності на навчальному наборі. На навчальному наборі він може бути дуже високим, але на наборі перевірки – низьким. Точність набору перевірок є важливим показником для вибору оптимальної кількості ітерацій алгоритмів машинного навчання.

Ефект узагальнення машинного навчання полягає в тому, що алгоритм регресії фіксує закономірності, що існують у цілому наборі магазинів чи товарів. Якщо продажі мають виражені закономірності, то узагальнення дозволяє отримати більш точні результати, стійкі до шуму продажів. У прикладі узагальнення машинного навчання використано наступні додаткові функції щодо попереднього прикладу: середня вартість продажів за певний період історичних даних, державні та шкільні прапори, відстань від магазину до магазину конкурента, тип асортименту магазину . На рисунку 5.10 показаний прогноз у випадку історичних даних із тривалим періодом часу (2 роки) для конкретного магазину

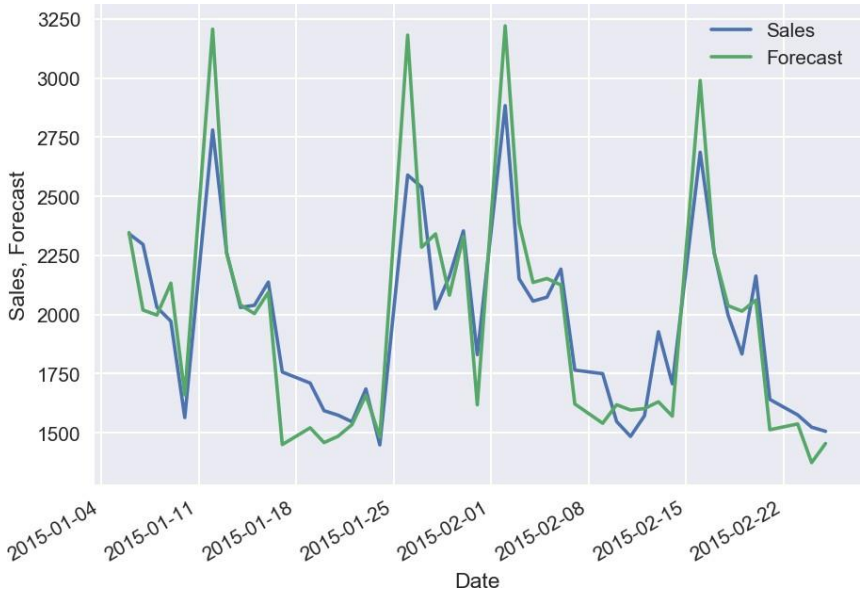

Рисунок 5.10 – Прогнозування продажів з давніми історичними даними (2 роки), похибка = 7,1%

На рисунку 5.11 показано прогноз у випадку історичних даних із коротким періодом часу (3 дні) для того самого конкретного магазину.

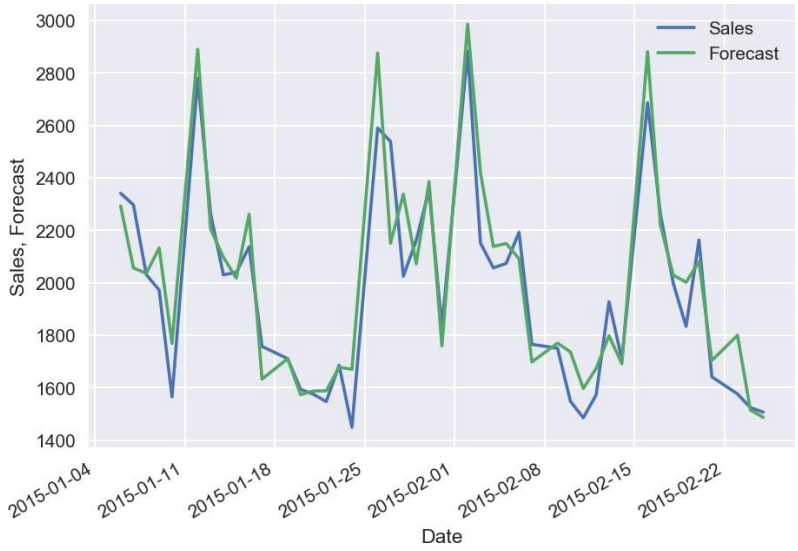

Рисунок 5.11 – Прогнозування продажів з коротким часом (3 дні), історичні дані, похибка = 5,3%

У випадку короткого періоду часу можна отримати ще більш точні результати. Ефект узагальнення машинного навчання дозволяє робити прогнози на випадок дуже малої кількості історичних даних про продажі, що важливо, коли запускається новий продукт або магазин.

Якщо необхідно прогнозувати продажі нових товарів, можна зробити експертну корекцію, помноживши прогноз на коефіцієнт, який залежить від часу, щоб врахувати перехідні процеси, наприклад, процес канібалізації товару, коли нові продукти замінюють інші продукти.

Маючи різні прогнозні моделі з різними наборами функцій, корисно об’єднати всі ці результати в одне ціле. Розглянемо методи укладання [41–45] для побудови ансамблю прогнозних моделей.

У такому підході результати прогнозування набору перевірки розглядаються як вхідні регресори для моделей наступного рівня. Як модель наступного рівня можна розглянути лінійну модель або інший тип алгоритму машинного навчання, наприклад, випадковий ліс або нейронну мережу.

Важливо зазначити, що у випадку прогнозування часових рядів не можна використовувати звичайний підхід перехресної перевірки, необхідно розділити історичний набір даних на навчальний набір та набір перевірок, використовуючи поділ періодів, тому дані навчання будуть лежати в перший часовий період, а перевірка встановлена в наступному.

На рисунку 5.12 показані прогнози часових рядів на наборах перевірки, отриманих з використанням різних моделей. Вертикальна пунктирна лінія на рисунку 5.12 відокремлює набір перевірки та набір поза вибіркою, який не використовується в процесах навчання та перевірки моделі. На наборі поза вибіркою можна розрахувати помилки укладання.

Прогнози на наборах перевірки розглядаються як регресори для лінійної моделі з регуляризацією Лассо.

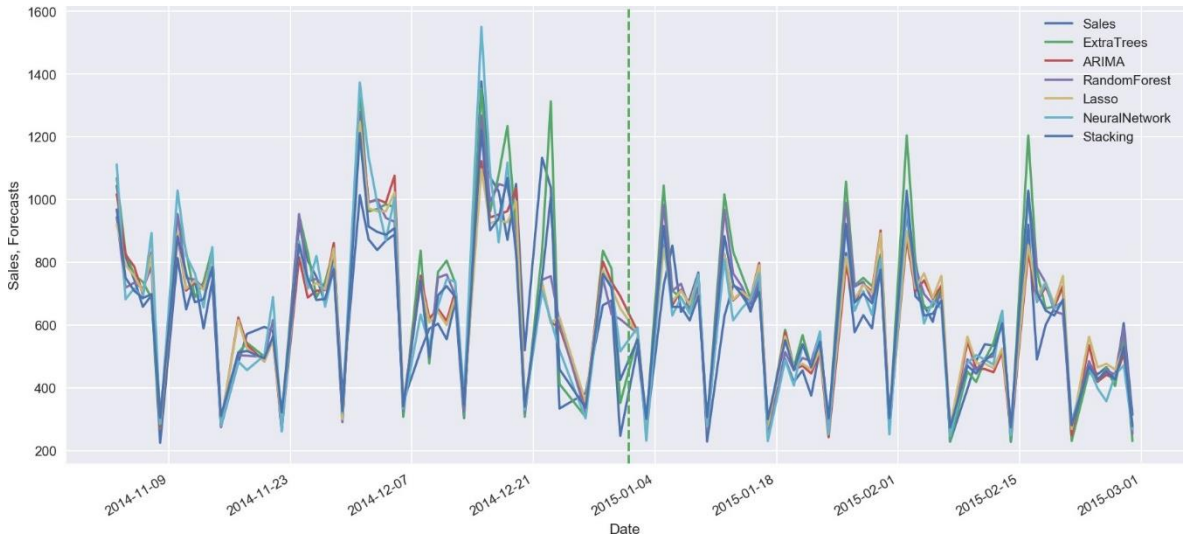

Рисунок 5.12 – Прогнозування часових рядів на наборах перевірки, отриманих з використанням різних моделей

На рисунку 5.13 показані результати, отримані за моделлю регресії Лассо другого рівня. Лише три моделі першого рівня (ExtraTree, Lasso, Neural Network) мають ненульові коефіцієнти для своїх результатів. Для інших випадків наборів даних про продажі результати можуть бути різними, коли інші моделі можуть відігравати більш важливу роль у прогнозуванні

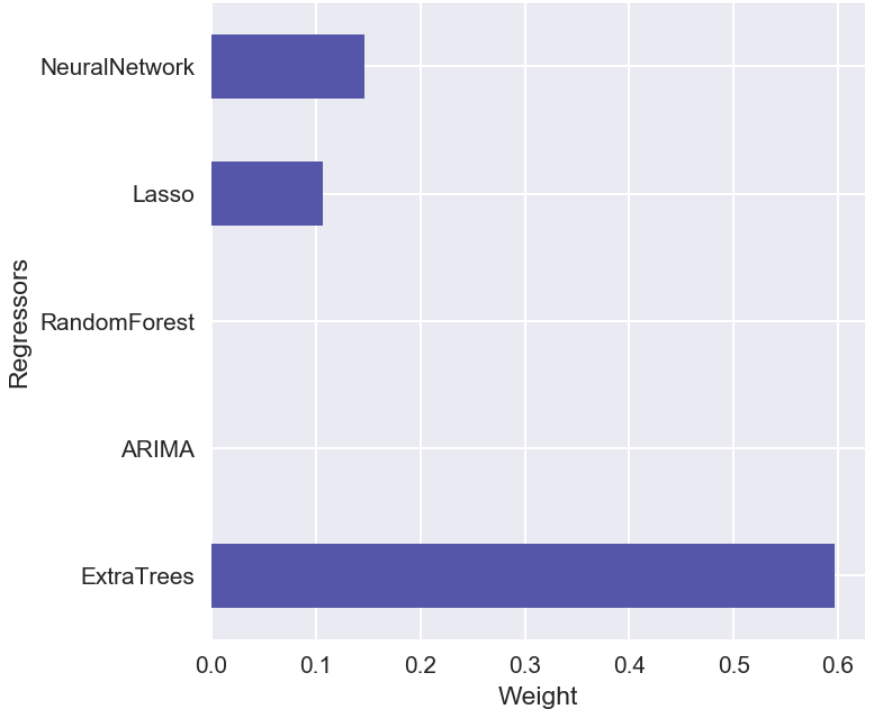

Рисунок 5.13 – Ваги укладання для регресорів

Таблиця 7 – Помилки прогнозування різних моделей.

Model	**Validation Error**	**Out-of-Sample Error**
ExtraTree	15.6%	13.9%
ARIMA	13.8%	11.4%
RandomForest	13.6%	11.9%
Lasso	13.4%	11.5%
Neural Network	13.6%	11.3%
Stacking	12.6%	10.2%

. У таблиці 7 наведено помилки в наборах перевірки та поза вибіркою. Ці результати показують, що підхід до укладання може підвищити точність перевірки та наборів, що не належать до вибірки.

Дане рішення базується на трирівневій моделі (рис. 5.14). На першому рівні використано багато одиночних моделей, більшість з яких базувались на алгоритмі машинного навчання XGBoost [46]. Для другого рівня укладання використано дві моделі з пакету scikit-learn Python – модель ExtraTree та лінійну модель від, а також модель нейронної мережі. Результати з другого рівня були підсумовані з вагами на третьому рівні.

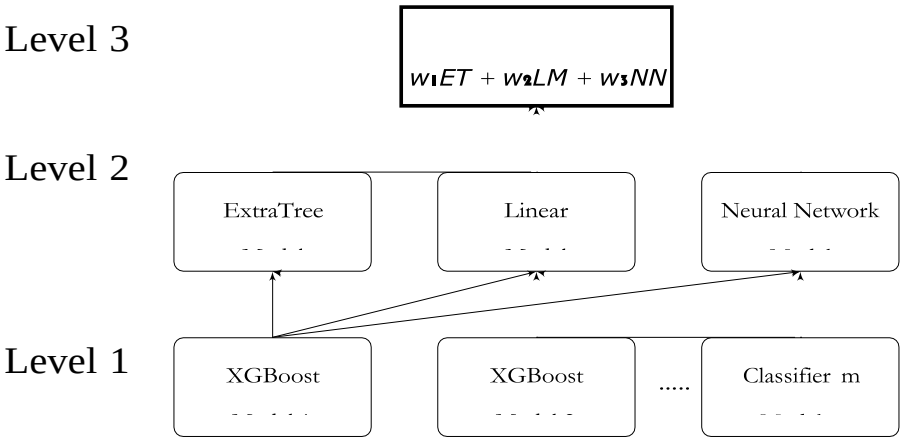

Рисунок 5.14 – Модель машинного навчання на багаторівневому рівні для прогнозування часових рядів продажів.

У даному прикладі розглянуто різні підходи до машинного навчання для прогнозування часових рядів. Прогнозування продажів – це швидше проблема регресії, ніж проблема часових рядів. Використання підходів регресії для прогнозування продажів часто може дати нам кращі результати порівняно з методами часових рядів. Одне з основних припущень методів регресії полягає в тому, що закономірності в історичних даних будуть повторюватися в майбутньому.

Точність набору перевірок є важливим показником для вибору оптимальної кількості ітерацій алгоритмів машинного навчання. Ефект узагальнення машинного навчання полягає в тому, що фіксуються закономірності у цілому наборі даних. Цей ефект можна використовувати для прогнозування продажів, коли існує невелика кількість історичних даних для конкретних часових рядів продажів у випадку запуску нового продукту чи магазину.

У підході до стекування результати кількох прогнозів моделей на наборі перевірки розглядаються як вхідні регресори для моделей наступного рівня. Як наступна модель рівня, може бути використана регресія Лассо. Використання укладання дозволяє врахувати різницю в результатах для кількох моделей з різними наборами параметрів та покращити точність перевірки та наборів даних, що не належать до вибірки.

ВИСНОВКИ

У ході виконання науково-дослідницької магістерської атестаційної роботи, було проведено дослідження існуючих методів та алгоритмів предиктивної аналітики для аналізу даних користувачів сайтів електронної комерції, проведено аналіз предметної області, запропонована модель використання предиктивного аналізу. На підставі аналізу предметної області була проведена постановка завдання.

У цій роботі було проведено дослідження декілька ключових факторів, які впливають на процес прийняття рішень клієнтами в контексті електронної комерції, включаючи потреби клієнтів, популярність продуктів та переваги споживачів. Крім того, використовуючи дані про закупівлі та рейтинги товарів, запропоновані методи кількісної оцінки сили цих факторів.

Предиктивна аналітика є одним з напрямків по обробці великих даних, дозволяє компаніям приймати більш зважені і коректні рішення сьогодні для досягнення кращих результатів завтра. Шляхом аналізу даних компанії отримують цінну інформацію і можуть вибудовувати міцні відносини зі споживачами, знаходити нові можливості, передбачити загрози, запобігати шахрайству, захищаючи доходи і репутацію. Залишається відкритим питання збереження даних, забезпечення безпеки інформаційних систем, організованих в середині компаній, а також адекватної інтерпретації даних, отриманих з різних джерел. Крім того, детального вивчення потребує питання оцінки економічних наслідків впровадження предиктивної аналітики.

В дослідженні розглянуті існуючі методи, які можна використовувати для прямих пропозицій електронних продаж. Проаналізовано основні галузі, в яких використовувалися роботи з великою кількістю даних. Можна сказати що використання методів предиктивого аналізу у електронній комерції неможливо без використання методів Data Science у галузі розробки програмного забезпечення.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Что такое e-commerce? [Електронний ресурс] – Режим доступу до ресурсу: https://trends.rbc.ru/trends/industry/5ddbf53e9a7947d0568ef37c (дата звернення: 06.09.2020).

2. E-commerce (electronic commerce) [Електронний ресурс] // Techtarget. – 2020. – Режим доступу до ресурсу: https://searchcio.techtarget.com/definition/e-commerce (дата звернення: 06.09.2020)..

3. What is eCommerce? [Електронний ресурс] // Oberlo. – 2018. – Режим доступу до ресурсу: https://www.oberlo.com/ecommerce-wiki/ecommerce (дата звернення: 06.09.2020).

4. Edwards J. What is predictive analytics? Transforming data into future insights [Електронний ресурс] / John Edwards // CIO. – 2019. – Режим доступу до ресурсу: https://www.cio.com/article/3273114/what-is-predictive-analytics-transforming-data-into-future-insights.html (дата звернення: 08.09.2020).

5. Predictive Analytics History & Current Advances [Електронний ресурс] // SAS – Режим доступу до ресурсу: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html (дата звернення: 09.09.2020).

6. Bradlow, E.T., Gangwar, M., Kopalle, P., Voleti, S.: The role of big data and predictive analytics in retailing. Journal of Retailing 93(1) (2017) 79–95

7. Abbott D. (2014) Applied predictive analytics: Principles and techniques for the professional data analyst. Indianapolis, IN: John Wiley & Sons. 456 p.

8. The Importance Predictive Analytics for E-commerce Stores [Електронний ресурс] // The Startup. – 2018. – Режим доступу до ресурсу:

https://medium.com/swlh/the-importance-of-predictive-analytics-for-e-commerce-stores-d7ef0ce2d32e (дата звернення: 09.09.2020).

9. Buytendijk F., Trepanier L. (2010) Predictive Analytics: Bringing the tools to the data/Oracle Corporation. Redwood Shores, CA.

10. Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897-904.

11. Natalia Kravets, Khrystova A. Using lambda architecture for big data analysis // Abstracts of VI International Scientific and Practical Conference. Milan, Italy 2020. pp. 491-494 pp.

12. Here's How Big Data Analytics Has Changed the eCommerce Industry [Електронний ресурс] // Smart Data Collective. – 2019. – Режим доступу до ресурсу: https://www.smartdatacollective.com/how-big-data-analytics-has-changed-ecommerce-industry/ (дата звернення: 12.09.2020).

13. Which Ecommerce Metrics You Should Measure (And Why They're Important) [Електронний ресурс] // Ecommerce Replatforming Guidebook. – 2018. – Режим доступу до ресурсу: https://www.bigcommerce.com/blog/ecommerce-metrics/#conclusion.

14. Панов М. М. Оценка деятельности и система управления компанией на основе KPI / М. М. Панов. – Москва: Инфра-М, 2013. – 255 с.

15. Heathman B. Conversion Marketing: The Online Marketing Economy / Bryan Heathman., 2012. – (Conversionmarketingbook).

16. Top 5 Predictive Analytics Models and Algorithms [Електронний ресурс] // Logi Analytics. – 2019. – Режим доступу до ресурсу: https://www.logianalytics.com/predictive-analytics/predictive-algorithms-and-models/.

17. A predictive model for customer purchase behavior in e-commerce context. // e Pacific Asia Conference on Information Systems (PACIS). – 2014. – С. 389.

18. adidas Official Website [Електронний ресурс] – Режим доступу до ресурсу: https://www.adidas.com/.

19. Leshchynskyi V. Principles of explanation in e-commerce system based on sales dynamics / Volodymyr Leshchynskyi. – Kharkiv: COMPUTER AND INFORMATION SYSTEMS AND TECHNOLOGIES, 2020. – c.76-77.

20. Method of forming recommendations using temporal constraints in a situation of cyclic cold start of the recommender system Chalyi, S., Leshchynskyi, V., Leshchynska, I. EUREKA, Physics and Engineering, 2019, 2019(4), c. 34-40

21. Sha Yang, Greg M. Allenby (2003). Modeling Interdependent Consumer Preferences. Journal of Marketing Research, 40(3), 282-294.

22. Субботин С.В., Большаков Д.Ю. Применение байесовского классификатора для распознавания классов целей. // "Журнал Радиоэлектроники" – М., 2006. – № 4.

23. Andrieu C. An Introduction to MCMC for Machine Learning / C. Andrieu, N. De Freita, A. Doucet., 2003.

24. Temporal modeling of user preferences in recommender system Chalyi, S., Leshchynskyi, V. CEUR Workshop Proceedings, 2020, 2711, c. 518-528

25. Jan R. Magnus, Heinz N. (2007). Matrix Differential Calculus with Applications in Statistics and Econometrics, JOHN WILEY & SONS, New York.'

26. Liebeskind M. 5 Machine Learning Techniques for Sales Forecasting [Електронний ресурс] / Molly Liebeskind // Towards Data Science. – 2019. – Режим доступу до ресурсу: https://towardsdatascience.com/5-machine-learning-techniques-for-sales-forecasting-598e4984b109.

27. Asteriou D. ARIMA Models and the Box–Jenkins Methodology / D. Asteriou, S. G. Hall., 2016. – 275 с. – (Applied Econometrics).

28. Документація scikit-learn [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/auto_examples/index.html.

29. Understanding LSTM Networks [Електронний ресурс]. – Режим доступу: http://colah.github.io/posts/2015-08-Understanding-LSTMs.

30. Pandas [Електронний ресурс] – https://pandas.pydata.org/

31. Introduction to Boosted Trees. [Електронний ресурс] – Режим доступу: https://xgboost.readthedocs.io/en/latest/tutorials/model.html

32. Mentzer, J.T.; Moon, M.A. Sales Forecasting Management: A Demand Management Approach; Sage: Thousand Oaks, CA, USA, 2004.

33. Efendigil, T.; Önüt, S.; Kahraman, C. A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. Expert Syst. Appl. 2009, 36, 6697–6707.

34. Zhang, G.P. Neural Networks in Business Forecasting; IGI Global: Hershey, PA, USA, 2004.

35. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning; Springer: Cham, Switzerland, 2013; Volume 112.

36. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

37. Friedman, J.H. Stochastic gradient boosting. Comput. Stat. Data Anal. 2002, 38, 367–378.

38. Pavlyshenko, B.M. Linear, machine learning and probabilistic approaches for time series analysis. In Proceedings of the IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 23–27 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 377–381.

39. Kaggle: Your Home for Data Science. Available online: http://kaggle.com (accessed on 3 November 2018).

40. Wolpert, D.H. Stacked generalization. Neural Netw. 1992, 5, 241–259. [CrossRef]

41. Rokach, L. Ensemble-based classifiers. Artif. Intell. Rev. 2010, [CrossRef]

42. Sagi, O.; Rokach, L. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2018,8, e1249. [CrossRef]

43. Gomes, H.M.; Barddal, J.P.; Enembreck, F.; Bifet, A. A survey on ensemble learning for data stream classification. ACM Comput. Surv. (CSUR) 2017, 50, 23. [CrossRef]

44. Dietterich, T.G. Ensemble methods in machine learning. In Proceedings of the International Workshop on Multiple Classifier Systems, Cagliari, Italy, 21–23 June 2000; Springer: Cham, Switzerland, 2000; pp. 1–15.

45. Rokach, L. Ensemble methods for classifiers. Data Mining and Knowledge Discovery Handbook; Springer: Cham, Switzerland, 2005; pp. 957–980.

46. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.